The appendix is organized as follows.

- Appendix A. Proof of Proposition 3.2.
- Appendix B. Proofs of results about off-policy evaluation in Section 4.
  - Appendix B.1. Proof of Theorem 4.3.
  - Appendix B.2. Proof of Lemma 4.6.
  - Appendix B.3. Proof of Theorem 4.9.
- Appendix C. Proof of Theorem 5.1: upper bound of batch policy optimization.
- Appendix D. Proof of Theorem 5.2: minimax lower bound of batch policy optimiztion.
- Appendix E. Proofs of auxiliary lemmas.
- Appendix F. Some supporting lemmas.

# A. Proof of Proposition 3.2

Proof of Proposition 3.2. Fix a policy  $\pi$ . Since M is a linear MDP, for every  $i \in [d]$  there exist  $w_i \in \mathbb{R}^d$  such that  $P\phi_i^{\pi} = g_{w_i}$ . Thus, for any  $(x, a) \in \mathcal{X} \times \mathcal{A}$ ,

$$\mathbb{E}_{x' \sim P(\cdot|x,a)}[\phi^{\pi}(x')^{\top}|x,a] = (P\phi_{1}^{\pi}, \dots, P\phi_{d}^{\pi}) = (\phi(x,a)^{\top}w_{1}, \dots, \phi(x,a)^{\top}w_{d})$$
$$= \phi(x,a)^{\top}(w_{1}, \dots, w_{d}).$$

Thus, Eq. (3.1) holds if we choose

$$K^{\pi} = (w_1, \cdots, w_d)$$
.

Under the sparsity assumption, Assumption 2.1, there exists  $\mathcal{K} \subset [d]$  such that  $w_{ij} = 0$  when  $j \notin \mathcal{K}$ . This shows that all but  $|\mathcal{K}|$  rows of  $K^{\pi}$  are identically zero, finishing the proof.

# **B.** Proofs of Off-Policy Evaluation

### B.1. Proof of Theorem 4.3: Lasso fitted Q-evaluation

Recall that we split the whole dataset  $\mathcal{D}$  into T folds and each fold consists of R episodes or RL sample transitions. At tth phase, only the fresh fold of dataset  $\mathcal{D}_t = \{(x_i^{(t)}, a_i^{(t)}, x_i^{(t')})\}_{i=1}^{RL}$  is used.

Step 1: Approximate value iteration. We first show that the execution of Algorithm 1 is equivalent to approximate value iteration. Denote a Lasso estimator with respect to a function V at tth phase:

$$\widehat{w}_t(V) = \operatorname*{argmin}_{w \in \mathbb{R}^d} \left( \frac{1}{RL} \sum_{i=1}^{RL} \left( V(x_i^{(t)'}) - \phi(x_i^{(t)}, a_i^{(t)})^\top w \right)^2 + \lambda_1 \|w\|_1 \right).$$
(B.1)

Note that  $\widehat{w}_t(\cdot)$  only depends data collected at the *t*th phase. Define the parameterized value function as

$$V_w^{\pi}(x) = \sum_a \pi(a|x) \big( r(x,a) + \gamma \phi(x,a)^{\top} w \big).$$

Define an approximate Bellman operator for target policy  $\pi$ , i.e.  $\widehat{\mathcal{T}}_{\pi}^{(t)} : \mathbb{R}^{\mathcal{X}} \to \mathbb{R}^{\mathcal{X}}$  as:

$$[\widehat{\mathcal{T}}_{\pi}^{(t)}V](x) := V_{\widehat{w}_{t}(V)}^{\pi}(x) = \sum_{a} \pi(a|x) \Big( r(x,a) + \gamma \phi(x,a)^{\top} \widehat{w}_{t}(V) \Big).$$
(B.2)

Note this  $\hat{\mathcal{T}}_{\pi}^{(t)}$  is a randomized operator that only depends data in the *t*th fold. It is easy to see that if  $(\hat{w}_t)_{t=1}^T$  is the sequence of weights computed in Algorithm 1 then  $\hat{w}_t = \hat{w}_t(\prod_{[0,1/(1-\gamma)]} V_{\hat{w}_{t-1}}^{\pi})$  and also

$$V_{\widehat{w}_t}^{\pi} = \widehat{\mathcal{T}}_{\pi}^{(t)} \Pi_{[0,1/(1-\gamma)]} V_{\widehat{w}_{t-1}}^{\pi} .$$
(B.3)

Step 2: Linear representation of Bellman operator. Recall that the true Bellman operator for target policy  $\pi$ , i.e.  $\mathcal{T}_{\pi} : \mathbb{R}^{\mathcal{X}} \to \mathbb{R}^{\mathcal{X}}$  is defined as

$$[\mathcal{T}_{\pi}V](x) := \sum_{a} \pi(a|x) \Big( r(x,a) + \gamma \sum_{x'} P(x'|x,a)V(x') \Big).$$
(B.4)

We first verify for each phase  $t \in [T]$ ,  $\mathcal{T}_{\pi} \Pi_{[0,1/(1-\gamma)]} V_{\widehat{w}_t}^{\pi}$  has a linear representation. From Assumption 2.1, there exists a vector  $\overline{w}_t \in \mathbb{R}^d$  with  $\overline{w}_{t,k} = 0$  if  $k \notin \mathcal{K}$  such that

$$[\mathcal{T}_{\pi}\Pi_{[0,1/(1-\gamma)]}V_{\widehat{w}_{t}}^{\pi}](x) = \sum_{a} \pi(a|x) \Big( r(x,a) + \gamma \phi(x,a)^{\top} \overline{w}_{t} \Big).$$
(B.5)

It shows that  $\mathcal{T}_{\pi}\Pi_{[0,1/(1-\gamma)]}V_{\widehat{w}_t}^{\pi}$  has a linear representation if the reward could also be linearly represented. For notation simplicity, we drop the supscript of  $x_i^{(t)}$  and  $a_i^{(t)}$  for the following derivations when there is no ambiguity.

Step 3: Sparse linear regression. We interpret  $\bar{w}_t$  as the ground truth of the lasso estimator in Algorithm 1 at phase t, in terms of the following sparse linear regression:

$$\Pi_{[0,1/(1-\gamma)]} V_{\widehat{w}_{t-1}}^{\pi}(x_i') = \phi(x_i, a_i)^{\top} \bar{w}_t + \varepsilon_i, i = 1..., RL,$$
(B.6)

where  $\varepsilon_i = \prod_{[0,1/(1-\gamma)]} V_{\widehat{w}_{t-1}}^{\pi}(x'_i) - \phi(x_i, a_i)^{\top} \overline{w}_t$ . Define a filtration  $\{\mathcal{F}_i\}_{i=1,...,RL}$  with  $\mathcal{F}_i$  generated by  $\{(x_1, a_1), \ldots, (x_i, a_i)\}$ . By the definition of  $V_{\widehat{w}_{t-1}}$ , we have

$$\mathbb{E}[\varepsilon_i|\mathcal{F}_i] = \mathbb{E}\left[\Pi_{[0,1/(1-\gamma)]}V_{\widehat{w}_{t-1}}^{\pi}(x_i')|\mathcal{F}_i\right] - \phi(x_i, a_i)^{\top} \overline{w}_t$$
  
$$= \sum_{x'} [\Pi_{[0,1/(1-\gamma)]}V_{\widehat{w}_{t-1}}^{\pi}](x')P(x'|x_i, a_i)' - \phi(x_i, a_i)^{\top} \overline{w}_t$$
  
$$= \sum_{k \in \mathcal{K}} \phi_k(x_i, a_i) \sum_{x'} [\Pi_{[0,1/(1-\gamma)]}V_{\widehat{w}_{t-1}}^{\pi}](x')\psi_k(x')' - \phi(x_i, a_i)^{\top} \overline{w}_t = 0.$$

Therefore,  $\{\varepsilon_i\}_{n=1}^{RL}$  is a sequence of martingale difference noises and  $|\varepsilon_i| \leq 1/(1-\gamma)$  due to the truncation operator  $\Pi_{[0,1/(1-\gamma)]}$ . The next lemma bounds the difference between  $\hat{w}_t$  and  $\bar{w}_t$  within  $\ell_1$ -norm. The proof is deferred to Appendix E.2.

**Lemma B.1.** Consider the sparse linear regression described in Eq. (B.6). Suppose the restricted minimum eigenvalue of  $\Sigma$  satisfy  $C_{\min}(\Sigma, s) > 0$  and the number of episodes used in phase t satisfies

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

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

$$\|\widehat{w}_t - \overline{w}_t\|_1 \le \frac{16\sqrt{2}s}{C_{\min}(\Sigma, s)} \frac{1}{1 - \gamma} \sqrt{\frac{\log(2d/\delta)}{RL}}.$$
 (B.7)

Note that the samples we use between phases are mutually independent. Thus, Eq. (B.7) uniformly holds for all  $t \in [T]$  with probability at least  $1 - T\delta$ .

**Step 4: Error decomposition.** Recall that  $\hat{v}_{\hat{w}_T}^{\pi} = \frac{1}{m} \sum_{u=1}^m \prod_{[0,1/(1-\gamma)]} (Q_{\hat{w}_T}(\tilde{x}_u, \tilde{a}_u))$  and we denote  $\bar{v}_{\hat{w}_T}^{\pi} = \sum_x V_{\hat{w}_T}^{\pi}(x)\xi_0(x)$ . According to Eq. (B.3), we decompose the policy evaluation error by Monte Carlo error, estimation error

and approximation error as follows:

Since  $\tilde{x}_u, \tilde{a}_u$  is i.i.d sampled from  $\xi_0$  and  $\pi$ , standard Hoeffding's inequality shows that

$$\left|\widehat{v}_{\widehat{w}_{T}}^{\pi} - \overline{v}_{\widehat{w}_{T}}^{\pi}\right| \leq \sqrt{\frac{\log(1/\delta)}{m(1-\gamma)^{2}}} = \sqrt{\frac{\log(1/\delta)}{N(1-\gamma)^{2}}},\tag{B.9}$$

with probability at least  $1 - \delta$ .

