A Proof of Theorem 4.2

Proof. In this section, we prove the regret bound of online Lasso fitted-Q-iteration. We need a notion of restricted eigenvalue that is common in high-dimensional statistics [Bickel et al., 2009, Bühlmann and Van De Geer, 2011].

Definition A.1 (Restricted eigenvalue). Given a positive semi-definite matrix $Z \in \mathbb{R}^{d\times d}$ and integer $s \geq 1$, define the restricted minimum eigenvalue of Z as $C_{\min}(Z, s) :=$

$$\min_{\mathcal{S}\subset[d],|\mathcal{S}|\leq s}\min_{\boldsymbol{\beta}\in\mathbb{R}^d}\left\{\frac{\langle\boldsymbol{\beta}, Z\boldsymbol{\beta}\rangle}{\|\boldsymbol{\beta}_{\mathcal{S}}\|_2^2} : \|\boldsymbol{\beta}_{\mathcal{S}^c}\|_1 \leq 3\|\boldsymbol{\beta}_{\mathcal{S}}\|_1\right\}.$$

Recall that π_e is an exploratory policy that satisfies Definition 3.1, e.g.,

$$\sigma_{\min}\left(\mathbb{E}^{\pi_e}\left[\frac{1}{H}\sum_{h=1}^{H}\phi(x_h, a_h)\phi(x_h, a_h)^\top\right]\right) > 0,$$

where $x_1 \sim \xi_0, a_h \sim \pi(\cdot|x_h), x_{h+1} \sim P(\cdot|x_h, a_h)$ and $\mathbb{E}^{\pi_e}$ denotes expectation over the sample path generated under policy π_e. Recall that N_1 is the number of episodes in exploration phase that will be specified later. Denote π_{N_1} as the greedy policy with respect to the estimated Q-value calculated from the Lasso fitted-Q-iteration in Algorithm 1. According to the design of Algorithm 1, we keep using π_{N_1} for the remaining $N - N_1$ episodes after exploration phase. From the definition of the cumulative regret in Eq. (2.3), we decompose R_N according to the exploration phase and exploitation phase:

$$R_N = \sum_{n=1}^{N}\left(V_1^*(x_1^n) - V_1^{\pi_n}(x_1^n)\right) = \underbrace{\sum_{n=1}^{N_1}\left(V_1^*(x_1^n) - V_1^{\pi_e}(x_1^n)\right)}_{I_1\text{: regret during exploring}} + \underbrace{\sum_{n=N_1+1}^{N}\left(V_1^*(x_1^n) - V_1^{\pi_{N_1}}(x_1^n)\right)}_{I_2\text{:regret during exploiting}}.$$

Since we assume $r \in [0, 1]$, from the definition of value functions, it is easy to see $0 \leq V_1^*(x), V_1^{\pi_e}(x) \leq H$ for any $x \in \mathcal{X}$. Thus, we can upper bound I_1 by

$$I_1 \leq N_1 H. \tag{A.1}$$

To bound I_2, we will bound $\|V_1^* - V_1^{\pi_{N_1}}\|_\infty$ first using the following lemma. The detailed proof is deferred to Lemma B.4. Recall that $C_{\min}(\Sigma^{\pi_e}, s)$ is the restricted eigenvalue in Definition A.1 and we split the exploratory dataset into H folds with R episodes per fold.

Lemma A.2. Suppose the number of episodes in the exploration phase satisfies

$$N_1 \geq \frac{C_1 s^2 H \log(3d^2/\delta)}{C_{\min}(\Sigma^{\pi_e}, s)},$$

for some sufficiently large constant C_1 and $\lambda_1 = H\sqrt{\log(2d/\delta)/(RH)}$. Then we have with probability at least $1-\delta$,

$$\left\|V_1^{\widehat{\pi}_{N_1}} - V_1^*\right\|_\infty \leq \frac{32\sqrt{2}sH^3}{C_{\min}(\Sigma^{\pi_e}, s)}\sqrt{\frac{\log(2dH/\delta)}{N_1}}.$$

According to Lemma A.2, we have

$$I_2 \leq N\left\|V_1^{\widehat{\pi}_{N_1}} - V_1^*\right\|_\infty \leq N\frac{32\sqrt{2}sH^3}{C_{\min}(\Sigma^{\pi_e}, s)}\sqrt{\frac{\log(2dH/\delta)}{N_1}}. \tag{A.2}$$

Putting the regret bound during exploring (Eq. (A.1)) and the regret bound during exploiting (Eq. (A.2)), we have

$$R_N \leq N_1 H + N\frac{32\sqrt{2}sH^3}{C_{\min}(\Sigma^{\pi_e}, s)}\sqrt{\frac{\log(2dH/\delta)}{N_1}}.$$

We optimize N_1 by letting

$$N_1 H = N\frac{32\sqrt{2}sH^3}{C_{\min}(\Sigma^{\pi_e}, s)}\sqrt{\frac{\log(2dH/\delta)}{N_1}} \Rightarrow N_1 = \left(\frac{2048s^2H^4N^2}{C_{\min}(\Sigma^{\pi_e}, s)^2}\log(2dH/\delta)\right)^{1/3}. \tag{A.3}$$

With this choice of N_1, we have with probability at least $1-\delta$

$$R_N \leq 2H\left(\frac{2048s^2H^4N^2}{C_{\min}(\Sigma^{\pi_e}, s)^2}\log(2dH/\delta)\right)^{1/3}.$$

□

Remark A.3. The optimal choice of N_1 in Eq. (A.3) requires the knowledge of s and $C_{\min}(\Sigma, s)$ that is typically not available in practice. Thus, we can choose a relatively conservative N_1 as

$$N_1 = \left(512H^4N^2\log(2dH/\delta)\right)^{1/3},$$

such that

$$R_N \leq 4\frac{s}{C_{\min}(\Sigma^{\pi_e}, s)}H\left(512s^2H^4N^2\log(2dH/\delta)\right)^{1/3}.$$

B Additional proofs

B.1 Feature constructions

Specifically, let

$$\phi(x_0, a_k^0) = (\underbrace{0,\ldots,0}_{d+2}, \underbrace{0,\ldots,0}_{k-1}, 1, \underbrace{0,\ldots,0}_{d-k}, 1) \in \mathbb{R}^{2d+3},$$
$$\phi(x_0, a_j^0) = (\underbrace{0,\ldots,0}_{d+2}, \underbrace{0,\ldots,0}_{j-1}, 1, \underbrace{0,\ldots,0}_{d-j}, 1) \in \mathbb{R}^{2d+3}.$$

for $j \in [d]$ but $j \neq k$. In addition, we let $\psi(x_{\mathrm{i}}) = (\bar{\theta}^{(k)\top}, 0) \in \mathbb{R}^{2d+3}$ and $\psi(x_{\mathrm{u}}) = (-\bar{\theta}^{(k)\top}, 1) \in \mathbb{R}^{2d+3}$. Now we can verify for a_k^0:

$$\mathbb{P}(x_{\mathrm{u}}|x_0, a_k^0) = \phi(x_0, a_k^0)^\top\psi(x_{\mathrm{u}}) = 0,$$
$$\mathbb{P}(x_{\mathrm{i}}|x_0, a_k^0) = \phi(x_0, a_k^0)^\top\psi(x_{\mathrm{i}}) = 1,$$

and for a_j^0 $(j \neq k)$:

$$\mathbb{P}(x_{\mathrm{u}}|x_0, a_j^0) = \phi(x_0, a_j^0)^\top\psi(x_{\mathrm{u}}) = 1,$$
$$\mathbb{P}(x_{\mathrm{i}}|x_0, a_j^0) = \phi(x_0, a_j^0)^\top\psi(x_{\mathrm{i}}) = 0,$$

