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, Buhlmann 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 $C_{\text{min}}(Z, s) :=$

$$\min_{S \subset [d]} \min_{0 \leq \beta \leq \mathbb{R}^s} \left\{ \frac{\langle \beta, Z \beta \rangle}{\| \beta \|_2^2} : \| \beta_S \|_1 \leq 3 \| \beta \|_1 \right\}.$$ 

Recall that $\pi_e$ is an exploratory policy that satisfies Definition 3.1, e.g.,

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

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 $\pi_e$. Recall that $N_1$ is the number of episodes in exploration phase that will be specified later. Denote $\pi_{N_1}$ 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 $\pi_{N_1}$ 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:

$$R_N = \sum_{n=1}^{N} \left( V^*_1(x^n_1) - V^{\pi_{N_1}}_1(x^n_1) \right) = \sum_{n=1}^{N_1} \left( V^*_1(x^n_1) - V^{\pi_{N_1}}_1(x^n_1) \right) + \sum_{n=N_1+1}^{N} \left( V^*_1(x^n_1) - V^{\pi_{N_1}}_1(x^n_1) \right).$$

Since we assume $r \in [0, 1]$, from the definition of value functions, it is easy to see $0 \leq V^*_1(x), V^{\pi_{N_1}}_1(x) \leq H$ for any $x \in \mathcal{X}$. Thus, we can upper bound $I_1$ by

$$I_1 \leq N_1 H. \quad (A.1)$$

To bound $I_2$, we will bound $\| V^*_1 - V^{\pi_{N_1}}_1 \|_\infty$ first using the following lemma. The detailed proof is deferred to Lemma B.4. Recall that $C_{\text{min}}(\Sigma^{\pi_e}, s)$ 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

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

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

$$\| V^*_1 - V^{\pi_{N_1}}_1 \|_\infty \leq \frac{32 \sqrt{2} s H^3}{C_{\text{min}}(\Sigma^{\pi_e}, s)} \sqrt{\frac{\log(2dH/\delta)}{N_1}}.$$

According to Lemma A.2, we have

$$I_2 \leq N \| V^*_1 - V^{\pi_{N_1}}_1 \|_\infty \leq N \frac{32 \sqrt{2} s H^3}{C_{\text{min}}(\Sigma^{\pi_e}, s)} \sqrt{\frac{\log(2dH/\delta)}{N_1}}. \quad (A.2)$$

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

$$R_N \leq N_1 H + N \frac{32 \sqrt{2} s H^3}{C_{\text{min}}(\Sigma^{\pi_e}, s)} \sqrt{\frac{\log(2dH/\delta)}{N_1}}.$$
We optimize \( N_1 \) by letting
\[
N_1 H = N \frac{32 \sqrt {2 s H^3}}{C_{\min}(\Sigma^e, s)} \sqrt{\log(2 d H/\delta)} \Rightarrow N_1 = \left( \frac{2048 s^2 H^4 N^2}{C_{\min}(\Sigma^e, s)^2} \log(2 d H/\delta) \right)^{1/3}.
\] (A.3)

With this choice of \( N_1 \), we have with probability at least \( 1 - \delta \)
\[
R_N \leq 2 H \left( \frac{2048 s^2 H^4 N^2}{C_{\min}(\Sigma^e, s)^2} \log(2 d H/\delta) \right)^{1/3}.
\]

\[\square\]

**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
\[
N_1 = \left( 512 H^4 N^2 \log(2 d H/\delta) \right)^{1/3},
\]
such that
\[
R_N \leq 4 \frac{s}{C_{\min}(\Sigma^e, s)} H \left( 512 s^2 H^4 N^2 \log(2 d H/\delta) \right)^{1/3}.
\]

**B Additional proofs**

**B.1 Feature constructions**

Specifically, let
\[
\phi(x_0, a^0_k) = (0, \ldots, 0, 0, \ldots, 0, 1, 0, \ldots, 0, 1) \in \mathbb{R}^{2d+3},
\]
\[
\phi(x_0, a^0_j) = (0, \ldots, 0, 0, \ldots, 0, 1, 0, \ldots, 0, 0) \in \mathbb{R}^{2d+3}.
\]