Recall that  $\nu_t^{\pi} = \mathbb{E}^{\pi}[\phi(x_t, a_t)|x_0 \sim \xi_0]$ . To bound the estimation error, combining Eqs. (B.2) and (B.5) together, we have

$$\sum_{x} \left( [\widehat{\mathcal{T}}_{\pi}^{(T)} \Pi_{[0,1/(1-\gamma)]} V_{\widehat{w}_{T-1}}^{\pi}](x) - [\mathcal{T}_{\pi} \Pi_{[0,1/(1-\gamma)]} V_{\widehat{w}_{T-1}}^{\pi}](x) \right) \xi_{0}(x)$$
  
= $\gamma \sum_{x} \sum_{a} \pi(a|x) \phi(x,a)^{\top} (\widehat{w}_{T-1} - \bar{w}_{T-1}) \xi_{0}(x)$   
= $\gamma (\nu_{0}^{\pi})^{\top} (\widehat{w}_{T-1} - \bar{w}_{T-1}).$  (B.10)

To bound approximation error, we expand it by Eq. (B.4):

$$\sum_{x} \left[ \mathcal{T}_{\pi} (\Pi_{[0,1/(1-\gamma)]} V_{\widehat{w}_{T-1}}^{\pi} - v^{\pi}) \right] (x) \xi_{0}(x)$$
  
= $\gamma \sum_{x} \left( \sum_{a} \pi(a|x) \sum_{x'} P(x'|x,a) (\Pi_{[0,1/(1-\gamma)]} V_{\widehat{w}_{T-1}}^{\pi} - v^{\pi}) (x') \right) \xi_{0}(x).$ 

According to Eq. (B.10), we decompose

$$\begin{aligned} & \left(\Pi_{[0,1/(1-\gamma)]}V_{\widehat{w}_{T-1}}^{\pi} - v^{\pi}\right)(x) \leq \left(V_{\widehat{w}_{T-1}}^{\pi} - v^{\pi}\right)(x) \\ &= \left[\widehat{\mathcal{T}}_{\pi}^{(T-1)}\Pi_{[0,1/(1-\gamma)]}V_{\widehat{w}_{T-2}}^{\pi} - \mathcal{T}_{\pi}\Pi_{[0,1/(1-\gamma)]}V_{\widehat{w}_{T-2}}^{\pi} + \mathcal{T}_{\pi}\Pi_{[0,1/(1-\gamma)]}V_{\widehat{w}_{T-2}}^{\pi} - \mathcal{T}_{\pi}v^{\pi}\right](x) \\ &= \gamma \sum_{a} \pi(a|x)\phi(x,a)^{\top}(\widehat{w}_{T-2} - \overline{w}_{T-2}) + \left[\mathcal{T}_{\pi}(\Pi_{[0,1/(1-\gamma)]}V_{\widehat{w}_{T-2}}^{\pi} - v^{\pi})\right](x). \end{aligned}$$

This implies

$$\begin{split} &\sum_{x} \left[ \mathcal{T}_{\pi} (\Pi_{[0,1/(1-\gamma)]} V_{\widehat{w}_{T-1}}^{\pi} - v^{\pi}) \right] (x) \xi_{0}(x) \\ &\leq \gamma^{2} \sum_{x} \left( \sum_{a} \pi(a|x) \gamma \sum_{x'} P(x'|x,a) \left( \sum_{a} \pi(a|x') \phi(x',a)^{\top} (\widehat{w}_{T-2} - \overline{w}_{T-2}) \right) \right) \xi_{0}(x) \\ &+ \sum_{x} \left( \sum_{a} \pi(a|x) \gamma \sum_{x'} \left[ \mathcal{T}_{\pi} (\Pi_{[0,1/(1-\gamma)]} V_{\widehat{w}_{T-2}}^{\pi} - v^{\pi}) \right] (x') P(x'|x,a) \right) \xi_{0}(x) \end{split}$$
(B.11)  
$$&= \gamma^{2} \mathbb{E}^{\pi} [\phi(x_{1},a_{1})|x \sim \xi_{0}]^{\top} (\widehat{w}_{T-2} - \overline{w}_{T-2}) \\ &+ \gamma \mathbb{E}^{\pi} \left[ \left[ \mathcal{T}_{\pi} (\Pi_{[0,1/(1-\gamma)]} V_{\widehat{w}_{T-2}}^{\pi} - v^{\pi}) \right] (x_{1}) |x_{0} \sim \xi_{0} \right] . \end{split}$$

Combining Eqs. (B.8), (B.10) and (B.11) together, we have

$$\begin{aligned} |\bar{v}_{\widehat{w}_{T}}^{\pi} - v^{\pi}| \leq & |\gamma(\nu_{0}^{\pi})^{\top} (\widehat{w}_{T-1} - \bar{w}_{T-1})| + |\gamma^{2} (\nu_{1}^{\pi})^{\top} (\widehat{w}_{T-2} - \bar{w}_{T-2})| \\ & + \left| \gamma \mathbb{E}^{\pi} \Big[ [\mathcal{T}_{\pi} (\Pi_{[0,1/(1-\gamma)]} V_{\widehat{w}_{T-2}}^{\pi} - v^{\pi})](x_{1}) | x_{0} \sim \xi_{0} \Big] \Big|. \end{aligned}$$

Iteratively implementing the above decomposition, we have

$$\begin{aligned} |\bar{v}_{\hat{w}_{T}}^{\pi} - v^{\pi}| &\leq \sum_{t=0}^{T-1} \gamma^{t+1} |(\nu_{t}^{\pi})^{\top} (\widehat{w}_{T-t-1} - \bar{w}_{T-t-1})| + \gamma^{T} \Big| \mathbb{E}^{\pi} \Big[ (\Pi_{[0,1/(1-\gamma)]} V_{\hat{w}_{0}}^{\pi} - v^{\pi}) (x_{T}) |x_{0} \sim \xi_{0} \Big] \Big| \\ &\leq \sum_{t=0}^{T-1} \gamma^{t+1} |(\nu_{t}^{\pi})^{\top} (\widehat{w}_{T-t-1} - \bar{w}_{T-t-1})| + \frac{2\gamma^{T}}{1-\gamma} \\ &\leq \sum_{t=0}^{T-1} \gamma^{t+1} ||\nu_{t}^{\pi}) ||_{\infty} ||\widehat{w}_{T-t-1} - \bar{w}_{T-t-1}||_{1} + \frac{2\gamma^{T}}{1-\gamma}. \end{aligned}$$

Since we assume  $\|\phi(x,a)\|_{\infty} \leq 1$ , then  $\|\nu_t^{\pi}\|_{\infty} \leq 1$  as well. Using the fact that  $\sum_{t=0}^{T-1} \gamma^t \leq 1/(1-\gamma)$ , we have

$$|\bar{v}_{\widehat{w}_T}^{\pi} - v^{\pi}| \le \frac{1}{1 - \gamma} \max_{t=0,\dots,T-1} \|\widehat{w}_t - \bar{w}_t\|_1 + \frac{2\gamma^T}{1 - \gamma}.$$

Suppose the sample size satisfies

$$N \ge \frac{C_1 \log(3d^2/\delta)s^2}{C_{\min}(\Sigma, s)} \frac{L}{1-\gamma} \log(N/(1-\gamma))),$$

for a sufficient large constant  $C_1 > 0$ . Applying Lemma B.1 over  $t = 0, \ldots, T - 1$ , it implies

$$|\bar{v}_{\widehat{w}_{T}}^{\pi} - v^{\pi}| \leq \frac{1}{(1-\gamma)^{2}} \frac{16\sqrt{2}s}{C_{\min}(\Sigma, s)} \sqrt{\frac{\log(2d/\delta)}{RL}} + \frac{2\gamma^{T}}{1-\gamma}$$

holds with probability at least  $1 - T\delta$ . By elementary change of base formula and Taylor expansion, we have

$$\log_{1/\gamma}(x) = \frac{\log(x)}{\log(1/\gamma)} \approx \frac{\log(x)\lambda}{1-\gamma}$$

By properly choosing  $T = \Theta(\log(N/(1-\gamma))/(1-\gamma))$ , we have with probability at least  $1-\delta$ ,

$$|\bar{v}_{\widehat{w}_{T}}^{\pi} - v^{\pi}| \le \frac{1}{(1-\gamma)^{5/2}} \frac{32\sqrt{2}s}{C_{\min}(\Sigma, s)} \sqrt{\frac{\log(N/(1-\gamma))\log(2dT/\delta)}{N}},$$

where we use N = TRL. Combining with Monte Carlo approximation error Eq. (B.9) This ends the proof.

## **B.2.** Proof of Lemma 4.6: Feature selection

We study the feature screening and sparsity properties of the model selected by the regularized estimator  $\hat{K}^{\pi}$ . Recall that from the identity Eq. (3.1), we solve the following multivariate regression problem:

$$\phi^{\pi}(x_n')^{\top} = \phi(x_n, a_n)^{\top} K^{\pi} + \varepsilon_n, \ n \in [N],$$
(B.12)

where  $x'_n \sim P(\cdot|x_n, a_n)$  and  $\varepsilon_n = \phi^{\pi}(x'_n)^{\top} - \mathbb{E}[\phi^{\pi}(x'_n)^{\top}] \in \mathbb{R}^d$  is the noise vector. Define a filtration  $\{\mathcal{F}_n\}_{n=1,...,N}$  with  $\mathcal{F}_n$  generated by  $\{(x_1, a_1), \ldots, (x_n, a_n)\}$ . It is easy to see  $\mathbb{E}[\varepsilon_n|\mathcal{F}_n] = 0$  such that  $\{\varepsilon_n\}_{n=1}^N$  are martingale difference vectors. We introduce some notations for simplicities:

- Denote the response  $Y_j = (\phi_j^{\pi}(x_1'), \dots, \phi_j^{\pi}(x_N'))^{\top}$  for  $j \in [d]$  where  $\phi_j^{\pi}(\cdot)$  is *j*th coordinate of  $\phi^{\pi}(\cdot)$ . And  $Y = (Y_1^{\top}, \dots, Y_d^{\top})^{\top} \in \mathbb{R}^{Nd \times 1}$ .
- Denote the noise  $W_j = (\varepsilon_{1j}, \dots, \varepsilon_{Nj})^\top$  where  $\varepsilon_{nj}$  is the *j*th coordinate of  $\varepsilon_n$ , and  $W = (W_1^\top, \dots, W_d^\top)^\top \in \mathbb{R}^{Nd \times 1}$ .
- Denote the design matrix as

$$\Phi = \begin{pmatrix} \phi(x_1, a_1)^\top \\ \vdots \\ \phi(x_N, a_N)^\top \end{pmatrix} \in \mathbb{R}^{N \times d}; X = \begin{pmatrix} \Phi & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \Phi \end{pmatrix} \in \mathbb{R}^{Nd \times d^2}.$$

Note that X is a block diagonal matrix.