B.2 Proof of Claim 3.6

Proof. We prove the first part. To simplify the notation, we write φ_{nj} short for $\varphi_j(x_{\mathrm{u}}, A_2^n)$. From Eq. (3.6), we have

$$R_N(\mathcal{M}_k) \geq (H-1)\mathbb{E}_k\Big[\Big((\tau_k - 1)(s-1)\varepsilon - \sum_{n=1}^{\tau_k}\sum_{j=1}^{s-1}\varphi_{nj}\varepsilon\Big)\mathbb{I}(\mathcal{D}_k)\Big]$$
$$\geq \frac{Hs\varepsilon}{8}\mathbb{E}_k\Big[\frac{\tau_k(s-1)\varepsilon}{2}\mathbb{I}(\mathcal{D}_k)\Big].$$

Second, we derive a regret lower bound of alternative MDP $\widetilde{\mathcal{M}}_k$. Define $\widetilde{a}^* = \mathrm{argmax}_{a_j^{\mathrm{u}} \in \mathcal{A}_2} \varphi(x_{\mathrm{u}}, a_j^{\mathrm{u}})^\top \widetilde{\theta}^{(k)}$ as the optimal action when the learner is at state x_{u} in MDP $\mathcal{M}_k$. By a similar decomposition in Eq. (3.6),

$$\begin{aligned} R_N(\widetilde{\mathcal{M}}_k) &\geq (H-1)\Big(\widetilde{\mathbb{E}}_k\Big[\sum_{n=1}^{\tau_k-1}\langle\varphi(x_{\mathrm{u}},\widetilde{a}^*),\widetilde{\theta}^{(k)}\rangle\Big] - \widetilde{\mathbb{E}}_k\Big[\sum_{n=1}^{\tau_k-1}\langle\varphi_n,\widetilde{\theta}^{(k)}\rangle\Big]\Big) \\ &= (H-1)\widetilde{\mathbb{E}}_k\Big[2\tau_k(s-1)\varepsilon - \sum_{n=1}^{\tau_k}\langle\varphi_n,\widetilde{\theta}^{(k)}\rangle\Big]. \end{aligned} \tag{B.1}$$

Next, we will find an upper bound for $\sum_{n=1}^{\tau_k-1}\langle\varphi_n,\widetilde{\theta}^{(k)}\rangle$. From the definition of $\widetilde{\theta}^{(k)}$ in Eq. (3.5),

$$\begin{aligned} \sum_{n=1}^{\tau_k-1}\langle\varphi_n,\widetilde{\theta}^{(k)}\rangle &= \sum_{n=1}^{\tau_k}\langle\varphi_n,\theta+2\varepsilon\widetilde{z}^{(k)}\rangle \\ &= \sum_{n=1}^{\tau_k-1}\langle\varphi_n,\theta\rangle + 2\varepsilon\sum_{n=1}^{\tau_k-1}\langle\varphi_n,\widetilde{z}^{(k)}\rangle \\ &\leq \sum_{n=1}^{\tau_k-1}\langle\varphi_n,\theta\rangle + 2\varepsilon\sum_{n=1}^{\tau_k-1}\sum_{j\in\mathrm{supp}(\widetilde{z}^{(k)})}|\varphi_{nj}|, \end{aligned} \tag{B.2}$$

where the last inequality is from the definition of $\widetilde{z}^{(k)}$ in Eq. (3.5). To bound the first term, we have

$$\begin{aligned} \sum_{n=1}^{\tau_k-1}\langle\varphi_n,\theta\rangle &= \sum_{n=1}^{\tau_k-1}\sum_{j=1}^{s-1}\varphi_{nj}\varepsilon \\ &\leq \varepsilon\sum_{n=1}^{\tau_k-1}\sum_{j=1}^{s-1}|\varphi_{nj}|. \end{aligned} \tag{B.3}$$

Since all the φ_n come from $\mathcal{S}$ which is a $(s-1)$-sparse set, we have

$$\sum_{n=1}^{\tau_k-1}\sum_{j=1}^{d}|\varphi_{nj}| = (s-1)\tau_k,$$

which implies

$$\begin{aligned} &\sum_{n=1}^{\tau_k-1}\Big(\sum_{j=1}^{s-1}|\varphi_{nj}| + \sum_{j\in\mathrm{supp}(\widetilde{x})}|\varphi_{nj}|\Big) \leq \sum_{n=1}^{\tau_k-1}\sum_{j=1}^{d}|\varphi_{nj}| = (s-1)(\tau_k-1), \\ &\sum_{n=1}^{\tau_k-1}\sum_{j=1}^{s-1}|\varphi_{nj}| \leq (s-1)(\tau_k-1) - \sum_{n=1}^{\tau_k-1}\sum_{j\in\mathrm{supp}(\widetilde{x})}|\varphi_{nj}|. \end{aligned} \tag{B.4}$$

Combining with Eq. (B.3),

$$\sum_{n=1}^{\tau_k-1}\langle\varphi_n,\theta\rangle \leq \varepsilon\Big((s-1)(\tau_k-1) - \sum_{n=1}^{\tau_k-1}\sum_{j\in\mathrm{supp}(\widetilde{x})}|\varphi_{nj}|\Big)$$

Plugging the above bound into Eq. (B.2), it holds that

$$\sum_{n=1}^{\tau_k-1}\langle\varphi_n,\widetilde{\theta}\rangle \leq \varepsilon(s-1)(\tau_k-1) + \varepsilon\sum_{n=1}^{\tau_k}\sum_{j\in\mathrm{supp}(\widetilde{x})}|\varphi_{nj}|. \tag{B.5}$$

When the event $\mathcal{D}_k^c$ (the complement event of $\mathcal{D}_k$) happen, we have

$$\sum_{n=1}^{\tau_k-1}\sum_{j=1}^{s-1}|\varphi_{nj}| \geq \sum_{n=1}^{\tau_k-1}\sum_{j=1}^{s-1}\varphi_{nj} \geq \frac{(\tau_k-1)(s-1)}{2}.$$