for \( j \in [d] \) but \( j \neq k \). In addition, we let \( \psi(x) = (\vec{\phi}(k)^T, 0) \in \mathbb{R}^{2d+3} \) and \( \psi(x) = (-\vec{\phi}(k)^T, 1) \in \mathbb{R}^{2d+3} \). Now we can verify for \( a^0_k \):
\[
\mathbb{P}(x_u|x_0, a^0_k) = \phi(x_0, a^0_k)^T \psi(x_u) = 0,
\]
\[
\mathbb{P}(x_i|x_0, a^0_k) = \phi(x_0, a^0_k)^T \psi(x_i) = 1,
\]
and for \( a^0_j (j \neq k) \):
\[
\mathbb{P}(x_u|x_0, a^0_j) = \phi(x_0, a^0_j)^T \psi(x_u) = 1,
\]
\[
\mathbb{P}(x_i|x_0, a^0_j) = \phi(x_0, a^0_j)^T \psi(x_i) = 0.
\]

**B.2 Proof of Claim 3.6**

**Proof.** We prove the first part. To simplify the notation, we write \( \varphi_{nj} \) short for \( \varphi_j(x_u, A^0_n) \). From Eq. (3.6), we have
\[
R_N(M_k) \geq (H - 1) \mathbb{E}_k \left[ (\tau_k - 1)(s - 1)\varepsilon - \sum_{n=1}^{s-1} \sum_{j=1}^{\tau_k} \varphi_{nj}\varepsilon \right] \mathbb{I}(D_k)
\]
\[
\geq \frac{H s \varepsilon}{8} \mathbb{E}_k \left[ \frac{\tau_k(s - 1)\varepsilon}{2} \right] \mathbb{I}(D_k).
\]
Second, we derive a regret lower bound of alternative MDP $\widetilde{M}_k$. Define $\tilde{a}^* = \arg\max_{a \in A_k} \varphi(x_u, a^*_u)^\top \tilde{\theta}^{(k)}$ as the optimal action when the learner is at state $x_u$ in MDP $M_k$. By a similar decomposition in Eq. (3.6),

$$
R_N(\widetilde{M}_k) \geq (H - 1) \left[ \mathbb{E}_k \left[ \sum_{n=1}^{\tau_k-1} \langle \varphi(x_u, a^*_u), \tilde{\theta}^{(k)} \rangle \right] - \mathbb{E}_k \left[ \sum_{n=1}^{\tau_k-1} \langle \varphi_n, \tilde{\theta}^{(k)} \rangle \right] \right] 
$$

$$
= (H - 1) \mathbb{E}_k \left[ 2\tau_k(s-1)\varepsilon - \sum_{n=1}^{\tau_k} \langle \varphi_n, \tilde{\theta}^{(k)} \rangle \right].
$$

(B.1)

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

$$
\sum_{n=1}^{\tau_k-1} \langle \varphi_n, \tilde{\theta}^{(k)} \rangle = \sum_{n=1}^{\tau_k-1} \langle \varphi_n, \theta + 2\varepsilon \tilde{z}^{(k)} \rangle
$$

$$
= \sum_{n=1}^{\tau_k-1} \langle \varphi_n, \theta \rangle + 2\varepsilon \sum_{n=1}^{\tau_k-1} \langle \varphi_n, \tilde{z}^{(k)} \rangle
$$

(B.2)

$$
\leq \sum_{n=1}^{\tau_k-1} \langle \varphi_n, \theta \rangle + 2\varepsilon \sum_{n=1}^{\tau_k-1} \sum_{j \in \text{supp}(\tilde{z}^{(k)})} |\varphi_{nj}|,
$$

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

$$
\sum_{n=1}^{\tau_k-1} \langle \varphi_n, \theta \rangle = \sum_{n=1}^{\tau_k-1} \sum_{j=1}^{s-1} \varphi_{nj} \varepsilon
$$

$$
\leq \varepsilon \sum_{n=1}^{\tau_k-1} \sum_{j=1}^{s-1} |\varphi_{nj}|.
$$

(B.3)

Since all the $\varphi_n$ come from $S$ 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

$$
\sum_{n=1}^{\tau_k-1} \left( \sum_{j=1}^{s-1} |\varphi_{nj}| + \sum_{j \in \text{supp}(\tilde{z})} |\varphi_{nj}| \right) \leq \sum_{n=1}^{\tau_k-1} \sum_{j=1}^{d} |\varphi_{nj}| = (s - 1)(\tau_k - 1),
$$

(B.4)

$$
\sum_{n=1}^{\tau_k-1} \sum_{j=1}^{s-1} |\varphi_{nj}| \leq (s - 1)(\tau_k - 1) - \sum_{n=1}^{\tau_k-1} \sum_{j \in \text{supp}(\tilde{z})} |\varphi_{nj}|.
$$