- Let  $\beta_j^*$  as the *j*th column of  $K^{\pi}$  and  $\beta^* = (\beta_1^{*\top}, \dots, \beta_d^{*\top})^{\top} \in \mathbb{R}^{d^2 \times 1}$  as the regression coefficient.
- For every β ∈ ℝ<sup>d<sup>2</sup></sup>, we define β<sup>j</sup> = (β<sub>j+(i-1)d</sub> : i ∈ [d])<sup>T</sup> as the vector formed by the coefficients corresponding to the *j*th variable. For instance, β<sup>\*j</sup> is the *j*th row of K<sup>π</sup>. If J ⊆ [d], denote β<sub>J</sub> ∈ ℝ<sup>d<sup>2</sup></sup> by stacking the vectors β<sup>j</sup> 1{j ∈ J}. Write S(β) = {j : β<sup>j</sup> ≠ 0} as the relevant feature set of β.
- For a vector  $\boldsymbol{\beta} \in \mathbb{R}^{d^2}$ , define the  $\ell_{2,p}$ -norm for  $1 \leq p < \infty$  as:

$$\|\boldsymbol{\beta}\|_{2,p} = \left(\sum_{j=1}^{d} \left(\sum_{i=1}^{d} \beta_{j+(i-1)d}^{2}\right)^{p/2}\right)^{1/p},$$

and the  $\ell_{2,0}$ -norm as:

$$\|\boldsymbol{\beta}\|_{2,0} = \sum_{j=1}^{d} \mathbb{1} \left\{ \|\boldsymbol{\beta}^{j}\|_{2} \neq 0 \right\}$$

Therefore, we can rewrite Eq. (B.12) into an ordinary linear regression form with group sparse structure on the regression coefficients  $\beta^*$ :

$$Y = X\beta^* + W.$$

Note that  $S(\beta^*) = K$  where K is defined in Assumption 2.1 since  $K^{\pi}$  is row-sparse. The corresponding group lasso estimator defined in Eq. (3.2) can be rewritten into:

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \frac{1}{Nd} \left\| Y - X\boldsymbol{\beta} \right\|_{2}^{2} + \lambda_{2} \sum_{j=1}^{d} \|\boldsymbol{\beta}^{j}\|_{2} \right\},$$
(B.13)

and  $\mathcal{S}(\widehat{\boldsymbol{\beta}}) = \widehat{\mathcal{K}}$ . The regularization parameter is chosen as

$$\lambda_2 = 4\sqrt{\frac{2\log(2d^2/\delta)}{Nd}},\tag{B.14}$$

for some  $\delta > 0$ .

Now we study the feature screening property of  $\hat{\beta}$  in four steps.

**Step 1.** By the optimality of  $\hat{\beta}$ , we have

$$\frac{1}{Nd} \|Y - X\widehat{\beta}\|_{2}^{2} + \lambda_{2} \sum_{j=1}^{d} \|\widehat{\beta}^{j}\|_{2} \le \frac{1}{Nd} \|Y - X\beta^{*}\|_{2}^{2} + \lambda_{2} \sum_{j=1}^{d} \|\beta^{*j}\|_{2}.$$

Plugging in  $Y = X\beta^* + W$ ,

$$\frac{1}{Nd} \|X(\widehat{\beta} - \beta^{*})\|_{2}^{2} \leq \frac{2}{Nd} W^{\top} X(\widehat{\beta} - \beta^{*}) + \lambda_{2} \sum_{j=1}^{d} \left( \|\beta^{*j}\|_{2} - \|\widehat{\beta}^{j}\|_{2} \right) \\
\leq \frac{2}{Nd} \max_{j \in [d]} \|(X^{\top} W)^{j}\|_{2} \sum_{j=1}^{d} \|\widehat{\beta}^{j} - \beta^{*j}\|_{2} + \lambda_{2} \sum_{j=1}^{d} \left( \|\beta^{*j}\|_{2} - \|\widehat{\beta}^{j}\|_{2} \right) \tag{B.15}$$

where the last inequality is from Hölder's inequality.

Step 2. Next, we will bound the noise term:  $\|(X^{\top}W)^{j}\|_{2}$ . From the definitions of X and W, we write it explicitly as

$$||(X^{\top}W)^{j}||_{2} = \sqrt{\sum_{i=1}^{d} \left(\sum_{n=1}^{N} \phi_{j}(x_{n}, a_{n})\varepsilon_{ni}\right)^{2}}.$$

It is easy to verify that  $\{\phi_j(x_n, a_n)\varepsilon_{ni}\}_{n=1}^N$  is also a martingale difference sequence for any  $i, j \in [d]$  and  $|\phi_j(x_n, a_n)\varepsilon_{ni}| \le 1$  since we assume  $\|\phi(x, a)\|_{\infty} \le 1$  for any state-action pair. According to Azuma-Hoeffding inequality (Lemma F.2), for all  $\tilde{\delta} > 0$ ,

$$\mathbb{P}\Big(\Big|\sum_{n=1}^{N}\phi_j(x_n,a_n)\varepsilon_{ni}\Big| \ge \widetilde{\delta}\Big) \le 2\exp\Big(-\frac{\widetilde{\delta}^2}{2N}\Big).$$

Using the union bound twice, the following holds,

$$\mathbb{P}\Big(\max_{j\in[d]}\sqrt{\sum_{i=1}^{d}\Big(\sum_{n=1}^{N}\phi_{j}(x_{n},a_{n})\varepsilon_{ni}\Big)^{2}} \ge \sqrt{d\widetilde{\delta}}\Big) \le 2d^{2}\exp\Big(-\frac{\widetilde{\delta}^{2}}{2N}\Big).$$
(B.16)

Letting  $\delta = 2d^2 \exp(-\widetilde{\delta}^2/2N)$ , we have with probability at least  $1 - \delta$ ,

$$\frac{1}{Nd} \max_{j \in [d]} \left\| (X^\top W)^j \right\|_2 \le \frac{\lambda_2}{4}.$$

Define an event  $\mathcal{A}$  as

$$\mathcal{A} = \left\{ \frac{1}{Nd} \max_{j \in [d]} \left\| (X^\top W)^j \right\|_2 \le \frac{\lambda_2}{4} \right\}.$$

Then we have  $\mathbb{P}(\mathcal{A}) \geq 1 - \delta$ .

**Step 3.** According to Karush–Kuhn–Tucker (KKT) condition, the solution  $\hat{\beta}$  of the optimization problem Eq. (B.13) satisfies

$$\begin{cases} \left(X^{\top}(Y - X\widehat{\boldsymbol{\beta}})\right)^j / (Nd) = \lambda_2 \widehat{\boldsymbol{\beta}}^j / \|\widehat{\boldsymbol{\beta}}^j\|_2, & \text{if } \widehat{\boldsymbol{\beta}}^j \neq 0, \\ \|\left(X^{\top}(Y - X\widehat{\boldsymbol{\beta}})\right)^j\|_2 / (Nd) \le \lambda_2, & \text{if } \widehat{\boldsymbol{\beta}}^j = 0. \end{cases}$$

Under event  $\mathcal{A}$  and using KKT condition, we have if  $\hat{\beta}^j \neq 0$ , then

$$\begin{split} \lambda_{2} &= \left\| \frac{1}{Nd} \big( X^{\top} (Y - X\widehat{\beta}) \big)^{j} \right\|_{2} = \left\| \frac{1}{Nd} \big( X^{\top} X (\beta^{*} - \widehat{\beta}) \big)^{j} + \frac{1}{Nd} \big( X^{\top} W \big)^{j} \right\|_{2} \\ &\leq \left\| \frac{1}{Nd} \big( X^{\top} X (\beta^{*} - \widehat{\beta}) \big)^{j} \right\|_{2} + \left\| \frac{1}{Nd} \big( X^{\top} W \big)^{j} \right\|_{2} \\ &\leq \left\| \frac{1}{Nd} \big( X^{\top} X (\beta^{*} - \widehat{\beta}) \big)^{j} \right\|_{2} + \frac{1}{4} \lambda_{2}, \end{split}$$

which implies

$$\frac{1}{(Nd)^2} \left\| \left( X^\top X(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \right)_{\mathcal{S}(\widehat{\boldsymbol{\beta}})} \right\|_2^2 = \frac{1}{(Nd)^2} \sum_{j \in \mathcal{S}(\widehat{\boldsymbol{\beta}})} \left\| \left( X^\top X(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \right)^j \right\|_2^2 \\
\geq |\mathcal{S}(\widehat{\boldsymbol{\beta}})| \frac{9}{16} \lambda_2^2.$$
(B.17)

We define a notation of restricted maximum eigenvalue with respect to  $S(\beta^*)$  and X:

$$\widetilde{C}_{\max}(m) = \max_{\boldsymbol{\beta} \in \mathbb{R}^{d^2 \times d^2} : \|\boldsymbol{\beta}_{\mathcal{S}(\boldsymbol{\beta}^*)^c}\|_{2,0} \le m} \frac{\boldsymbol{\beta}^\top X^\top X \boldsymbol{\beta}}{N \|\boldsymbol{\beta}\|_2^2}.$$
(B.18)

Denote  $\widehat{m} = |\mathcal{S}(\widehat{\beta}) \setminus \mathcal{S}(\beta^*)|$ . Then we have

$$\begin{split} \left\| \left( X^{\top} X(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{*}) \right)_{\mathcal{S}(\widehat{\boldsymbol{\beta}})} \right\|_{2} &\leq \sup_{\|\boldsymbol{\alpha}_{\mathcal{S}(\boldsymbol{\beta}^{*})^{c}}\|_{2,0} \leq \widehat{\boldsymbol{m}}} \frac{|\boldsymbol{\alpha}^{\top} X^{\top} X(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{*})|}{\|\boldsymbol{\alpha}\|_{2}} \\ &\leq \sup_{\|\boldsymbol{\alpha}_{\mathcal{S}(\boldsymbol{\beta}^{*})^{c}}\|_{2,0} \leq \widehat{\boldsymbol{m}}} \frac{\|\boldsymbol{\alpha}^{\top} X^{\top} \|_{2} \|X(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{*})\|_{2}}{\|\boldsymbol{\alpha}\|_{2}} \\ &= \sup_{\|\boldsymbol{\alpha}_{\mathcal{S}(\boldsymbol{\beta}^{*})^{c}}\|_{2,0} \leq \widehat{\boldsymbol{m}}} \frac{\sqrt{|\boldsymbol{\alpha}^{\top} X^{\top} X \boldsymbol{\alpha}|}}{\|\boldsymbol{\alpha}\|_{2}} \|X(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{*})\|_{2} \\ &\leq \sqrt{NC_{\max}(\widehat{\boldsymbol{m}})} \|X(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{*})\|_{2}. \end{split}$$
(B.19)