Combining with Eq. (B.4), we have under event $\mathcal{D}_k^c$,

$$\sum_{n=1}^{\tau_k-1}\sum_{j\in\text{supp}(\widetilde{x})}|\varphi_{nj}| \le \frac{(\tau_k-1)(s-1)}{2}. \tag{B.6}$$

Putting Eqs. (B.1), (B.5), (B.6) together, it holds that

$$R_N(\widetilde{\mathcal{M}}_k) \ge (H-1)\widetilde{\mathbb{E}}_k\Big[\frac{(\tau_k-1)(s-1)\varepsilon}{2}\mathbb{I}(\mathcal{D}_k^c)\Big]. \tag{B.7}$$

Putting the lower bounds of $R_N(\mathcal{M}_k)$ and $R_N(\widetilde{\mathcal{M}}_k)$ together, we have

$$\begin{aligned}R_N(\mathcal{M}_k)+R_N(\widetilde{\mathcal{M}}_k) &\ge (H-1)\Big(\mathbb{E}_k\Big[\frac{(\tau_k-1)(s-1)\varepsilon}{2}\mathbb{I}(\mathcal{D}_k)\Big]+\widetilde{\mathbb{E}}_k\Big[\frac{(\tau_k-1)(s-1)\varepsilon}{2}\mathbb{I}(\mathcal{D}_k^c)\Big]\Big)\\ &= \frac{Hs\varepsilon}{8}\Big(\mathbb{E}_k\Big[\tau_k\Big(\mathbb{I}(\mathcal{D}_k)+\mathbb{I}(\mathcal{D}_k^c)\Big)\Big]+\widetilde{\mathbb{E}}_k[\tau_k\mathbb{I}(\mathcal{D}_k^c)]-\mathbb{E}_k[\tau_k\mathbb{I}(\mathcal{D}_k^c)]\Big)\\ &= \frac{Hs\varepsilon}{8}\Big(\mathbb{E}_k[\tau_k]+\widetilde{\mathbb{E}}_k[\tau_k\mathbb{I}(\mathcal{D}_k^c)]-\mathbb{E}_k[\tau_k\mathbb{I}(\mathcal{D}_k^c)]\Big).\end{aligned}$$

This ends the proof. □

B.3 Proof of Claim 3.7

Proof. The KL-calculation is inspired by Jaksch et al. [2010], but with novel stopping time argument. Denote the state-sequence up to nth episode, hth step as $\mathbb{S}_h^n = \{S_1^1,\dots,S_H^1,\dots,S_1^n,\dots,S_h^n\}$ and write $\mathcal{X}_h^n=\{x_0,x_{\rm i},x_{\rm u},x_{\rm g},x_{\rm b}\}^{(n-1)H+h}$. For a fixed policy π interacting with the environment for n episodes, we denote $\mathbb{P}_k(\cdot)$ as the distribution over $\mathbb{S}^n$, where $S_1^n=x_0$, $A_h^n\sim\pi(\cdot|S_h^n)$, $S_{h+1}^n\sim\mathbb{P}_k(\cdot|S_h^n,A_h^n)$. Let $\mathbb{E}_k$ denote the expectation w.r.t. distribution $\mathbb{P}_k$. By the chain rule, we can decompose the KL divergence as follows:

$$\text{KL}(\widetilde{\mathbb{P}}_k\|\mathbb{P}_k)=\mathbb{E}\Bigg[\sum_{n=1}^{\tau_k-1}\sum_{h=1}^{H}\text{KL}\Big[\widetilde{\mathbb{P}}_k(S_{h+1}^n|\mathbb{S}_h^n)\Big\|\mathbb{P}_k(S_{h+1}^n|\mathbb{S}_h^n)\Big]\Bigg]. \tag{B.8}$$

Given a random variable x, the KL divergence over two conditional probability distributions is defined as

$$\text{KL}\big(p(y|x),q(y|x)\big)=\sum_x\sum_y p(x,y)\log\Big(\frac{p(y|x)}{q(y|x)}\Big).$$

Then the KL divergence between $\widetilde{\mathbb{P}}_k(S_{h+1}^n|\mathbb{S}_h^n)$ and $\mathbb{P}_k(S_{h+1}^n|\mathbb{S}_h^n)$ can be calculated as follows:

$$\begin{aligned}&\text{KL}\Big[\widetilde{\mathbb{P}}_k(S_{h+1}^n|\mathbb{S}_h^n)\Big\|\mathbb{P}_k(S_{h+1}^n|\mathbb{S}_h^n)\Big]\\ &=\sum_{\mathbb{S}_h^n\in\mathcal{X}_h^n}\sum_{x\in\mathcal{X}}\widetilde{\mathbb{P}}_k(S_{h+1}^n=x,\mathbb{S}_h^n)\log\Bigg(\frac{\widetilde{\mathbb{P}}_k(S_{h+1}^n=x|\mathbb{S}_h^n)}{\mathbb{P}_k(S_{h+1}^n=x|\mathbb{S}_h^n)}\Bigg)\\ &=\sum_{\mathbb{S}_h^n\in\mathcal{X}_h^n}\sum_{x\in\mathcal{X}}\widetilde{\mathbb{P}}_k(S_{h+1}^n=x|\mathbb{S}_h^n)\widetilde{\mathbb{P}}_k(\mathbb{S}_h^n)\log\Bigg(\frac{\widetilde{\mathbb{P}}_k(S_{h+1}^n=x|\mathbb{S}_h^n)}{\mathbb{P}_k(S_{h+1}^n=x|\mathbb{S}_h^n)}\Bigg)\\ &=\sum_{\mathbb{S}_{h-1}^n\in\mathcal{X}_{h-1}^n}\widetilde{\mathbb{P}}_k(\mathbb{S}_{h-1}^n)\sum_{x'\in\mathcal{X},a\in\mathcal{A}}\widetilde{\mathbb{P}}_k(S_h^n=x',A_h^n=a|\mathbb{S}_{h-1}^n)\\ &\quad\cdot\sum_{x\in\mathcal{X}}\widetilde{\mathbb{P}}_k(S_{h+1}^n=x|\mathbb{S}_{h-1}^n,S_h^n=x',A_h^n=a)\log\Bigg(\frac{\widetilde{\mathbb{P}}_k(S_{h+1}^n=x|\mathbb{S}_{h-1}^n,S_h^n=x',A_h^n=a)}{\mathbb{P}_k(S_{h+1}^n=x|\mathbb{S}_{h-1}^n,S_h^n=x',A_h^n=a)}\Bigg).\end{aligned} \tag{B.9}$$

According to the construction of $\mathcal{M}_k$ and $\widetilde{\mathcal{M}}_k$, the learner will remain staying at the current state when $x'=x_{\rm g}$ or $x_{\rm b}$, that implies

$$\widetilde{\mathbb{P}}_k(S_{h+1}^n=x|\mathbb{S}_{h-1}^n,S_h^n=x',A_h^n=a)=\mathbb{P}_k(S_{h+1}^n=x|\mathbb{S}_{h-1}^n,S_h^n=x',A_h^n=a).$$