Combining with Eq. (B.3),

$$
\sum_{n=1}^{\tau_k-1} \langle \varphi_n, \theta \rangle \leq \varepsilon \left( (s - 1)(\tau_k - 1) - \sum_{n=1}^{\tau_k-1} \sum_{j \in \text{supp}(\tilde{z})} |\varphi_{nj}| \right)
$$

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

$$
\sum_{n=1}^{\tau_k-1} \langle \varphi_n, \tilde{\theta} \rangle \leq \varepsilon(s - 1)(\tau_k - 1) + \varepsilon \sum_{n=1}^{\tau_k-1} \sum_{j \in \text{supp}(\tilde{z})} |\varphi_{nj}|.
$$

(B.5)

When the event $D_k^c$ (the complement event of $D_k$) 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 $D_k^c$,

$$\frac{\tau_k - 1}{n} \sum_{j \in \text{supp}(\pi)} |\varphi_{nj}| \leq \frac{(\tau_k - 1)(s - 1)}{2}. \quad \text{(B.6)}$$

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

$$R_N(\mathcal{M}_k) \geq (H - 1) \mathbb{E}_k \left[ \frac{(\tau_k - 1)(s - 1)}{2} \mathbb{I}(D_k^c) \right]. \quad \text{(B.7)}$$

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

$$R_N(\mathcal{M}_k) + R_N(\tilde{\mathcal{M}}_k) \geq (H - 1) \left( \mathbb{E}_k \left[ \frac{(\tau_k - 1)(s - 1)}{2} \mathbb{I}(D_k^c) \right] + \mathbb{E}_k \left[ \frac{(\tau_k - 1)(s - 1)}{2} \mathbb{I}(D_k^c) \right] \right)$$

$$= \frac{H \varepsilon}{S} \left( \mathbb{E}_k [\tau_k] \mathbb{I}(D_k^c) + \mathbb{E}_k [\tau_k \mathbb{I}(D_k^c)] \right)$$

$$= \frac{H \varepsilon}{S} \left( \mathbb{E}_k [\tau_k] + \mathbb{E}_k [\tau_k \mathbb{I}(D_k^c)] - \mathbb{E}_k [\tau_k \mathbb{I}(D_k^c)] \right).$$

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 $n$th episode, $h$th step as $S^n_h = \{S^n_1, S^n_2, \ldots, S^n_h, \ldots, S^n_n\}$ and write $X^n_h = \{x_0, x_1, x_2, x_3\}^{(n-1)H+h}$.

For a fixed policy $\pi$ interacting with the environment for $n$ episodes, we denote $P_k(\cdot)$ as the distribution over $S^n$, where $S^n_1 = x_0, A^n_h \sim \pi(\cdot|S^n_h), S^n_{h+1} \sim P_k(\cdot|S^n_h, A^n_h)$. Let $E_k$ denote the expectation w.r.t. distribution $P_k$. By the chain rule, we can decompose the KL divergence as follows:

$$\text{KL}(\tilde{P}_k||P_k) = E_k \left[ \sum_{n=1}^{H} \sum_{h=1}^{n-1} \text{KL} \left( P_k(S^n_{h+1}|S^n_h) || P_k(S^n_{h+1}|S^n_h) \right) \right]. \quad \text{(B.8)}$$

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

$$\text{KL}(p(y|x), q(y|x)) = \sum_x \sum_{y} p(x, y) \log \left( \frac{p(y|x)}{q(y|x)} \right).$$

Then the KL divergence between $\tilde{P}_k(S^n_{h+1}|S^n_h)$ and $P_k(S^n_{h+1}|S^n_h)$ can be calculated as follows:

$$\text{KL} \left( \tilde{P}_k(S^n_{h+1}|S^n_h) || P_k(S^n_{h+1}|S^n_h) \right)$$

$$= \sum_{S^n_h \in X^n_h} \sum_{x \in X} \tilde{P}_k(S^n_{h+1} = x, S^n_h) \log \left( \frac{\tilde{P}_k(S^n_{h+1} = x|S^n_h)}{P_k(S^n_{h+1} = x|S^n_h)} \right)$$

$$= \sum_{S^n_h \in X^n_h} \sum_{x \in X} \tilde{P}_k(S^n_{h+1} = x|S^n_h) \tilde{P}_k(S^n_h) \log \left( \frac{\tilde{P}_k(S^n_{h+1} = x|S^n_h)}{P_k(S^n_{h+1} = x|S^n_h)} \right)$$