Combining Eqs. (B.17) and (B.19) together, we have

$$|\mathcal{S}(\widehat{\boldsymbol{\beta}})| \le \frac{16C_{\max}(\widehat{\boldsymbol{m}})}{9Nd^2\lambda_2^2} \|X(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2, \tag{B.20}$$

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

**Step 4.** It remains to control the in-sample prediction error  $||X(\hat{\beta} - \beta^*)||_2^2$ . Under event  $\mathcal{A}$ , using Eq. (B.15) implies

$$\frac{1}{Nd} \left\| X(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \right\|_2^2 \le \frac{\lambda_2}{2} \sum_{j=1}^d \left\| \widehat{\boldsymbol{\beta}}^j - \boldsymbol{\beta}^{*j} \right\|_2 + \lambda_2 \sum_{j=1}^d \left( \| \boldsymbol{\beta}^{*j} \|_2 - \| \widehat{\boldsymbol{\beta}}^j \|_2 \right).$$

Adding  $\sum_{j=1}^{d} \|\widehat{\beta}^{j} - \beta^{*j}\|_{2} \lambda_{2}/2$  to both sides and using the fact that  $\|\beta^{*j} - \beta^{*j}\|_{2} + \|\widehat{\beta}^{j}\|_{2} - \|\widehat{\beta}^{j}\|_{2} = 0$  for  $j \neq S(\beta^{*})$ , we have

$$\frac{1}{Nd} \|X(\widehat{\beta} - \beta^*)\|_2^2 + \frac{\lambda_2}{2} \sum_{j=1}^d \|\widehat{\beta}^j - \beta^{*j}\|_2$$

$$\leq \lambda_2 \sum_{j \in \mathcal{S}(\beta^*)} \left( \|\widehat{\beta}^j - \beta^{*j}\|_2 + \|\beta^{*j}\|_2 - \|\widehat{\beta}^j\|_2 \right)$$

$$\leq 2\lambda_2 \sqrt{s} \|(\widehat{\beta} - \beta^*)_{\mathcal{S}(\beta^*)}\|_2,$$
(B.21)

where the last inequality is from Cauchy-Schwarz inequality. Recall that the expected uncentered covariance matrix is defined as

$$\Sigma = \mathbb{E}\Big[\frac{1}{L}\sum_{h=0}^{L-1}\phi(x_h^{(1)}, a_h^{(1)})\phi(x_h^{(1)}, a_h^{(1)})^{\top}\Big],$$

and we define the empirical uncentered covariance matrix as

$$\widehat{\Sigma} = \frac{1}{N} \sum_{n=1}^{N} \phi(x_n, a_n) \phi(x_n, a_n)^{\top} = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{L} \sum_{h=0}^{L-1} \phi(x_h^{(k)}, a_h^{(k)}) \phi(x_h^{(k)}, a_h^{(k)})^{\top},$$

with N = KL. Denote the expected and empirical uncentered covariance matrices for the multivariate linear regression as

$$\widehat{\Psi} = \begin{pmatrix} \widehat{\Sigma} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \widehat{\Sigma} \end{pmatrix} \in \mathbb{R}^{d^2 \times d^2}; \Psi = \begin{pmatrix} \Sigma & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \Sigma \end{pmatrix} \in \mathbb{R}^{d^2 \times d^2}.$$

We introduce a generalization of restricted eigenvalue condition (Definition 4.2) for multivariate linear regression.

**Definition B.2** ( $\ell_{2,1}$ -restricted eigenvalue). Given a symmetric matrix  $H \in \mathbb{R}^{d^2 \times d^2}$  and integer  $s \ge 1$ , the restricted eigenvalue of H is defined as

$$\widetilde{C}_{\min}(H,s) := \min_{\mathcal{S} \subset [d], |\mathcal{S}| \le s} \min_{\boldsymbol{\beta} \in \mathbb{R}^{d^2}} \Big\{ \frac{\langle \boldsymbol{\beta}, H \boldsymbol{\beta} \rangle}{\|\boldsymbol{\beta}_{\mathcal{S}}\|_2^2} : \boldsymbol{\beta} \in \mathbb{R}^{d^2}, \|\boldsymbol{\beta}_{\mathcal{S}^c}\|_{2,1} \le 3\|\boldsymbol{\beta}_{\mathcal{S}}\|_{2,1} \Big\}$$

Next lemma provides a lower bound for  $\widetilde{C}_{\min}(\widehat{\Phi}, s)$ . The proof is deferred to Appendix E.1.

**Lemma B.3.** Assume the  $\ell_{2,1}$ -restricted eigenvalue of  $\Psi$  satisfies  $\widetilde{C}_{\min}(\Psi, s) > 0$  for some  $\delta > 0$ . Suppose the sample size satisfies  $N \geq 32^2 L \log(3d^2/\delta)s^2/\widetilde{C}_{\min}(\Psi, s)^2$ . Then the  $\ell_{2,1}$ -restricted eigenvalue of  $\widehat{\Psi}$  satisfies  $\widetilde{C}_{\min}(\widehat{\Psi}, s) > \widetilde{C}_{\min}(\Psi, s)/2$  with probability at least  $1 - \delta$ .

On the other hand, from Eq. (B.21), we know that

$$\frac{1}{2}\sum_{j=1}^{d}\left\|\widehat{\boldsymbol{\beta}}^{j}-\boldsymbol{\beta}^{*j}\right\|_{2} \leq 2\sum_{j\in\mathcal{S}(\boldsymbol{\beta}^{*})}\left\|\widehat{\boldsymbol{\beta}}^{j}-\boldsymbol{\beta}^{*j}\right\|_{2},$$

and thus,

$$\sum_{j \in \mathcal{S}(\boldsymbol{\beta}^*)^c} \left\| \widehat{\boldsymbol{\beta}}^j - \boldsymbol{\beta}^{*j} \right\|_2 \le 3 \sum_{j \in \mathcal{S}(\boldsymbol{\beta}^*)} \left\| \widehat{\boldsymbol{\beta}}^j - \boldsymbol{\beta}^{*j} \right\|_2.$$
(B.22)

This implies that  $\|(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{\mathcal{S}(\boldsymbol{\beta}^*)^c}\|_{2,1} \leq 3 \|(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{\mathcal{S}(\boldsymbol{\beta}^*)}\|_{2,1}$ . Applying Lemma B.3, the following holds with probability at least  $1 - \delta$ ,

$$\frac{\left\|X(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^{*})\right\|_{2}^{2}}{N\left\|(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^{*})_{\mathcal{S}(\boldsymbol{\beta}^{*})}\right\|_{2}^{2}} = \frac{(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^{*})^{\top}\widehat{\Psi}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^{*})}{\left\|(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^{*})_{\mathcal{S}(\boldsymbol{\beta}^{*})}\right\|_{2}^{2}} \ge \widetilde{C}_{\min}(\widehat{\Psi},s) \ge \widetilde{C}_{\min}(\Psi,s)/2.$$
(B.23)

Plugging the above bound into Eq. (B.21),

$$\frac{1}{Nd} \| X(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \|_2^2 \le 2\lambda_2 \sqrt{s} \| (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{\mathcal{S}(\boldsymbol{\beta}^*)} \|_2$$
$$\le \frac{1}{\sqrt{N}} \| X(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \|_2 \frac{4\sqrt{s}\lambda_2}{\sqrt{\widetilde{C}_{\min}(\Psi, s)}}.$$

Combining with Eq. (B.20) and the choice of  $\lambda_2$  in Eq. (B.14), we reach

$$|\mathcal{S}(\widehat{\boldsymbol{\beta}})| \leq \frac{16\widetilde{C}_{\max}}{9Nd^2\lambda_2^2} \frac{16\lambda_2^2Nd^2s}{\widetilde{C}_{\min}(\Psi,s)} = \frac{256\widetilde{C}_{\max}(\widehat{m})s}{9\widetilde{C}_{\min}(\Psi,s)},$$

with probability at least  $1 - \delta$ , as long as  $N \ge 32^2 L \log(3d^2/\delta)s^2/\widetilde{C}_{\min}(\Psi, s)^2$ . If the vanilla restricted eigenvalue (Definition 4.2) of  $\Sigma$  satisfies  $C_{\min}(\Sigma, s) > 0$ , then we have for any  $\mathcal{S} \subset [d], |\mathcal{S}| \le s$ , and any  $\beta_j \in \mathbb{R}^d$  satisfying  $\|(\beta_j)_{\mathcal{S}^c}\|_1 \le 3\|(\beta_j)_{\mathcal{S}}\|_1$ ,

$$\frac{\boldsymbol{\beta}_{j}^{\top}\boldsymbol{\Sigma}\boldsymbol{\beta}_{j}}{\|(\boldsymbol{\beta}_{j})_{\mathcal{S}}\|_{2}^{2}} \geq C_{\min}(\boldsymbol{\Sigma},s) > 0.$$

Consider a sequence of vectors  $\beta_1, \ldots, \beta_d$  satisfying  $\|(\beta_j)_{\mathcal{S}^c}\|_1 \leq 3\|(\beta_j)_{\mathcal{S}}\|_1$ . Then for  $\beta = (\beta_1^\top, \ldots, \beta_d^\top)^\top$ , we have

$$\boldsymbol{\beta}^{\top} \boldsymbol{\Psi} \boldsymbol{\beta} = \sum_{j=1}^{d} \boldsymbol{\beta}_{j}^{\top} \boldsymbol{\Sigma} \boldsymbol{\beta}_{j} \geq C_{\min}(\boldsymbol{\Sigma}, s) \sum_{j=1}^{d} \left\| (\boldsymbol{\beta}_{j})_{\mathcal{S}} \right\|_{2}^{2} = C_{\min}(\boldsymbol{\Sigma}, s) \left\| \boldsymbol{\beta}_{\mathcal{S}} \right\|_{2}^{2}.$$

Therefore, we conclude  $\widetilde{C}_{\min}(\Psi, s) \ge C_{\min}(\Sigma, s) > 0$  such that

$$|\mathcal{S}(\widehat{\boldsymbol{\beta}})| \le \frac{256\widetilde{C}_{\max}(\widehat{m})s}{9C_{\min}(\Sigma, s)},\tag{B.24}$$

with probability at least  $1 - \delta$ , as long as  $N \ge 32^2 L \log(3d^2/\delta)s^2/C_{\min}^2(\Sigma, s)$ .