In addition, from the definition of stopping time τ_k, the learner will never transit to the informative state x_{i}. Therefore,

$$
\begin{aligned}
&\mathrm{KL}\Big[\widetilde{\mathbb{P}}_k(S_{h+1}^n|\mathbb{S}_h^n)\Big\|\mathbb{P}_k(S_{h+1}^n|\mathbb{S}_h^n)\Big] \\
&= \sum_{\mathbb{S}_{h-1}^n\in\mathcal{X}^{t-1}} \widetilde{\mathbb{P}}_k(\mathbb{S}_{h-1}^n) \sum_{x'=x_0,x_{\mathrm{i}},x_{\mathrm{u}}} \sum_{a\in\mathcal{A}} \widetilde{\mathbb{P}}_k(S_h^n=x', A_h^n=a|\mathbb{S}_{h-1}^n) \\
&\qquad \cdot \sum_{x\in\mathcal{X}} \widetilde{\mathbb{P}}_k(S_{h+1}^n=x|\mathbb{S}_{h-1}^n, S_h^n=x', A_h^n=a)\log\left(\frac{\widetilde{\mathbb{P}}_k(S_{h+1}^n=x|\mathbb{S}_{h-1}^n, S_h^n=x', A_h^n=a)}{\mathbb{P}_k(S_{h+1}^n=x|\mathbb{S}_{h-1}^n, S_h^n=x', A_h^n=a)}\right) \\
&= \sum_{a\in\mathcal{A}_2} \widetilde{\mathbb{P}}_k(S_h^n=x_{\mathrm{u}}, A_h^n=a) \sum_{x=x_{\mathrm{g}},x_{\mathrm{b}}} \widetilde{\mathbb{P}}_k(S_{h+1}^n=x|S_h^n=x_{\mathrm{u}}, A_h^n=a)\log\left(\frac{\widetilde{\mathbb{P}}_k(S_{h+1}^n=x|S_h^n=x_{\mathrm{u}}, A_h^n=a)}{\mathbb{P}_k(S_{h+1}^n=x|S_h^n=x_{\mathrm{u}}, A_h^n=a)}\right) \\
&= \sum_{a\in\mathcal{A}_2} \widetilde{\mathbb{P}}_k(S_h^n=x_{\mathrm{u}}, A_h^n=a)\Big(\langle\varphi(x_{\mathrm{u}},a),\widetilde{\theta}^{(k)}\rangle\log\Big(\frac{\langle\varphi(x_{\mathrm{u}},a),\widetilde{\theta}^{(k)}\rangle}{\langle\varphi(x_{\mathrm{u}},a),\theta\rangle}\Big) + (1-\langle\varphi(x_{\mathrm{u}},a),\widetilde{\theta}^{(k)}\rangle)\log\Big(\frac{1-\langle\varphi(x_{\mathrm{u}},a),\widetilde{\theta}^{(k)}\rangle}{1-\langle\varphi(x_{\mathrm{u}},a),\theta\rangle}\Big)\Big),
\end{aligned}
$$

where $\mathcal{A}_2$ is the action set associated to state x_{u}. Moreover, we will use Lemma C.4 to bound the above last term. Letting $q=\langle\varphi(x_{\mathrm{u}},a),\widetilde{\theta}^{(k)}\rangle$ and $\epsilon=\langle\varphi(x_{\mathrm{u}},a),\theta-\widetilde{\theta}^{(k)}\rangle$, it is easy to verify the conditions in Lemma C.4 as long as $\varepsilon\le(10(s-1))^{-1}$. Then we have

$$
\begin{aligned}
\mathrm{KL}\Big[\widetilde{\mathbb{P}}_k(S_{h+1}^n|\mathbb{S}_h^n)\Big\|\mathbb{P}_k(S_{h+1}^n|\mathbb{S}_h^n)\Big] &\le \sum_{a\in\mathcal{A}_2}\widetilde{\mathbb{P}}_k(S_h^n=x_{\mathrm{u}}, A_h^n=a)\frac{2\langle\widetilde{\theta}^{(k)}-\theta,\varphi(x_{\mathrm{u}},a)\rangle^2}{\langle\widetilde{\theta}^{(k)},\varphi(x_{\mathrm{u}},a)\rangle} \\
&= \sum_{a\in\mathcal{A}_2}\widetilde{\mathbb{P}}_k(S_h^n=x_{\mathrm{u}}, A_h^n=a)\frac{8\varepsilon^2\langle\widetilde{z}^{(k)},\varphi(x_{\mathrm{u}},a)\rangle^2}{\langle\widetilde{\theta},\varphi(x_{\mathrm{u}},a)\rangle}\,.
\end{aligned}
$$

Back to the KL-decomposition in Eq. (B.8), we have

$$
\mathrm{KL}(\widetilde{\mathbb{P}}_k\|\mathbb{P}_k)\le 8\varepsilon^2\widetilde{\mathbb{E}}_k\Big[\sum_{n=1}^{\tau_k-1}\langle\varphi(x_{\mathrm{u}},A_2^n),\widetilde{z}\rangle^2\Big]\,.
$$

To simplify the notations, we let $\varphi_n=\varphi(x_{\mathrm{u}},A_2^n)$.

Next, we use a simple argument "minimum is always smaller than the average". We decompose the following summation over action set $\mathcal{S}'$ defined in Eq. (3.4),

$$
\begin{aligned}
\sum_{z\in\mathcal{S}'}\sum_{n=1}^{\tau_k-1}\langle\varphi_n,z\rangle^2 &= \sum_{z\in\mathcal{S}'}\sum_{n=1}^{\tau_k-1}\Big(\sum_{j=1}^d z_j\varphi_{nj}\Big)^2 \\
&= \sum_{z\in\mathcal{S}'}\sum_{n=1}^{\tau_k-1}\Big(\sum_{j=1}^d\big(z_j\varphi_{nj}\big)^2+2\sum_{i<j}z_iz_j\varphi_{ni}\varphi_{nj}\Big).
\end{aligned}
$$

We bound the above two terms separately. To bound the first term, we observe that

$$
\sum_{z\in\mathcal{S}'}\sum_{n=1}^{\tau_k-1}\sum_{j=1}^d\big(z_j\varphi_{nj}\big)^2=\sum_{z\in\mathcal{S}'}\sum_{n=1}^{\tau_k-1}\sum_{j=1}^d|z_j\varphi_{nj}|, \tag{B.10}
$$

since both z_j,φ_{nj} can only take $-1,0,+1$. In addition, $\sum_{t=1}^{\tau_k-1}\sum_{j=1}^d|\varphi_{nj}|=(s-1)\tau_k$. Since $z\in\mathcal{S}'$ that is $(s-1)$-sparse, we have $\sum_{j=1}^d|z_j\varphi_{nj}|\le s-1$. Therefore, we have

$$
\sum_{z\in\mathcal{S}'}\sum_{n=1}^{\tau_k-1}\sum_{j=1}^d|z_j\varphi_{nj}|\le(s-1)(\tau_k-1)\binom{d-s-1}{s-2}. \tag{B.11}
$$