$$= \sum_{S^n_{h+1} \in X^n_{h+1}} \sum_{x' \in X, a \in A} \tilde{P}_k(S^n_{h+1} = x', A^n_h = a|S^n_{h-1})$$

$$\cdot \sum_{x \in X} \tilde{P}_k(S^n_{h+1} = x|S^n_{h-1}, S^n_h = x', A^n_h = a) \log \left( \frac{\tilde{P}_k(S^n_{h+1} = x|S^n_{h-1}, S^n_h = x', A^n_h = a)}{P_k(S^n_{h+1} = x|S^n_{h-1}, S^n_h = x', A^n_h = a)} \right). \quad \text{(B.9)}$$

According to the construction of $\mathcal{M}_k$ and $\tilde{\mathcal{M}}_k$, the learner will remain staying at the current state when $x' = x_e$ or $x_h$, that implies

$$\tilde{P}_k(S^n_{h+1} = x|S^n_{h-1}, S^n_h = x', A^n_h = a) = P_k(S^n_{h+1} = x|S^n_{h-1}, S^n_h = x', A^n_h = a).$$
In addition, from the definition of stopping time $\tau_k$, the learner will never transit to the informative state $x_t$. Therefore,

$$\text{KL}\left[\bar{P}_k(S^n_{h+1}|S^n_h)\left|\bar{P}_k(S^n_{h+1}|S^n_h)\right.\right] = \sum_{S^n_{h-1} \in \mathcal{X}^{-1}} \sum_{x''=x_0, x_u, a \in \mathcal{A}} \sum_{x'=x_0, x_u, a \in \mathcal{A}} \bar{P}_k(S^n_h = x\prime, A^n_h = a) \log \left( \frac{\bar{P}_k(S^n_{h+1} = x|S^n_{h-1}, S^n_h = x\prime, A^n_h = a)}{\bar{P}_k(S^n_{h+1} = x|S^n_{h-1}, S^n_h = x', A^n_h = a)} \right)
$$

and

$$\text{KL}\left[\bar{P}_k(S^n_{h+1}|S^n_h)\left|\bar{P}_k(S^n_{h+1}|S^n_h)\right.\right] \leq \sum_{a \in \mathcal{A}_2} \sum_{S^n_h = x_u, A^n_h = a} 2\left( \frac{\langle \varphi(x_u, a), \varphi(x_u, a) \rangle}{\langle \varphi(x_u, a) \rangle} \right) \frac{\sum_{x' \in \mathcal{X}} \sum_{a \in \mathcal{A}} \bar{P}_k(S^n_h = x_u, A^n_h = a) \sum_{x'' = x_u, \theta} \bar{P}_k(S^n_h = x_u, A^n_h = a) \langle \langle \varphi(x_u, a), \varphi(x_u, a) \rangle \rangle \log \left( \frac{1 - \langle \varphi(x_u, a), \varphi(x_u, a) \rangle}{1 - \langle \varphi(x_u, a), \varphi(x_u, a) \rangle} \right)}{\langle \theta, \varphi(x_u, a) \rangle}.
$$

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

$$\text{KL}(\bar{P}_k\|\bar{P}_k) \leq 8\varepsilon^2 \varepsilon_k \left[ \sum_{n=1}^{\tau_k-1} \langle \varphi(x_u, A^n_u), z \rangle^2 \right].
$$

To simplify the notations, we let $\varphi_n = \varphi(x_u, A^n_u)$. 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),

$$\sum_{z \in \mathcal{S}} \sum_{n=1}^{\tau_k-1} \langle \varphi_n, z \rangle^2 = \sum_{z \in \mathcal{S}'} \sum_{n=1}^{\tau_k-1} \left( \sum_{j=1}^d z_j \varphi_nj \right)^2
$$

$$= \sum_{z \in \mathcal{S}'} \sum_{n=1}^{\tau_k-1} \left( \sum_{j=1}^d z_j \varphi_nj \right)^2 + 2 \sum_{i<j} z_i z_j \varphi_{ni} \varphi_{nj}.
$$