For any  $\beta = (\beta_1^{\top}, \dots, \beta_d^{\top})^{\top}$  satisfying  $\|\beta_{\mathcal{S}(\beta^*)^c}\|_{2,0} \leq m_0$ , we have  $\|(\beta_j)_{\mathcal{S}(\beta^*)^c}\|_0 \leq m_0$  for any  $j \in [d]$ . Given a positive semi-definite matrix  $Z \in \mathbb{R}^{d \times d}$  and integer  $s \geq 1$ , the restricted maximum eigenvalue of Z are defined as

$$C_{\max}(Z,s) := \max_{\mathcal{S} \subset [d], |\mathcal{S}| \le s} \max_{\boldsymbol{\beta} \in \mathbb{R}^d} \Big\{ \frac{\langle \boldsymbol{\beta}, Z \boldsymbol{\beta} \rangle}{\|\boldsymbol{\beta}\|_2^2} : \boldsymbol{\beta} \in \mathbb{R}^d, \|\boldsymbol{\beta}_{\mathcal{S}^c}\|_0 \le s \Big\}.$$

Using the definition of  $C_{\max}(\widehat{\Sigma}, s)$ , it holds that

$$\boldsymbol{\beta}_j^{\top} \widehat{\boldsymbol{\Sigma}} \boldsymbol{\beta}_j \leq C_{\max}(\widehat{\boldsymbol{\Sigma}}, s) \| \boldsymbol{\beta}_j \|_2^2$$
, for any  $j \in [d]$ .

Summing the above inequality from 1 to d,

$$\boldsymbol{\beta}^{\top} \frac{X^{\top} X}{N} \boldsymbol{\beta} = \sum_{j=1}^{d} \boldsymbol{\beta}_{j}^{\top} \widehat{\boldsymbol{\Sigma}} \boldsymbol{\beta}_{j} \le C_{\max}(\widehat{\boldsymbol{\Sigma}}, s) \sum_{j=1}^{d} \|\boldsymbol{\beta}_{j}\|_{2}^{2} = C_{\max}(\widehat{\boldsymbol{\Sigma}}, s) \|\boldsymbol{\beta}\|_{2}^{2}.$$

This implies  $\widetilde{C}_{\max}(\widehat{m}) \leq C_{\max}(\widehat{\Sigma}, \widehat{m})$ . As shown in the Lemma 1 in Belloni et al. (2013), we have  $C_{\max}(\widehat{\Sigma}, m) \leq 4C_{\max}(\Sigma, m)$  for any  $m + s \leq \log(n)$  as long as  $n \gtrsim s$ .

Step 5. Recall that  $\widehat{m} = |S(\widehat{\beta}) \setminus S(\beta^*)|$  and denote

$$\mathcal{M} = \Big\{ m \in \mathbb{N}^+ : m > \frac{2048sC_{\max}(\Sigma, m)}{9C_{\min}(\Sigma, s)} \Big\}.$$

Suppose there is a  $m_0 \in \mathcal{M}$  such that  $\hat{m} > m_0$ . From Eq. (B.24), we know that

$$\widehat{m} \leq |\mathcal{S}(\widehat{\boldsymbol{\beta}})| \leq C_{\max}(\Sigma, m) \frac{1024s}{9C_{\min}(\Sigma, s)}$$

According to Lemma 3 in Belloni et al. (2013) for the sublinearity of sparse maximum eigenvalues, we have

$$C_{\max}(\Sigma, \widehat{m}) \leq \lceil \widehat{m}/m_0 \rceil C_{\max}(\Sigma, m_0) \leq 2C_{\max}(\Sigma, m_0) \widehat{m}/m_0.$$

where the last inequality we use  $\lceil \kappa \rceil \leq 2\kappa$ . Putting the above two results together, we have

$$m_0 \le \frac{2048sC_{\max}(\Sigma, m_0)}{9C_{\min}(\Sigma, s)}.$$

This leads a contradiction with the definition of  $\mathcal{M}$ . Therefore,  $\widehat{m} \leq m_0$  for all  $m_0 \in \mathcal{M}$ . This implies

$$\begin{aligned} |\mathcal{S}(\widehat{\boldsymbol{\beta}})| &\leq \min_{m_0 \in \mathcal{M}} C_{\max}(\Sigma, m_0) \frac{1024s}{9C_{\min}(\Sigma, s)} \\ &= \Big[ \frac{1024 \min_{m_0 \in \mathcal{M}} C_{\max}(\Sigma, m_0)}{9C_{\min}(\Sigma, s)} \Big] s \lesssim s. \end{aligned}$$

The term  $\min_{m_0 \in \mathcal{M}} C_{\max}(\Sigma, m_0)/C_{\min}(\Sigma, s)$  essentially characterizes the condition number of  $\Sigma$  on a restricted support and is upper bounded by the condition number defined in the full support. Now we finish the proof of the first part of Lemma 4.6 and start to prove the second part of Lemma 4.6 under separability condition.

According to Eq. (B.22), under event A we have

$$\begin{split} \left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_{2,1} &= \sum_{j=1}^a \left\|\widehat{\boldsymbol{\beta}}^j - \boldsymbol{\beta}^{*j}\right\|_2 = \sum_{j \in \mathcal{S}(\boldsymbol{\beta}^*)} \left\|\widehat{\boldsymbol{\beta}}^j - \boldsymbol{\beta}^{*j}\right\|_2 + \sum_{j \in \mathcal{S}(\boldsymbol{\beta}^*)^c} \left\|\widehat{\boldsymbol{\beta}}^j - \boldsymbol{\beta}^{*j}\right\|_2 \\ &\leq 4 \sum_{j \in \mathcal{S}(\boldsymbol{\beta}^*)} \left\|\widehat{\boldsymbol{\beta}}^j - \boldsymbol{\beta}^{*j}\right\|_2. \end{split}$$

From Eq. (B.23),

$$\begin{split} \sum_{j \in \mathcal{S}(\boldsymbol{\beta}^*)} \left\| \widehat{\boldsymbol{\beta}}^j - \boldsymbol{\beta}^{*j} \right\|_2 &\leq \sqrt{s} \left\| (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{\mathcal{S}(\boldsymbol{\beta}^*)} \right\|_2 \\ &\leq \sqrt{\frac{2s}{\widetilde{C}_{\min}(\Sigma, s)}} \frac{1}{\sqrt{N}} \left\| X(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \right\|_2 \\ &\leq \frac{4\sqrt{2}sd\lambda_2}{\widetilde{C}_{\min}(\Sigma, s)} \leq \frac{4\sqrt{2}sd\lambda_2}{C_{\min}(\Sigma, s)}. \end{split}$$

Combining the above two inequality together and plugging in the choice of  $\lambda_2$ , we can bound

$$\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_{2,1} \le \frac{64\sqrt{2}s\sqrt{d}}{C_{\min}(\Sigma,s)}\sqrt{\frac{2\log(2d^2/\delta)}{N}}.$$

with probability at least  $1 - \delta$ . Under Assumption 4.5, the following holds that with probability at least  $1 - \delta$ ,

$$\min_{j\in\mathcal{S}(\boldsymbol{\beta}^*)}\left\|\boldsymbol{\beta}^{*j}\right\|_2 > \left\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\right\|_{2,1} \ge \left\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\right\|_{2,\infty}.$$

If there is a  $j \in \mathcal{S}(\beta^*)$  but  $j \notin \mathcal{S}(\widehat{\beta})$ , we have

$$\left\|\widehat{\beta}^{j} - \beta^{*j}\right\|_{2} = \left\|\beta^{*j}\right\|_{2} > \left\|\widehat{\beta} - \beta^{*}\right\|_{2,\infty}$$

On the other hand,

$$\left\|\widehat{oldsymbol{eta}}^{j}-oldsymbol{eta}^{*j}
ight\|_{2}\leq\left\|\widehat{oldsymbol{eta}}-oldsymbol{eta}^{*}
ight\|_{2,\infty}$$

which leads a contradiction. Now we conclude that  $\mathcal{S}(\widehat{\beta}) \supseteq \mathcal{K}$ . This ends the proof.

#### B.3. Proof of Theorem 4.9: Instance-dependent upper bound

We restate the instance-dependent error bound error bound of vanilla fitted Q-evaluation algorithm on the full support.

**Theorem B.4** (Theorem 5 in Duan and Wang (2020)). Suppose Assumption 2.1 holds. Suppose  $\phi(x, a)^{\top} \Sigma^{-1} \phi(x, a) \lesssim d$  for any pair of (x, a). Let  $\delta \in (0, 1)$  and Algorithm 2 without feature selection stage takes N samples satisfying

$$N \gtrsim \frac{\gamma^2 \log(d/\delta)d}{(1-\gamma)^3}.$$

Set regularization parameter  $\lambda_3 = \lambda_{\min}(\Sigma) \log(12d/\delta)C_1d/(1-\gamma)$ . Letting the number of iteration  $T \to \infty$ , the following holds with probability at least  $1 - \delta$ ,

$$\widehat{v}_{\widehat{w}}^{\pi} - v^{\pi} \leq \frac{1}{1 - \gamma} \sum_{t=0}^{\infty} \gamma^{t+1} \sqrt{(\nu_t^{\pi})^\top \Sigma^{-1}(\nu_t^{\pi})} \sqrt{\frac{\log(1/\delta)}{N}} + \frac{\gamma \ln(12d/\delta)d}{N(1 - \gamma)^{3.5}},$$

where  $\nu_t^{\pi} = \mathbb{E}^{\pi}[\phi(x_t, a_t) | x \sim \xi_0].$ 

If the true relevant feature set  $\mathcal{K}$  is known in an oracle case, we could directly run the algorithm on  $\mathcal{K}$  such that all the dependency on d can be reduced to s and the instance-dependent term turns to be defined in the  $\mathcal{K}$  that is much sharper than the original one. Fortunately, Lemma 4.6 implies  $\hat{\mathcal{K}} \supseteq \mathcal{K}$  and  $|\hat{\mathcal{K}}| \leq s$ . Suppose

$$N \gtrsim \frac{\gamma^2 \log(s/\delta)s}{(1-\gamma)^3} \gtrsim \frac{\gamma^2 \log(|\widehat{\mathcal{K}}|/\delta)|\widehat{\mathcal{K}}|}{(1-\gamma)^3}.$$