Putting Eqs. (B.10) and (B.11) together,

$$
\sum_{z\in\mathcal{S}'}\sum_{n=1}^{\tau_k-1}\sum_{j=1}^d\big(z_j\varphi_{nj}\big)^2\le(s-1)(\tau_k-1)\binom{d-s-1}{s-2}. \tag{B.12}
$$

To bound the second term, we observe

$$\sum_{z\in\mathcal{S}'}\sum_{n=1}^{\tau_k-1}2\sum_{i<j}z_iz_j\varphi_{ni}\varphi_{nj}=2\sum_{n=1}^{\tau_k-1}\sum_{i<j}\sum_{z\in\mathcal{S}'}z_iz_j\varphi_{ni}\varphi_{nj}.$$

From the definition of $\mathcal{S}'$, z_iz_j can only take values of $\{1*1, 1*-1, -1*1, -1*-1, 0\}$. This symmetry implies

$$\sum_{z\in\mathcal{S}'}z_iz_j\varphi_{ni}\varphi_{nj}=0,$$

which implies

$$\sum_{z\in\mathcal{S}'}\sum_{n=1}^{\tau_k-1}2\sum_{i<j}z_iz_j\varphi_{ni}\varphi_{nj}=0. \tag{B.13}$$

Combining Eqs. (B.12) and (B.13) together, we have

$$\sum_{z\in\mathcal{S}'}\sum_{n=1}^{\tau_k-1}\langle\varphi_n,z\rangle^2=\sum_{z\in\mathcal{S}'}\sum_{n=1}^{\tau_k-1}\sum_{j=1}^{d}|z_j\varphi_{nj}|\le(s-1)(\tau_k-1)\binom{d-s-1}{s-2}.$$

In the end, we use the fact that the minimum of τ_k-1 points is always smaller than its average,

$$\begin{aligned}\widetilde{\mathbb{E}}_k\Big[\sum_{n=1}^{\tau_k-1}\langle\varphi_n,\widetilde{z}\rangle^2\Big]&=\min_{z\in\mathcal{S}'}\widetilde{\mathbb{E}}_k\Big[\sum_{n=1}^{\tau_k-1}\langle\varphi_n,z\rangle^2\Big]\\&\le\frac{1}{|\mathcal{S}'|}\sum_{z\in\mathcal{S}'}\widetilde{\mathbb{E}}_k\Big[\sum_{n=1}^{\tau_k-1}\langle\varphi_n,z\rangle^2\Big]\\&=\widetilde{\mathbb{E}}_k\Big[\frac{1}{|\mathcal{S}'|}\sum_{z\in\mathcal{S}'}\sum_{n=1}^{\tau_k-1}\langle\varphi_n,z\rangle^2\Big]\\&\le\frac{(s-1)\widetilde{\mathbb{E}}_k[\tau_k-1]\binom{d-s-1}{s-2}}{\binom{d-s}{s-1}}\\&\le\frac{(s-1)^2\widetilde{\mathbb{E}}_k[\tau_k-1]}{d}.\end{aligned}$$

Therefore, we reach

$$\mathrm{KL}(\widetilde{\mathbb{P}}_k\|\mathbb{P}_k)\le\frac{8\varepsilon^2(s-1)^2\widetilde{\mathbb{E}}_k[\tau_k-1]}{d}\le\frac{8\varepsilon^2(s-1)^2N}{d}\le 8\varepsilon^2(s-1)^2,$$

since we consider the data-poor regime that $N\le d$. It is obvious to see $\mathrm{KL}(\mathbb{P}_0\|\mathbb{P}_k)=0$ from Eq. (B.9). This ends the proof. □

B.4 Proof of Lemma A.2

Proof. Recall that in the learning phase, we split the data collected in the exploration phase into H folds and each fold consists of R episodes or RH sample transitions. For the update of each step h, we use a fresh fold of samples.

Step 1. We verify that the execution of Lasso fitted-Q-iteration is equivalent to the approximate value iteration. Recall that a generic Lasso estimator with respect to a function V at step h is defined in Eq. (4.1) as

$$\widehat{w}_h(V)=\operatorname*{argmin}_{w\in\mathbb{R}^d}\Big(\frac{1}{RH}\sum_{i=1}^{RH}\Big(\Pi_{[0,H]}V(x_i^{(h)\prime})-\phi(x_i^{(h)},a_i^{(h)})^\top w\Big)^2+\lambda_1\|w\|_1\Big).$$

Denote $V_w(x)=\max_{a\in\mathcal{A}}(r(x,a)+\phi(x,a)^\top w)$. For simplicity, we write $\widehat{w}_h:=\widehat{w}_h(V_{\widehat{w}_{h+1}})$ for short. Define an approximate Bellman optimality operator $\widehat{\mathcal{T}}^{(h)}:\mathcal{X}\to\mathcal{X}$ as:

$$[\widehat{\mathcal{T}}^{(h)}V](x):=\max_a\Big[r(x,a)+\phi(x,a)^\top\widehat{w}_h(V)\Big]. \tag{B.14}$$

Note this $\widehat{\mathcal{T}}^{(h)}$ is a randomized operator that only depends data from hth fold. The Lasso fitted-Q-iteration in learning phase of Algorithm 1 is equivalent to the following approximate value iteration:

$$[\widehat{\mathcal{T}}^{(h)}\Pi_{[0,H]}V_{\widehat{w}_{h+1}}](x) = \max_a \Big[r(x,a) + \phi(x,a)^\top \widehat{w}_h\Big] = \max_a Q_{\widehat{w}_h}(x,a) = V_{\widehat{w}_h}(x). \tag{B.15}$$

Recall that the true Bellman optimality operator in state space $\mathcal{T}: \mathcal{X} \to \mathcal{X}$ is defined as

$$[\mathcal{T}V](x) := \max_a \Big[r(x,a) + \sum_{x'} P(x'|x,a)V(x')\Big]. \tag{B.16}$$