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

$$\sum_{z \in \mathcal{S}'} \sum_{n=1}^{\tau_k-1} \sum_{j=1}^d \left( z_j \varphi_nj \right)^2 = \sum_{z \in \mathcal{S}'} \sum_{n=1}^{\tau_k-1} \sum_{j=1}^d | z_j \varphi_nj |^2
$$

since both $z_j, \varphi_nj$ can only take $-1, 0, +1$. In addition, $\sum_{j=1}^{\tau_k-1} \sum_{n=1}^{\tau_k-1} | \varphi_nj | = (s-1)\tau_k$. Since $z \in \mathcal{S}'$ that is $(s-1)$-sparse, we have $\sum_{j=1}^d | z_j \varphi_nj | \leq s - 1$. Therefore, we have

$$\sum_{z \in \mathcal{S}'} \sum_{n=1}^{\tau_k-1} \sum_{j=1}^d | z_j \varphi_nj | \leq (s-1)(\tau_k-1) \left( \frac{d-s-1}{s-2} \right).
$$

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

$$\sum_{z \in \mathcal{S}'} \sum_{n=1}^{\tau_k-1} \sum_{j=1}^d (z_j \varphi_nj)^2 \leq (s-1)(\tau_k-1) \left( \frac{d-s-1}{s-2} \right).
$$
To bound the second term, we observe
\[
\sum_{z \in S'} \sum_{n=1}^{\tau_k-1} 2 \sum_{i<j} z_i z_j \varphi_{ni} \varphi_{nj} = 2 \sum_{n=1}^{\tau_k-1} \sum_{i<j} z_i z_j \varphi_{ni} \varphi_{nj}.
\]
From the definition of $S'$, $z_i z_j$ can only take values of $\{1 \ast 1, 1 \ast -1, -1 \ast 1, -1 \ast -1, 0\}$. This symmetry implies
\[
\sum_{z \in S'} z_i z_j \varphi_{ni} \varphi_{nj} = 0,
\]
which implies
\[
\sum_{z \in S'} \sum_{n=1}^{\tau_k-1} 2 \sum_{i<j} z_i z_j \varphi_{ni} \varphi_{nj} = 0. \tag{B.13}
\]
Combining Eqs. (B.12) and (B.13) together, we have
\[
\sum_{z \in S'} \sum_{n=1}^{\tau_k-1} \langle \varphi_n, z \rangle^2 = \sum_{z \in S'} \sum_{n=1}^{\tau_k-1} \sum_{j=1}^{d} |z_j \varphi_{nj}| \leq (s-1)(\tau_k - 1)\left( \frac{d-s-1}{s-2} \right).
\]
In the end, we use the fact that the minimum of $\tau_k - 1$ points is always smaller than its average,
\[
\bar{\mathbb{E}}_k \left[ \sum_{n=1}^{\tau_k-1} \langle \varphi_n, z \rangle^2 \right] = \min_{z \in S'} \bar{\mathbb{E}}_k \left[ \sum_{n=1}^{\tau_k-1} \langle \varphi_n, z \rangle^2 \right] \\
\leq \frac{1}{|S'|} \sum_{z \in S'} \bar{\mathbb{E}}_k \left[ \sum_{n=1}^{\tau_k-1} \langle \varphi_n, z \rangle^2 \right] \\
= \bar{\mathbb{E}}_k \left[ \frac{1}{|S'|} \sum_{z \in S'} \sum_{n=1}^{\tau_k-1} \langle \varphi_n, z \rangle^2 \right] \\
\leq \frac{(s-1)\bar{\mathbb{E}}_k[\tau_k - 1](d-s-1)}{(s-1)(d-s-1)} \\
\leq \frac{(s-1)^2 \bar{\mathbb{E}}_k[\tau_k - 1]}{d}.
\]
Therefore, we reach
\[
\text{KL}(\bar{\mathbb{P}}_k | \mathbb{P}_k) \leq \frac{8\varepsilon^2(s-1)^2 \bar{\mathbb{E}}_k[\tau_k - 1]}{d} \leq \frac{8\varepsilon^2(s-1)^2 N}{d} \leq \frac{8\varepsilon^2(s-1)^2}{d},
\]
since we consider the data-poor regime that $N \leq d$. It is obvious to see $\text{KL}(\mathbb{P}_0 | \mathbb{P}_k) = 0$ 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
\[
\hat{w}_h(V) = \arg\min_{w \in \mathbb{R}^d} \left( \frac{1}{RH} \sum_{i=1}^{RH} \left( \Pi_{[0,R]} V(x_i^{(h)}) - \phi(x_i^{(h)}, a_i^{(h)})^T w \right)^2 + \lambda_1 \|w\|_1 \right).
\]
Denote $V_w(x) = \max_{a \in A}(r(x,a) + \phi(x,a)^T w)$. For simplicity, we write $\hat{w}_h := \hat{w}_h(V_{\hat{w}_{h+1}})$ for short. Define an approximate Bellman optimality operator $\hat{T}^{(h)} : \mathcal{X} \to \mathcal{X}$ as:
\[
[\hat{T}^{(h)} V](x) := \max_a \left[ r(x,a) + \phi(x,a)^T \hat{w}_h(V) \right]. \tag{B.14}
\]
Note this $\hat{T}^{(h)}$ is a randomized operator that only depends data from $h$th fold. The Lasso fitted-Q-iteration in learning phase of Algorithm 1 is equivalent to the following approximate value iteration:

$$\hat{T}^{(h)} \Pi_{[0,H]} V_{\bar{\omega}_{k+1}}(x) = \max_a \left[ r(x,a) + \phi(x,a) \right] = \max_a Q_{\bar{\omega}_{k}}(x,a) = V_{\bar{\omega}_{k}}(x).$$ \hspace{1cm} (B.15)

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

$$[TV](x) := \max_a \left[ r(x,a) + \sum_{x'} P(x'|x,a) V(x') \right].$$ \hspace{1cm} (B.16)

**Step 2.** We verify that the true Bellman operator on $\Pi_{[0,H]} V_{\bar{\omega}_{k+1}}$ 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

$$P(x'|x,a) = \sum_{k \in \mathcal{K}} \phi_k(x,a) \psi_k(x'),$$ \hspace{1cm} (B.17)

where $\mathcal{K} \subseteq [d]$ and $|\mathcal{K}| \leq s$. For a vector $\bar{\omega}_{k} \in \mathbb{R}^d$, we define its $k$th coordinate as

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

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

$$[T \Pi_{[0,H]} V_{\bar{\omega}_{k+1}}](x) = \max_a \left[ r(x,a) + \sum_{x'} P(x'|x,a) \Pi_{[0,H]} V_{\bar{\omega}_{k+1}}(x') \right]$$

$$= \max_a \left[ r(x,a) + \sum_{x'} \phi(x,a) \sum_{k \in \mathcal{K}} \phi_k(x,a) \psi_k(x') \Pi_{[0,H]} V_{\bar{\omega}_{k+1}}(x') \right]$$

$$= \max_a \left[ r(x,a) + \sum_{k \in \mathcal{K}} \phi_k(x,a) \sum_{x'} \psi_k(x') \Pi_{[0,H]} V_{\bar{\omega}_{k+1}}(x') \right]$$

$$= \max_a \left[ r(x,a) + \phi(x,a) \Pi_{[0,H]} V_{\bar{\omega}_{k+1}}(x') \right].$$ \hspace{1cm} (B.19)

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

$$\Pi_{[0,H]} V_{\bar{\omega}_{k+1}}(x'_i) = \phi(x_i, a_i) \top \bar{\omega}_{k} + \epsilon_i, i = 1 \ldots, RH,$$ \hspace{1cm} (B.20)

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

$$\mathbb{E}[\epsilon_i | \mathcal{F}_i] = \mathbb{E}[\Pi_{[0,H]} V_{\bar{\omega}_{k+1}}(x'_i) | \mathcal{F}_i] - \phi(x_i, a_i) \top \bar{\omega}_{k}$$

$$= \sum_{x'} [\Pi_{[0,H]} V_{\bar{\omega}_{k+1}}(x')] P(x'|x_i, a_i) - \phi(x_i, a_i) \top \bar{\omega}_{k}$$

$$= \sum_{k \in \mathcal{K}} \phi_k(x_i, a_i) \sum_{x'} [\Pi_{[0,H]} V_{\bar{\omega}_{k+1}}(x')] \psi_k(x') - \phi(x_i, a_i) \top \bar{\omega}_{k} = 0.$$

Therefore, $\{\epsilon_i\}_{i=1}^{RH}$ is a sequence of martingale difference noises and $|\epsilon_i| \leq H$ due to the truncation operator $\Pi_{[0,H]}$. The next lemma bounds the difference between $\bar{\omega}_{k}$ and $\bar{\omega}_{k}$ within $\ell_1$-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