Rewriting Theorem **B.4** with respect to  $\widehat{\mathcal{K}}$ , we have

$$|\widehat{v}_{\widehat{w}}^{\pi} - v^{\pi}| \lesssim \frac{1}{1 - \gamma} \sum_{t=0}^{\infty} \gamma^{t+1} \sqrt{(\widetilde{\nu}_t^{\pi})^{\top} \widetilde{\Sigma}^{-1}(\widetilde{\nu}_t^{\pi})} \sqrt{\frac{\log(1/\delta)}{N}},$$

		1

where  $\tilde{\nu}_t^{\pi} = [\nu_t^{\pi}]_{\widehat{\mathcal{K}}}$  and  $\tilde{\Sigma} = \Sigma_{\widehat{\mathcal{K}} \times \widehat{\mathcal{K}}}$ . The corresponding condition  $\phi(x, a)^{\top} \Sigma^{-1} \phi(x, a) \lesssim |\widehat{\mathcal{K}}|$  can be satisfied due to  $C_{\min}(\Sigma, s) > 0$  and  $\|\phi(x, a)\|_{\infty} \leq 1$ . From Definitions 4.7, 4.8 and Lemma B.2 in (Duan and Wang, 2020), we have

$$\begin{split} \sqrt{1 + \chi^2_{\mathcal{G}(\widehat{\mathcal{K}})}(\mu^{\pi}, \bar{\mu})} &= \sum_{t=0}^{\infty} \gamma^t \sup_{f \in \mathcal{G}(\widehat{\mathcal{K}})} \frac{(1 - \gamma) \mathbb{E}^{\pi}[f(x_t, a_t) | x_0 \sim \xi_0]}{\sqrt{\mathbb{E}[\frac{1}{L} \sum_{h=0}^{L-1} f^2(x_{1h}, a_{1h})]}} \\ &= (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t \sqrt{(\widetilde{\nu}_t^{\pi})^\top \widetilde{\Sigma}^{-1}(\widetilde{\nu}_t^{\pi})}. \end{split}$$

Now we end the proof.

# C. Proof of Theorem 5.1: Lasso Fitted Q-iteration

# C.1. Proof of Theorem 5.1

The main structure of this proof is similar to the proof of Theorem 4.3 in Appendix B.1 but we need to utilize the contraction property of Bellman optimality operator. Recall that we split the whole dataset into T folds and each fold consists of R episodes or RL sample transitions. The overall sample size is N = TRL.

**Step 1.** We verify that the execution of Algorithm 3 is equivalent to the approximate value iteration. Recall that a generic Lasso estimator with respect to a function V at tth phase is defined in Eq. (B.1) as

$$\widehat{w}_t(V) = \operatorname*{argmin}_{w \in \mathbb{R}^d} \left( \frac{1}{RL} \sum_{i=1}^{RL} \left( \Pi_{[0,1/(1-\gamma)]} V(x_i^{(t)'}) - \phi(x_i^{(t)}, a_i^{(t)})^\top w \right)^2 + \lambda_1 \|w\|_1 \right).$$

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

$$[\widehat{\mathcal{T}}^{(t)}V](x) := \max_{a} \Big[ r(x,a) + \gamma \phi(x,a)^{\top} \widehat{w}_t(V) \Big].$$
(C.1)

Note this  $\hat{\mathcal{T}}^{(t)}$  is a randomized operator that only depends data collected at *t*th phase. Algorithm 3 is equivalent to the following approximate value iteration:

$$\left[\widehat{\mathcal{T}}^{(t)}\Pi_{[0,1/(1-\gamma)]}V_{\widehat{w}_{t-1}}\right](x) = \max_{a} \left[r(x,a) + \gamma\phi(x,a)^{\top}\widehat{w}_{t}\right] = \max_{a} Q_{\widehat{w}_{t}}(x,a) = V_{\widehat{w}_{t}}(x).$$
(C.2)

Recall that the true Bellman optimality operator  $\mathcal{T}: \mathbb{R}^{\mathcal{X}} \to \mathbb{R}^{\mathcal{X}}$  is defined as

$$[\mathcal{T}V](x) := \max_{a} \Big[ r(x,a) + \gamma \sum_{x'} P(x'|x,a) V(x')' \Big].$$
(C.3)

**Step 2.** We verify that the true Bellman operator on  $\Pi_{[0,1/(1-\gamma)]}V_{\widehat{w}_{t-1}}$  can also be written as a linear form. From Assumption 2.1, there exists a vector  $\overline{w}_t \in \mathbb{R}^d$  with  $\overline{w}_{t,k} = 0$  if  $k \notin \mathcal{K}$  such that

$$[\mathcal{T}\Pi_{[0,1/(1-\gamma)]}V_{\widehat{w}_{t-1}}](x) = \max_{a} \Big[r(x,a) + \gamma \phi(x,a)^{\top} \bar{w}_t\Big].$$
(C.4)

**Step 3.** We start to bound  $||V_{\widehat{w}_t} - v^*||_{\infty}$  for each phase t. By the approximate value iteration form Eq. (C.2) and the definition of optimal value function,

$$\begin{aligned} \left\| V_{\widehat{w}_{t}} - v^{*} \right\|_{\infty} &= \left\| \widehat{\mathcal{T}}^{(t)} \Pi_{[0,1/(1-\gamma)]} V_{\widehat{w}_{t-1}} - \mathcal{T} v^{*} \right\|_{\infty} \\ &= \left\| \widehat{\mathcal{T}}^{(t)} \Pi_{[0,1/(1-\gamma)]} V_{\widehat{w}_{t-1}} - \mathcal{T} \Pi_{[0,1/(1-\gamma)]} V_{\widehat{w}_{t-1}} \right\|_{\infty} \\ &+ \left\| \mathcal{T} \Pi_{[0,1/(1-\gamma)]} V_{\widehat{w}_{t-1}} - \mathcal{T} v^{*} \right\|_{\infty}. \end{aligned}$$
(C.5)

The first term mainly captures the error between approximate Bellman optimality operator and true Bellman optimality operator while the second term can be bounded by the contraction of true Bellman operator. From linear forms Eqs. (C.2) and (C.4), it holds for any  $x \in \mathcal{X}$ ,

$$\begin{aligned} & [\widehat{\mathcal{T}}^{(t)}\Pi_{[0,1/(1-\gamma)]}V_{\widehat{w}_{t-1}}](x) - [\mathcal{T}\Pi_{[0,1/(1-\gamma)]}V_{\widehat{w}_{t-1}}](x) \\ &= \max_{a} \left[ r(x,a) + \gamma\phi(x,a)^{\top}\widehat{w}_{t} \right] - \max_{a} \left[ r(x,a) + \gamma\phi(x,a)^{\top}\overline{w}_{t} \right] \\ &\leq \gamma \max_{a} \left| \phi(x,a)^{\top}(\widehat{w}_{t} - \overline{w}_{t}) \right| \\ &\leq \gamma \max_{a,x} \|\phi(x,a)\|_{\infty} \|\widehat{w}_{t} - \overline{w}_{t}\|_{1}. \end{aligned}$$

$$(C.6)$$

Applying Lemma B.1, with the choice of  $\lambda_1 = (1 - \gamma)^{-1} \sqrt{\log(2d/\delta)/RL}$ , the following error bound holds with probability at least  $1 - \delta$ ,

$$\left\|\widehat{w}_t - \overline{w}_t\right\|_1 \le \frac{16\sqrt{2}s}{C_{\min}(\Sigma, s)} \frac{1}{1 - \gamma} \sqrt{\frac{\log(2d/\delta)}{RL}},\tag{C.7}$$

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

Note that the samples we use between phases are mutually independent. Thus Eq. (C.7) uniformly holds for all  $t \in [T]$  with probability at least  $1 - T\delta$ . Plugging it into Eq. (C.6), we have for any phase  $t \in [T]$ ,

$$\left\|\widehat{\mathcal{T}}^{(t)}\Pi_{[0,1/(1-\gamma)]}V_{\widehat{w}_{t-1}} - \mathcal{T}\Pi_{[0,1/(1-\gamma)]}V_{\widehat{w}_{t-1}}\right\|_{\infty} \le \gamma \frac{16\sqrt{2}s}{C_{\min}(\Sigma,s)} \frac{1}{1-\gamma} \sqrt{\frac{\log(2dT/\delta)}{RL}},\tag{C.8}$$

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

To bound the second term in Eq. (C.5), we use the contraction property of true Bellman operator such that

$$\left\|\mathcal{T}\Pi_{[0,1/(1-\gamma)]}V_{\widehat{w}_{t-1}} - \mathcal{T}v^*\right\|_{\infty} \le \gamma \left\|\Pi_{[0,1/(1-\gamma)]}V_{\widehat{w}_{t-1}} - v^*\right\|_{\infty}.$$
(C.9)