Step 2. We verify that the true Bellman operator on $\Pi_{[0,H]}V_{\widehat{w}_{h+1}}$ can also be written as a linear form. From Definition 2.1, there exists some functions $\psi(\cdot) = (\psi_k(\cdot))_{k\in\mathcal{K}}$ such that for every x, a, x', the transition function can be represented as

$$P(x'|x,a) = \sum_{k\in\mathcal{K}} \phi_k(x,a)\psi_k(x'), \tag{B.17}$$

where $\mathcal{K} \subseteq [d]$ and $|\mathcal{K}| \le s$. For a vector $\bar{w}_h \in \mathbb{R}^d$, we define its kth coordinate as

$$\bar{w}_{h,k} = \sum_{x'} \Pi_{[0,H]}V_{\widehat{w}_{h+1}}(x')\psi_k(x'), \text{ if } k \in \mathcal{K}, \tag{B.18}$$

and $\bar{w}_{h,k} = 0$ if $k \notin \mathcal{K}$. By the definition of true Bellman optimality operator in Eq. (B.16) and Eq. (B.17),

$$\begin{aligned}
[\mathcal{T}\Pi_{[0,H]}V_{\widehat{w}_{h+1}}](x) &= \max_a \Big[r(x,a) + \sum_{x'} P(x'|x,a)\Pi_{[0,H]}V_{\widehat{w}_{h+1}}(x')'\Big] \\
&= \max_a \Big[r(x,a) + \sum_{x'} \phi(x,a)^\top \psi(x')\Pi_{[0,H]}V_{\widehat{w}_{h+1}}(x')'\Big] \\
&= \max_a \Big[r(x,a) + \sum_{x'}\sum_{k\in\mathcal{K}} \phi_k(x,a)\psi_k(x')\Pi_{[0,H]}V_{\widehat{w}_{h+1}}(x')'\Big] \\
&= \max_a \Big[r(x,a) + \sum_{k\in\mathcal{K}} \phi_k(x,a)\sum_{x'} \psi_k(x')\Pi_{[0,H]}V_{\widehat{w}_{h+1}}(x')'\Big] \\
&= \max_a \Big[r(x,a) + \phi(x,a)^\top \bar{w}_h\Big].
\end{aligned} \tag{B.19}$$

We interpret $\bar{w}_h$ as the ground truth of the Lasso estimator in Eq. (4.1) at step h in terms of the following sparse linear regression:

$$\Pi_{[0,H]}V_{\widehat{w}_{h+1}}(x_i') = \phi(x_i,a_i)^\top \bar{w}_h + \varepsilon_i, i = 1\ldots, RH, \tag{B.20}$$

where $\varepsilon_i = \Pi_{[0,H]}V_{\widehat{w}_{h+1}}(x_i') - \phi(x_i,a_i)^\top \bar{w}_h$. Define the filtration $\mathcal{F}_i$ generated by $\{(x_1,a_1),\ldots,(x_i,a_i)\}$ and also the data in folds $h+1$ to H. By the definition of $V_{\widehat{w}_{h+1}}$ and $\bar{w}_h$, we have

$$\begin{aligned}
\mathbb{E}[\varepsilon_i|\mathcal{F}_i] &= \mathbb{E}\big[\Pi_{[0,H]}V_{\widehat{w}_{h+1}}(x_i')|\mathcal{F}_i\big] - \phi(x_i,a_i)^\top \bar{w}_h \\
&= \sum_{x'}[\Pi_{[0,H]}V_{\widehat{w}_{h+1}}](x')P(x'|x_i,a_i) - \phi(x_i,a_i)^\top \bar{w}_h \\
&= \sum_{k\in\mathcal{K}} \phi_k(x_i,a_i)\sum_{x'}[\Pi_{[0,H]}V_{\widehat{w}_{h+1}}](x')\psi_k(x') - \phi(x_i,a_i)^\top \bar{w}_h = 0.
\end{aligned}$$

Therefore, $\{\varepsilon_i\}_{i=1}^{RH}$ is a sequence of martingale difference noises and $|\varepsilon_i| \le H$ due to the truncation operator $\Pi_{[0,H]}$. The next lemma bounds the difference between $\widehat{w}_h$ and $\bar{w}_h$ within ℓ_1-norm. The proof is deferred to Appendix B.5.

Lemma B.1. Consider the sparse linear regression described in Eq. (B.20). Suppose the number of episodes used in step h satisfies

$$R \ge \frac{C_1 \log(3d^2/\delta)s^2}{C_{\min}(\Sigma^{\pi_e}, s)},$$

for some absolute constant $C_1 > 0$. With the choice of $\lambda_1 = H\sqrt{\log(2d/\delta)/(RH)}$, the following holds with probability at least $1-\delta$,

$$\big\|\widehat{w}_h - \bar{w}_h\big\|_1 \le \frac{16\sqrt{2}s}{C_{\min}(\Sigma^{\pi_e}, s)} H\sqrt{\frac{\log(2d/\delta)}{RH}}. \tag{B.21}$$

Step 3. We start to bound $\|V_{\widehat{w}_h} - V_h^*\|_\infty$ for each step h. By the approximate value iteration form Eq. (B.15) and the definition of optimal value function,

$$\begin{aligned}\big\|V_{\widehat{w}_h} - V_h^*\big\|_\infty &= \big\|\widehat{\mathcal{T}}^{(h)}\Pi_{[0,H]}V_{\widehat{w}_{h+1}} - \mathcal{T}V_{h+1}^*\big\|_\infty \\ &= \big\|\widehat{\mathcal{T}}^{(h)}\Pi_{[0,H]}V_{\widehat{w}_{h+1}} - \mathcal{T}\Pi_{[0,H]}V_{\widehat{w}_{h+1}}\big\|_\infty + \big\|\mathcal{T}\Pi_{[0,H]}V_{\widehat{w}_{h+1}} - \mathcal{T}V_{h+1}^*\big\|_\infty.\end{aligned} \tag{B.22}$$

The first term mainly captures the error between approximate Bellman optimality operator and true Bellman optimality operator. From linear forms Eqs. (B.15) and (B.19), it holds for any $x \in \mathcal{X}$,

$$\begin{aligned}& [\widehat{\mathcal{T}}^{(h)}\Pi_{[0,H]}V_{\widehat{w}_{h+1}}](x) - [\mathcal{T}\Pi_{[0,H]}V_{\widehat{w}_{h+1}}](x) \\ = \ & \max_a \Big[r(x,a) + \phi(x,a)^\top \widehat{w}_h\Big] - \max_a \Big[r(x,a) + \phi(x,a)^\top \bar{w}_h\Big] \\ \le \ & \max_a \big|\phi(x,a)^\top(\widehat{w}_h - \bar{w}_h)\big| \\ \le \ & \max_{a,x} \|\phi(x,a)\|_\infty \|\widehat{w}_h - \bar{w}_h\|_1.\end{aligned} \tag{B.23}$$

Applying Lemma B.1, the following error bound holds with probability at least $1-\delta$,

$$\big\|\widehat{w}_h - \bar{w}_h\big\|_1 \le \frac{16\sqrt{2}s}{C_{\min}(\Sigma^{\pi_e}, s)} H\sqrt{\frac{\log(2d/\delta)}{RH}}, \tag{B.24}$$

where R satisfies $R \ge C_1 \log(3d^2/\delta)s^2/C_{\min}(\Sigma^{\pi_e}, s)$.

Note that the samples we use between phases are mutually independent. Thus Eq. (B.24) uniformly holds for all $h \in [H]$ with probability at least $1 - H\delta$. Plugging it into Eq. (B.23), we have for any stage $h \in [H]$,

$$\big\|\widehat{\mathcal{T}}^{(h)}\Pi_{[0,H]}V_{\widehat{w}_{h+1}} - \mathcal{T}\Pi_{[0,H]}V_{\widehat{w}_{h+1}}\big\|_\infty \le \frac{16\sqrt{2}s}{C_{\min}(\Sigma^{\pi_e}, s)} H\sqrt{\frac{\log(2dH/\delta)}{RH}}, \tag{B.25}$$

holds with probability at least $1-\delta$.