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

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 $1 - \delta$,

$$\|\bar{\omega}_{h} - \bar{\omega}_{k}\|_1 \leq \frac{16 \sqrt{2} s}{C_{\min}(\Sigma \pi, s)} H \sqrt{\frac{\log(2d/\delta)}{RH}}.$$ \hspace{1cm} (B.21)
Step 3. We start to bound $\|V_{\hat{\omega}_h} - V^*_h\|_\infty$ for each step $h$. By the approximate value iteration form Eq. (B.15) and the definition of optimal value function,

$$
\|V_{\hat{\omega}_h} - V^*_h\|_\infty = \|\hat{T}(h)\Pi_{[0,H]}V_{\hat{\omega}_{h+1}} - TV^*_h\|_\infty
= \|\hat{T}(h)\Pi_{[0,H]}V_{\hat{\omega}_{h+1}} - TP\Pi_{[0,H]}V_{\hat{\omega}_{h+1}}\|_\infty + \|TP\Pi_{[0,H]}V_{\hat{\omega}_{h+1}} - TV^*_h\|_\infty. 
$$

(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 $x \in \mathcal{X}$,

$$
\|\hat{T}(h)\Pi_{[0,H]}V_{\hat{\omega}_{h+1}}(x) - TP\Pi_{[0,H]}V_{\hat{\omega}_{h+1}}(x)\|
\leq \max_a \left[ r(x,a) + \phi(x,a)^\top \hat{\omega}_h \right] - \max_a \left[ r(x,a) + \phi(x,a)^\top \hat{\omega}_h \right]
\leq \max_a \left| \phi(x,a)^\top (\hat{\omega}_h - \hat{\omega}_h) \right|
\leq \max_{a,x} \left| \phi(x,a) \right| \|\hat{\omega}_h - \hat{\omega}_h\|_1. 
$$

(B.23)

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

$$
\|\hat{\omega}_h - \hat{\omega}_h\|_1 \leq \frac{16\sqrt{2s}}{C_{\min}(\Sigma^\pi, s)} H \sqrt{\frac{\log(2d/\delta)}{RH}}, 
$$

(B.24)

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

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 $h \in [H]$,

$$
\|\hat{T}(h)\Pi_{[0,H]}V_{\hat{\omega}_{h+1}} - TP\Pi_{[0,H]}V_{\hat{\omega}_{h+1}}\|_\infty \leq \frac{16\sqrt{2s}}{C_{\min}(\Sigma^\pi, s)} H \sqrt{\frac{\log(2dH/\delta)}{RH}}, 
$$

(B.25)

holds with probability at least $1 - \delta$. To bound the second term in Eq. (B.22), we observe that

$$
\|TP\Pi_{[0,H]}V_{\hat{\omega}_{h+1}} - TV^*_h\|_\infty = \max_x \left| TP\Pi_{[0,H]}V_{\hat{\omega}_{h+1}}(x) - TV^*_h(x) \right|
\leq \max_x \max_{x'} \left| \sum_{x'} P(x'|x,a)\Pi_{[0,H]}V_{\hat{\omega}_{h+1}}(x') - \sum_{x'} P(x'|x,a)\Pi_{[0,H]}V^*_h(x') \right|
\leq \|\Pi_{[0,H]}V_{\hat{\omega}_{h+1}} - V^*_h\|_\infty. 
$$

(B.26)

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

$$
\|V_{\hat{\omega}_h} - V^*_h\|_\infty \leq \frac{16\sqrt{2s}}{C_{\min}(\Sigma^\pi, s)} H \sqrt{\frac{\log(2dH/\delta)}{RH}} + \|\Pi_{[0,H]}V_{\hat{\omega}_{h+1}} - V^*_h\|_\infty, 
$$

(B.27)

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

$$
\|\Pi_{[0,H]}V_{\hat{\omega}_1} - V^*_1\|_\infty \leq \|V_{\hat{\omega}_1} - V^*_1\|_\infty
\leq \frac{16\sqrt{2s}}{C_{\min}(\Sigma^\pi, s)} H \sqrt{\frac{\log(2dH/\delta)}{RH}} + \|\Pi_{[0,H]}V_{\hat{\omega}_2} - V^*_2\|_\infty
\leq \|\Pi_{[0,H]}V_{\hat{\omega}_{H+1}} - V^*_h\|_\infty + H^2 \frac{16\sqrt{2s}}{C_{\min}(\Sigma^\pi, s)} H \sqrt{\frac{\log(2dH/\delta)}{RH}}
= H^2 \frac{16\sqrt{2s}}{C_{\min}(\Sigma^\pi, s)} H \sqrt{\frac{\log(2dH/\delta)}{RH}}, 
$$