Plugging Eqs. (C.8) and (C.9) into Eq. (C.5), it holds that

$$\left\| V_{\widehat{w}_{t}} - v^{*} \right\|_{\infty} \leq \gamma \frac{16\sqrt{2}s}{C_{\min}(\Sigma, s)} \frac{1}{1 - \gamma} \sqrt{\frac{\log(2dT/\delta)}{RL}} + \gamma \left\| \Pi_{[0, 1/(1 - \gamma)]} V_{\widehat{w}_{t-1}} - v^{*} \right\|_{\infty}, \tag{C.10}$$

with probability at least  $1 - \delta$ . Recursively using Eq. (C.10), the following holds with probability  $1 - \delta$ ,

$$\begin{split} \left\| \Pi_{[0,1/(1-\gamma)]} V_{\widehat{w}_{T-1}} - v^* \right\|_{\infty} &\leq \left\| V_{\widehat{w}_{T-1}} - v^* \right\|_{\infty} \\ &= \gamma \frac{16\sqrt{2}s}{C_{\min}(\Sigma,s)} \frac{1}{1-\gamma} \sqrt{\frac{\log(2dT/\delta)}{RL}} + \gamma \left\| \Pi_{[0,1/(1-\gamma)]} V_{\widehat{w}_{T-2}} - v^* \right\|_{\infty} \\ &\leq \gamma^{T-1} \left\| \Pi_{[0,1/(1-\gamma)]} V_{\widehat{w}_0} - v^* \right\|_{\infty} + \sum_{t=1}^{T-1} \gamma^t \frac{16\sqrt{2}s}{C_{\min}(\Sigma,s)} \frac{1}{1-\gamma} \sqrt{\frac{\log(2dT/\delta)}{RL}} \\ &\leq \frac{2\gamma^{T-1}}{1-\gamma} + \frac{1}{(1-\gamma)^2} \frac{16\sqrt{2}s}{C_{\min}(\Sigma,s)} \sqrt{\frac{\log(2dT/\delta)}{RL}}, \end{split}$$

where the first inequality is due to that  $\Pi_{[0,1/(1-\gamma)]}$  can only make error smaller and the last inequality is from  $\sum_{t=1}^{T-1} \gamma^t \le 1/(1-\gamma)$ . By properly choosing  $T = \Theta(\log(N/(1-\gamma))/(1-\gamma))$ , it implies

$$\left\|\Pi_{[0,1/(1-\gamma)]}V_{\widehat{w}_{T-1}} - v^*\right\|_{\infty} \le \frac{1}{(1-\gamma)^{5/2}} \frac{32\sqrt{2}s}{C_{\min}(\Sigma,s)} \sqrt{\frac{\log(2dT/\delta)\log(N/(1-\gamma))}{N}}$$

holds with probability at least  $1 - \delta$ . From Proposition 2.14 in Bertsekas (1995),

$$\left\| v^{\widehat{\pi}_{T}} - v^{*} \right\|_{\infty} \leq \frac{1}{1 - \gamma} \left\| Q_{\widehat{w}_{T}} - Q^{*} \right\|_{\infty} \leq \frac{2}{1 - \gamma} \left\| \Pi_{[0, 1/(1 - \gamma)]} V_{\widehat{w}_{T-1}} - v^{*} \right\|_{\infty}.$$
 (C.11)

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

$$\|v^{\widehat{\pi}_T} - v^*\|_{\infty} \le \frac{64\sqrt{2}s}{C_{\min}(\Sigma, s)} \sqrt{\frac{\log(2dT/\delta)\log(N/(1-\gamma))}{N(1-\gamma)^7}}$$

when the sample size N satisfies

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

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

# **D.** Proof of Theorem 5.2: Minimax Lower Bound of Policy Optimization

### D.1. Reducing to likelihood test

We prove the minimax lower bound by conducting likelihood test. Similar to Lemma C.1 in Duan and Wang (2020), we have Lemma D.1 below.

**Lemma D.1.** Let  $M_{\alpha}$  and  $M_{\beta}$  be two MDP instances with transition kernels  $p_{\alpha}(x' | x, a)$  and  $p_{\beta}(x' | x, a)$ . Suppose Assumption 2.2 holds. Define likelihood functions

$$\mathcal{L}_{i}(\mathcal{D}) := \prod_{k=1}^{K} \bar{\xi}_{0}(x_{0}^{(k)}) \prod_{h=0}^{L-1} \bar{\pi}(a_{h}^{(k)} | x_{h}^{(k)}) p_{i}(x_{h+1}^{(k)} | x_{h}^{(k)}, a_{h}^{(k)}), \qquad i = \alpha, \beta.$$

Denote  $\mathbb{P}_{\alpha}$  the probability space generated by running  $M_{\alpha}$  following the behavioral policy  $\bar{\pi}$ . If  $\mathbb{P}_{\alpha}\left(\frac{\mathcal{L}_{\beta}(\mathcal{D})}{\mathcal{L}_{\alpha}(\mathcal{D})} \geq \frac{1}{2}\right) \geq \frac{1}{2}$  and there exist scalars  $\rho_{\alpha}, \rho_{\beta} \geq 0$  such that

$$\left\{\text{policy }\pi \mid v_{M_{\alpha},\xi_{0}}^{*}-v_{M_{\alpha},\xi_{0}}^{\pi} \geq \rho_{\alpha}\right\} \cap \left\{\text{policy }\pi \mid v_{M_{\beta},\xi_{0}}^{*}-v_{M_{\beta},\xi_{0}}^{\pi} \geq \rho_{\beta}\right\} = \emptyset,\tag{D.1}$$

then for any policy learning algorithm  $\hat{\pi}$ ,

$$\mathbb{P}_{\alpha}\left(v_{M_{\alpha},\xi_{0}}^{*}-v_{M_{\alpha},\xi_{0}}^{\widehat{\pi}(\mathcal{D})}\geq\rho_{\alpha}\right)\geq\frac{1}{6}\qquad\text{or}\qquad\mathbb{P}_{\beta}\left(v_{M_{\beta},\xi_{0}}^{*}-v_{M_{\beta},\xi_{0}}^{\widehat{\pi}(\mathcal{D})}\geq\rho_{\beta}\right)\geq\frac{1}{6}.$$
(D.2)

We learn from Lemma D.1 that as long as  $\mathbb{P}_{\alpha}\left(\frac{\mathcal{L}_{\beta}(\mathcal{D})}{\mathcal{L}_{\alpha}(\mathcal{D})} \geq \frac{1}{2}\right) \geq \frac{1}{2}$  and (D.1) hold, the lower bound is achieved at model  $M_{\alpha}$  or  $M_{\beta}$ . In the following, we construct MDP models and analyze these two conditions separately.

# **D.2.** Constructing MDP instances

We assume without loss of generality that the number of active features s is even. We consider a simplest case where the MDP only consists of two states, *i.e.*  $\mathcal{X} = \{\overline{x}, \underline{x}\}$ . At each state, the agent chooses from  $\frac{s}{2} + s(d - s)$  actions  $\mathcal{A} = \{a_1, a_2, \ldots, a_{\frac{s}{2}}\} \cup \{\overline{a}_{i,k} \mid i = 1, 2, \ldots, \frac{s}{2}, k = \pm 1, \pm 2, \ldots, \pm (d - s)\}$ . Here, we only use  $\overline{a}_{i,j}$  in collecting the dataset  $\mathcal{D}$ .

We first introduce Lemma D.2, which will be used in the construction of feature mapping  $\phi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$ . Lemma D.2. For any  $s \in \mathbb{Z}_+$ , there exists an *s*-by-*s* orthogonal matrix  $\Theta \in \mathbb{R}^{s \times s}$  satisfying

$$\sqrt{s} \cdot |\Theta_{i,j}| \le \sqrt{2}, \qquad \text{for } i, j = 1, 2, \dots, s.$$
(D.3)

*Proof.* Consider the discrete cosine transform (DCT) matrix  $\Theta \in \mathbb{R}^{s \times s}$ , given by

$$\Theta_{i,1} = \frac{1}{\sqrt{s}}, \ i = 1, 2, \dots, s, \qquad \Theta_{i,j} = \sqrt{\frac{2}{s}} \cos \frac{(2i-1)(j-1)\pi}{2s}, \ i = 1, 2, \dots, s, \ j = 2, 3, \dots, s.$$

 $\Theta$  is orthogonal and satisfies (D.3).

Let  $\Theta \in \mathbb{R}^{s \times s}$  be the orthogonal matrix given in Lemma D.2. We fix  $\mathcal{K} \subseteq [d]$  to be the active feature set and denote by  $\phi_{\mathcal{K}}$  the corresponding coordinates of  $\phi$ . We now construct  $\phi_{\mathcal{K}} : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^s$  as follows: For  $i = 1, 2, \ldots, \frac{s}{2}$ ,  $k = \pm 1, \pm 2, \ldots, \pm (d-s)$ , let

$$\begin{split} \phi_{\mathcal{K}}(\overline{x}, a_i) &:= \sqrt{\frac{s}{2}} \cdot \Theta \begin{pmatrix} 0 & 0 & \cdots & 1 & 0 & \cdots & 0 & 0 \end{pmatrix}^{\top} \in \mathbb{R}^s, \\ \phi_{\mathcal{K}}(\overline{x}, \overline{a}_{i,k}) &:= \sqrt{\frac{s}{2}} \cdot \Theta \begin{pmatrix} 0 & 0 & \cdots & 1 - \varsigma_1 & \varsigma_1 & \cdots & 0 & 0 \end{pmatrix}^{\top} \in \mathbb{R}^s, \\ \phi_{\mathcal{K}}(\underline{x}, a_i) &= \phi_{\mathcal{K}}(\underline{x}, \overline{a}_{i,k}) &:= \sqrt{\frac{s}{2}} \cdot \Theta \begin{pmatrix} 0 & 0 & \cdots & \varsigma_2 & 1 - \varsigma_2 & \cdots & 0 & 0 \end{pmatrix}^{\top} \in \mathbb{R}^s, \end{split}$$

where  $\varsigma_1, \varsigma_2 \in (0, 1)$  will be determined later. By construction, we have  $\|\phi_{\mathcal{K}}(x, a)\|_{\infty} \leq 1$  for any  $(x, a) \in \mathcal{X} \times \mathcal{A}$ . Note that  $\phi_{\mathcal{K}}$  abstracts all the dynamic informatrion for state-action pairs, and  $\phi_{\mathcal{K}^c}$  does not affect the transition model or reward function. Therefore, it is sufficient for us to use  $\phi_{\mathcal{K}}$  when identifying the optimal policy or calculate value functions.

We propose  $\frac{s}{2}$  MDP models  $M_1, M_2, \ldots, M_{\frac{s}{2}}$ , where  $M_i$  has transition kernel  $p_i(x' | x, a) = \phi_{\mathcal{K}}(x, a)^{\top} \psi_i(x')$  given by

$$\psi_i(\overline{x}) = \sqrt{\frac{2}{s}} \cdot \Theta \begin{pmatrix} 1 - \delta_1 & \delta_2 & \cdots & 1 & 0 & \cdots & 1 - \delta_1 & \delta_2 \end{pmatrix}^\top \in \mathbb{R}^s,$$
$$2i - 1 \quad 2i$$
$$\psi_i(\underline{x}) = \sqrt{\frac{2}{s}} \cdot \Theta \begin{pmatrix} \delta_1 & 1 - \delta_2 & \cdots & 0 & 1 & \cdots & \delta_1 & 1 - \delta_2 \end{pmatrix}^\top \in \mathbb{R}^s.$$

Here,  $\delta_1, \delta_2 \in [0, 2(1 - \gamma))$  are parameters reflecting the small differences among actions.