To bound the second term in Eq. (B.22), we observe that

$$\begin{aligned}\big\|\mathcal{T}\Pi_{[0,H]}V_{\widehat{w}_{h+1}} - \mathcal{T}V_{h+1}^*\big\|_\infty &= \max_x \big|\mathcal{T}\Pi_{[0,H]}V_{\widehat{w}_{h+1}}(x) - \mathcal{T}V_{h+1}^*(x)\big| \\ &\le \max_x \max_a \Big|\sum_{x'} P(x'|x,a)\Pi_{[0,H]}V_{\widehat{w}_{h+1}}(x') - \sum_{x'} P(x'|x,a)\Pi_{[0,H]}V_{h+1}^*(x')\Big| \\ &\le \big\|\Pi_{[0,H]}V_{\widehat{w}_{h+1}} - V_{h+1}^*\big\|_\infty.\end{aligned} \tag{B.26}$$

Plugging Eqs. (B.25) and (B.26) into Eq. (B.22), it holds that

$$\big\|V_{\widehat{w}_h} - V_h^*\big\|_\infty \le \frac{16\sqrt{2}s}{C_{\min}(\Sigma^{\pi_e}, s)} H\sqrt{\frac{\log(2dH/\delta)}{RH}} + \big\|\Pi_{[0,H]}V_{\widehat{w}_{h+1}} - V_{h+1}^*\big\|_\infty, \tag{B.27}$$

with probability at least $1-\delta$. Recursively using Eq. (B.27), the following holds with probability $1-\delta$,

$$\begin{aligned}\big\|\Pi_{[0,H]}V_{\widehat{w}_1} - V_1^*\big\|_\infty &\le \big\|V_{\widehat{w}_1} - V_1^*\big\|_\infty \\ &= \frac{16\sqrt{2}s}{C_{\min}(\Sigma^{\pi_e}, s)} H\sqrt{\frac{\log(2dH/\delta)}{RH}} + \big\|\Pi_{[0,H]}V_{\widehat{w}_2} - V_2^*\big\|_\infty \\ &\le \big\|\Pi_{[0,H]}V_{\widehat{w}_{H+1}} - V_{H+1}^*\big\|_\infty + H^2\frac{16\sqrt{2}s}{C_{\min}(\Sigma^{\pi_e}, s)}\sqrt{\frac{\log(2dH/\delta)}{RH}} \\ &= H^2\frac{16\sqrt{2}s}{C_{\min}(\Sigma^{\pi_e}, s)}\sqrt{\frac{\log(2dH/\delta)}{RH}},\end{aligned}$$

where the first inequality is due to that $\Pi_{[0,H]}$ can only make error smaller and the last inequality is due to $V_{\widehat{w}_{H+1}} = V_{H+1}^* = 0$. From Proposition 2.14 in Bertsekas [1995],

$$\big\|V_1^{\widehat{\pi}_{N_1}} - V_1^*\big\|_\infty \le H\big\|Q_{\widehat{w}_1} - Q_1^*\big\|_\infty \le 2H\big\|\Pi_{[0,H]}V_{\widehat{w}_1} - V_1^*\big\|_\infty. \tag{B.28}$$

Putting the above together, we have with probability at least $1-\delta$,

$$\big\|V_1^{\widehat{\pi}_{N_1}} - V_1^*\big\|_\infty \leq \frac{32\sqrt{2}sH^3}{C_{\min}(\Sigma^{\pi_e}, s)}\sqrt{\frac{\log(2dH/\delta)}{N_1}},$$

when the number of episodes in the exploration phase has to satisfy

$$N_1 \geq \frac{C_1 s^2 H \log(3d^2/\delta)}{C_{\min}(\Sigma^{\pi_e}, s)},$$

for some sufficiently large constant C_1. This ends the proof. □

B.5 Proof of Lemma B.1

Proof. Denote the empirical covariance matrix induced by the exploratory policy π_e and feature map ϕ as

$$\widehat{\Sigma}^{\pi_e} := \frac{1}{R}\sum_{r=1}^R \frac{1}{H}\sum_{h=1}^H \phi(x_h^r, a_h^r)\phi(x_h^r, a_h^r)^\top.$$

Recall that Σ^{π_e} is the population covariance matrix induced by the exploratory policy π_e defined in Eq. (3.1) and feature map ϕ with $\sigma_{\min}(\Sigma^{\pi_e}) > 0$. From the definition of restricted eigenvalue in (A.1) it is easy to verify $C_{\min}(\Sigma^{\pi_e}, s) \geq \sigma_{\min}(\Sigma^{\pi_e}) > 0$. For any $i, j \in [d]$, denote

$$v_{ij}^r = \frac{1}{H}\sum_{h=1}^H \phi_i(x_h^r, a_h^r)\phi_j(x_h^r, a_h^r) - \Sigma_{ij}^{\pi_e}.$$

It is easy to verify $\mathbb{E}[v_{ij}^r] = 0$ and $|v_{ij}^r| \leq 1$ since we assume $\|\phi(x,a)\|_\infty \leq 1$. Note that samples between different episodes are independent. This implies $v_{ij}^1, \ldots, v_{ij}^R$ are independent. By standard Hoeffding's inequality (Proposition 5.10 in Vershynin [2010]), we have

$$\mathbb{P}\Big(\Big|\sum_{r=1}^R v_{ij}^r\Big| \geq \delta\Big) \leq 3\exp\Big(-\frac{C_0\delta^2}{R}\Big),$$

for some absolute constant $C_0 > 0$. Applying an union bound over $i, j \in [d]$, we have

$$\mathbb{P}\Big(\max_{i,j}\Big|\sum_{r=1}^R v_{ij}^r\Big| \geq \delta\Big) \leq 3d^2\exp\Big(-\frac{C_0\delta^2}{R}\Big)$$
$$\Rightarrow \mathbb{P}\Big(\big\|\widehat{\Sigma}^{\pi_e} - \Sigma^{\pi_e}\big\|_\infty \geq \delta\Big) \leq 3d^2\exp\Big(-\frac{C_0\delta^2}{R}\Big).$$

It implies the following holds with probability $1-\delta$,

$$\big\|\widehat{\Sigma}^{\pi_e} - \Sigma^{\pi_e}\big\|_\infty \leq \sqrt{\frac{\log(3d^2/\delta)}{R}}.$$

When the number of episodes $R \geq 32^2\log(3d^2/\delta)s^2/C_{\min}(\Sigma^{\pi_e}, s)^2$, the following holds with probability at least $1-\delta$,

$$\big\|\widehat{\Sigma}^{\pi_e} - \Sigma^{\pi_e}\big\|_\infty \leq \frac{C_{\min}(\Sigma^{\pi_e}, s)}{32s}.$$

Next lemma shows that if the restricted eigenvalue condition holds for one positive semi-definite matrix Σ_0, then it holds with high probability for another positive semi-definite matrix Σ_1 as long as Σ_0 and Σ_1 are close enough in terms of entry-wise max norm.