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

$$
\|V_{\hat{\omega}_{H+1}} - V^*_1\|_\infty \leq H \|Q_{\hat{\omega}_1} - Q^*_1\|_\infty \leq 2H \|\Pi_{[0,H]}V_{\hat{\omega}_1} - V^*_1\|_\infty. 
$$

(B.28)
Putting the above together, we have with probability at least $1 - \delta$,

$$\|V_{N_1} - V_T\|_\infty \leq \frac{32\sqrt{2\delta H^3}}{C_{\min}(\Sigma_\pi, s)} \sqrt{\frac{\log(2dH/\delta)}{N_1}},$$

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

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

for some sufficiently large constant $C_1$. This ends the proof. \hfill \qed

### B.5 Proof of Lemma B.1

**Proof.** Denote the empirical covariance matrix induced by the exploratory policy $\pi_e$ and feature map $\phi$ as

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

Recall that $\Sigma_{\pi_e}$ is the population covariance matrix induced by the exploratory policy $\pi_e$ defined in Eq. (3.1) and feature map $\phi$ 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

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

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

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

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

$$\mathbb{P}\left(\max_{i, j} \left|\sum_{r=1}^{R} v_{ij}^r\right| \geq \delta\right) \leq 3d^2 \exp\left(-\frac{C_0\delta^2}{R}\right)$$

$$\Rightarrow \mathbb{P}\left(\|\hat{\Sigma}_{\pi_e} - \Sigma_{\pi_e}\|_\infty \geq \delta\right) \leq 3d^2 \exp\left(-\frac{C_0\delta^2}{R}\right).$$

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

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

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 $1 - \delta$,

$$\|\hat{\Sigma}_{\pi_e} - \Sigma_{\pi_e}\|_\infty \leq \frac{C_{\min}(\Sigma_{\pi_e}, s)}{32s},$$

Next lemma shows that if the restricted eigenvalue condition holds for one positive semi-definite matrix $\Sigma_0$, then it holds with high probability for another positive semi-definite matrix $\Sigma_1$ as long as $\Sigma_0$ and $\Sigma_1$ are close enough in terms of entry-wise max norm.

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

Applying Lemma B.2 with $\hat{\Sigma}^{\pi_\epsilon}$ and $\Sigma^{\pi_\epsilon}$, we have the restricted eigenvalue of $\hat{\Sigma}^{\pi_\epsilon}$ satisfies $C_{\min}(\hat{\Sigma}^{\pi_\epsilon}, s) > C_{\min}(\Sigma^{\pi_\epsilon}, s)/2$ 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,

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

Denote event $\mathcal{E}$ as

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

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

$$
\left\| \hat{\omega}_h - \bar{\omega}_h \right\|_1 \leq \frac{16\sqrt{2} \lambda_1}{C_{\min}(\Sigma^{\pi_\epsilon}, s)},
$$

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

\[\square\]

## Supporting lemmas

### Lemma C.1 (Pinsker’s inequality).

Denote $x = \{x_1, \ldots, x_T\} \in \mathcal{X}^T$ as the observed states from step 1 to $T$. Then for any two distributions $P_1$ and $P_2$ over $\mathcal{X}^T$ and any bounded function $f : \mathcal{X}^T \to [0, B]$, we have

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

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 $P$ and $\bar{P}$ be two probability measures on the same measurable space $(\Omega, \mathcal{F})$. Then for any event $D \in \mathcal{F}$,

$$
P(D) + \bar{P}(D^c) \geq \frac{1}{2} \exp \left( -\text{KL}(P, \bar{P}) \right),
$$

(C.1)

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

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

### Lemma C.3 (Divergence decomposition).

Let $P$ and $\bar{P}$ be two probability measures on the sequence $(A_1, Y_1, \ldots, A_n, Y_n)$ for a fixed bandit policy $\pi$ interacting with a linear contextual bandit with standard Gaussian noise and parameters $\theta$ and $\hat{\theta}$ respectively. Then the KL divergence of $P$ and $\bar{P}$ can be computed exactly and is given by

$$
\text{KL}(P, \bar{P}) = \frac{1}{2} \sum_{x \in \mathcal{X}} \mathbb{E}[T_x(n)] \langle x, \theta - \hat{\theta} \rangle^2,
$$

(C.2)

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

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

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