Lemma B.2 (Corollary 6.8 in [Bühlmann and Van De Geer, 2011])**.** Let Σ_0 and Σ_1 be two positive semi-definite block diagonal matrices. Suppose that the restricted eigenvalue of Σ_0 satisfies $C_{\min}(\Sigma_0, s) > 0$ and $\|\Sigma_1 - \Sigma_0\|_\infty \leq C_{\min}(\Sigma_0, s)/(32s)$. Then the restricted eigenvalue of Σ_1 satisfies $C_{\min}(\Sigma_1, s) > C_{\min}(\Sigma_0, s)/2$.

Applying Lemma B.2 with $\widehat{\Sigma}^{\pi_e}$ and Σ^{π_e}, we have the restricted eigenvalue of $\widehat{\Sigma}^{\pi_e}$ satisfies $C_{\min}(\widehat{\Sigma}^{\pi_e}, s) > C_{\min}(\Sigma^{\pi_e}, s)/2$ with high probability.

Note that $\{\varepsilon_i \phi_j(x_i, a_i)\}_{i=1}^{RH}$ is also a martingale difference sequence and $|\varepsilon_i \phi_j(x_i, a_i)| \leq H$. By Azuma-Hoeffding inequality,

$$\mathbb{P}\Big(\max_{j\in[d]}\Big|\frac{1}{RH}\sum_{i=1}^{RH}\varepsilon_i\phi_j(x_i,a_i)\Big| \leq H\sqrt{\frac{\log(2d/\delta)}{RH}}\Big) \geq 1-\delta.$$

Denote event $\mathcal{E}$ as

$$\mathcal{E} = \Big\{\max_{j\in[d]}\Big|\frac{1}{RH}\sum_{i=1}^{RH}\varepsilon_i\phi_j(x_i,a_i)\Big| \leq \lambda_1\Big\}.$$

Then $\mathbb{P}(\mathcal{E}) \geq 1-\delta$. Under event $\mathcal{E}$, applying (B.31) in Bickel et al. [2009], we have

$$\big\|\widehat{w}_h - \bar{w}_h\big\|_1 \leq \frac{16\sqrt{2}s\lambda_1}{C_{\min}(\Sigma^{\pi_e}, s)},$$

holds with probability at least $1-2\delta$. This ends the proof.

□

C Supporting lemmas

Lemma C.1 (Pinsker's inequality)**.** Denote $\mathbf{x} = \{x_1, \ldots, x_T\} \in \mathcal{X}^T$ as the observed states from step 1 to T. Then for any two distributions P_1 and P_2 over $\mathcal{X}^\top$ and any bounded function $f : \mathcal{X}^\top \to [0, B]$, we have

$$\mathbb{E}_1 f(\mathbf{x}) - \mathbb{E}_2 f(\mathbf{x}) \leq \sqrt{\log 2/2}B\sqrt{\mathrm{KL}(P_2\|P_1)},$$

where $\mathbb{E}_1$ and $\mathbb{E}_2$ are expectations with respect to P_1 and P_2.

Lemma C.2 (Bretagnolle-Huber inequality)**.** Let $\mathbb{P}$ and $\widetilde{\mathbb{P}}$ be two probability measures on the same measurable space $(\Omega, \mathcal{F})$. Then for any event $\mathcal{D} \in \mathcal{F}$,

$$\mathbb{P}(\mathcal{D}) + \widetilde{\mathbb{P}}(\mathcal{D}^c) \geq \frac{1}{2}\exp\left(-\mathrm{KL}(\mathbb{P}, \widetilde{\mathbb{P}})\right), \tag{C.1}$$

where $\mathcal{D}^c$ is the complement event of $\mathcal{D}$ ($\mathcal{D}^c = \Omega \setminus \mathcal{D}$) and $\mathrm{KL}(\mathbb{P}, \widetilde{\mathbb{P}})$ is the KL divergence between $\mathbb{P}$ and $\widetilde{\mathbb{P}}$, which is defined as $+\infty$, if $\mathbb{P}$ is not absolutely continuous with respect to $\widetilde{\mathbb{P}}$, and is $\int_\Omega d\mathbb{P}(\omega) \log \frac{d\mathbb{P}}{d\widetilde{\mathbb{P}}}(\omega)$ otherwise.

The proof can be found in the book of Tsybakov [2008]. When $\mathrm{KL}(\mathbb{P}, \widetilde{\mathbb{P}})$ is small, we may expect the probability measure $\mathbb{P}$ is close to the probability measure $\widetilde{\mathbb{P}}$. Note that $\mathbb{P}(\mathcal{D}) + \mathbb{P}(\mathcal{D}^c) = 1$. If $\widetilde{\mathbb{P}}$ is close to $\mathbb{P}$, we may expect $\mathbb{P}(\mathcal{D}) + \widetilde{\mathbb{P}}(\mathcal{D}^c)$ to be large.

Lemma C.3 (Divergence decomposition)**.** Let $\mathbb{P}$ and $\widetilde{\mathbb{P}}$ be two probability measures on the sequence $(A_1, Y_1, \ldots, A_n, Y_n)$ for a fixed bandit policy π interacting with a linear contextual bandit with standard Gaussian noise and parameters θ and $\widetilde{\theta}$ respectively. Then the KL divergence of $\mathbb{P}$ and $\widetilde{\mathbb{P}}$ can be computed exactly and is given by

$$\mathrm{KL}(\mathbb{P}, \widetilde{\mathbb{P}}) = \frac{1}{2}\sum_{x\in\mathcal{A}}\mathbb{E}[T_x(n)]\,\langle x, \theta - \widetilde{\theta}\rangle^2\,, \tag{C.2}$$

where $\mathbb{E}$ is the expectation operator induced by $\mathbb{P}$.

This lemma appeared as Lemma 15.1 in the book of Lattimore and Szepesvári [2020], where the reader can also find the proof.

Lemma C.4 (Lemma 20 in Jaksch et al. [2010])**.** Suppose $0 \leq q \leq 1/2$ and $\epsilon \leq 1-2q$, then

$$q\log\left(\frac{q}{q+\epsilon}\right) + (1-q)\log\left(\frac{1-q}{1-q-\epsilon}\right) \leq \frac{2\epsilon^2}{q}\,.$$

Lemma C.5 (Pinsker's inequality)**.** For measures P and Q on the same probability space $(\Omega, \mathcal{F})$, we have

$$\delta(P, Q) = \sup_{A\in\mathcal{F}}(P(A) - Q(A)) \leq \sqrt{\frac{1}{2}\mathrm{KL}(P, Q)}.$$