The reward functions are the same for all models and are chosen as

$$r(\overline{x}, a_i) = r(\overline{x}, \overline{a}_{i,k}) = 1, \qquad r(\underline{x}, a_i) = r(\underline{x}, \overline{a}_{i,k}) = 0,$$

for  $i = 1, 2, \dots, \frac{s}{2}, j = \pm 1, \pm 2, \dots, \pm (d - s).$ 

### D.3. Analyzing the concentration of the likelihood ratio

We devise a behavior policy  $\bar{\pi}$  and verify the likelihood ratio condition under the data collecting scheme in Assumption 2.2. We start from an initial distribution  $\bar{\xi}_0$  and take a behavior policy  $\bar{\pi}(\bar{a}_{i,k} | \bar{x}) = \bar{\pi}(\bar{a}_{i,k} | \underline{x}) = \frac{1}{s(d-s)}$  for any i, k. Under this specific  $\bar{\pi}$ , due to symmetry, all MDP models  $M_1, M_2, \ldots, M_{\frac{s}{2}}$  have the same marginal distribution at each time step  $l = 0, 1, \ldots, L-1$ , which we denote by  $\bar{\xi}_l = \left(\frac{\bar{\xi}_l(\bar{x})}{\bar{\xi}_l(\underline{x})}\right) \in \mathbb{R}^2$ . Define the average distribution as  $\bar{\xi} := \frac{1}{L} \sum_{l=0}^{L-1} \bar{\xi}_l \in \mathbb{R}^2$ .

Take

$$p_{\min} := \min\left\{p_i(\overline{x} \mid \overline{x}, \overline{a}_{i,k}), p_i(\underline{x} \mid \overline{x}, \overline{a}_{i,k}), p_i(\overline{x} \mid \underline{x}, \overline{a}_{i,k})), p_i(\overline{x} \mid \overline{x}, \overline{a}_{i,k})), \\ i = 1, 2, \dots, \frac{s}{2}, \ k = \pm 1, \pm 2 \dots, \pm (d-s)\right\}$$

and

$$\Sigma^{\circ} := \bar{\xi}(\overline{x}) \begin{pmatrix} 1 - \varsigma_1 \\ \varsigma_1 \end{pmatrix} (1 - \varsigma_1 \quad \varsigma_1) + \bar{\xi}(\underline{x}) \begin{pmatrix} \varsigma_2 \\ 1 - \varsigma_2 \end{pmatrix} (\varsigma_2 \quad 1 - \varsigma_2).$$

Parallel to Lemma C.3 in (Duan and Wang, 2020), we provide concentration results of the likelihood ratio in Lemma D.3. The proof can be found in Appendix E.2.1.

**Lemma D.3.** If we take  $\delta_1, \delta_2 \ge 0$  such that

$$\begin{pmatrix} \delta_1 & -\delta_2 \end{pmatrix} \Sigma^{\circ} \begin{pmatrix} \delta_1 \\ -\delta_2 \end{pmatrix} \le \frac{sp_{\min}}{100N}, \qquad \delta_1 \lor \delta_2 \le \frac{p_{\min}}{100\sqrt{L}},$$
(D.4)

then for any  $i, j = 1, 2, ..., s, i \neq j$ , it holds that

$$\mathbb{P}_i\left(\frac{\mathcal{L}_j(\mathcal{D})}{\mathcal{L}_i(\mathcal{D})} \ge \frac{1}{2}\right) \ge \frac{1}{2}.$$
(D.5)

Lemma D.3 suggests that as long as (D.4) is satisfied, the likelihood test in Lemma D.1 works for any pair of indices  $(\alpha, \beta) = (i, j), i \neq j$ .

#### D.4. Calculating the gap in values

For model  $M_i$ , the optimal policy is given by

$$\pi_i^*(\overline{x}) = a_i \quad \text{and} \quad \pi_i^*(\underline{x}) = \begin{cases} a_j \text{ for any } j \neq i, & \text{if } (1 - \varsigma_2)\delta_2 > \varsigma_2\delta_1, \\ a_i, & \text{otherwise.} \end{cases}$$

For computational simplicity, we take initial distribution  $\xi_0 := \begin{pmatrix} \xi_0(\overline{x}) \\ \xi_0(\underline{x}) \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \in \mathbb{R}^2$ . In the following Lemma D.4, we provide an estimation for the difference between values of optimal and sub-optimal policies. See Appendix E.2.2 for the proof.

**Lemma D.4.** If  $\delta_1 \leq \frac{1-\gamma}{\gamma}$ ,  $\delta_2 \leq \varsigma_2$ , then for any policy  $\pi$  such that  $\pi(\overline{x}) \neq a_i$ , it holds that

$$v^*_{M_i,\xi_0} - v^\pi_{M_i,\xi_0} \geq \frac{\gamma \delta_1}{2(1-\gamma)} \cdot \frac{1}{1-\gamma+2\gamma \varsigma_2}$$

According to Lemma D.4, if we take

$$\rho_i = \rho' := \frac{\gamma \delta_1}{2(1-\gamma)} \cdot \frac{1}{1-\gamma+2\gamma\varsigma_2}, \qquad i = 1, 2, \dots, \frac{s}{2},$$
(D.6)

then condition (D.1) in Lemma D.1 holds for any  $(\alpha, \beta) = (i, j), i \neq j$ .

# **D.5.** Choosing parameters

We now integrate Lemmas D.1, D.3 and D.4. Specifically, we choose parameters  $\varsigma_1, \varsigma_2, \delta_1, \delta_2$  and  $\bar{\xi}$  that maximize  $\rho'$  in (D.6) under the constraint (D.4).

We first consider the optimization problem

maximize 
$$\delta_1$$
, subject to  $\begin{pmatrix} \delta_1 & -\delta_2 \end{pmatrix} \Sigma^{\circ} \begin{pmatrix} \delta_1 \\ -\delta_2 \end{pmatrix} \leq \frac{sp_{\min}}{100N}$ 

It has solution

$$\delta_1 = \sqrt{\frac{\Sigma_{22}^\circ}{\det(\Sigma^\circ)}} \sqrt{\frac{sp_{\min}}{100N}}, \qquad \delta_2 = \sqrt{\frac{(\Sigma_{12}^\circ)^2}{\Sigma_{22}^\circ\det(\Sigma^\circ)}} \sqrt{\frac{sp_{\min}}{100N}}.$$
 (D.7)

Plugging (D.7) into (D.6) and assuming that  $p_{\min} \geq \frac{\varsigma_2}{2}$ , we have

$$\rho' \ge \frac{\gamma}{2(1-\gamma)} \cdot \frac{\sqrt{\varsigma_2}}{1-\gamma+2\gamma\varsigma_2} \cdot \sqrt{\frac{\Sigma_{22}^\circ}{\det(\Sigma^\circ)}} \sqrt{\frac{s}{200N}}.$$
(D.8)

We maximize the right hand side of (D.8) over  $\varsigma_2$ , and obtain

$$\varsigma_2 = \frac{1-\gamma}{2\gamma}, \qquad \rho' \ge \frac{\sqrt{\gamma}}{80} \sqrt{\frac{\Sigma_{22}^\circ}{\det(\Sigma^\circ)}} \sqrt{\frac{s}{N(1-\gamma)^3}}.$$

We further let  $\varsigma_1 \in \left[\frac{1-\gamma}{2\gamma}, 1-\frac{1-\gamma}{2\gamma}\right)$  and suppose the sample size

$$N \ge \frac{(\Sigma_{22}^{\circ} \lor \Sigma_{12}^{\circ})^2}{\Sigma_{22}^{\circ} \det(\Sigma^{\circ})} \frac{400sL}{1-\gamma}.$$
(D.9)

In this case,  $p_{\min} \ge \frac{\varsigma_2}{2}$  and  $\delta_1 \lor \delta_2 \le \frac{p_{\min}}{100\sqrt{L}} \le \varsigma_2 \le \frac{1-\gamma}{\gamma}$ .

In summary, if the sample size N satisfies (D.9) and we take

$$\gamma \geq \frac{1}{2}, \qquad \varsigma_1 \in \left[\frac{1-\gamma}{2\gamma}, 1-\frac{1-\gamma}{2\gamma}\right), \quad \varsigma_2 = \frac{1-\gamma}{2\gamma} \qquad \text{and} \qquad \delta_1, \delta_2 \text{ in (D.7)},$$

then the conditions in Lemmas D.3 and D.4 are satisfied and (D.2) holds for

$$\rho := \frac{1}{80\sqrt{2}} \sqrt{\frac{\Sigma_{22}^{\circ}}{\det(\Sigma^{\circ})}} \sqrt{\frac{s}{N(1-\gamma)^3}}.$$
(D.10)

Remark that under this construction, we still have the flexibility to take  $\varsigma_1 \nearrow 1 - \frac{1-\gamma}{2\gamma}$  so that  $\Sigma^{\circ}$  is very ill-conditioned. For instance, if we take  $\varsigma_1 = 1 - \frac{1-\gamma}{\gamma}$ , then  $\det(\Sigma^{\circ})$  or  $\lambda_{\min}(\Sigma^{\circ})$  at least has the order of  $(1-\gamma)^3$ .

In order that condition (D.9) is as weak as possible, we take  $\gamma \geq \frac{2}{3}$ ,  $\varsigma_1 = \frac{1-\gamma}{2\gamma}$  and  $\bar{\xi}(\bar{x}) = \bar{\xi}(\underline{x}) = \frac{1}{2}$ . In this setting, if  $N \geq 2000sL(1-\gamma)^{-1}$  then (D.9) holds.

# D.6. Relating to mismatch terms

In this part, we relate  $\frac{\Sigma_{22}^{\circ}}{\det(\Sigma^{\circ})}$  in (D.10) to mismatch terms  $\chi^2_{\mathcal{G}(\mathcal{K})}(\mu^*, \bar{\mu})$  and  $C_{\min}(\Sigma, s)$ .

# D.6.1. Restricted $\chi^2$ -divergence

According to Lemma B.2 in Duan and Wang (2020), we have

$$1 + \chi^{2}_{\mathcal{G}(\mathcal{K})}(\mu^{*},\bar{\mu}) = \sup_{f \in \mathcal{G}(\mathcal{K})} \frac{\mathbb{E}_{i} [f(x,a) | (x,a) \sim \mu^{*}]^{2}}{\mathbb{E}_{i} [f^{2}(x,a) | (x,a) \sim \bar{\mu}]} = (\nu^{*}_{\mathcal{K}})^{\top} \Sigma^{-1}_{\mathcal{K}} \nu^{*}_{\mathcal{K}},$$

where

$$\nu_{\mathcal{K}}^* := \mathbb{E}_i \left[ \phi_{\mathcal{K}}(x,a) \, \big| \, (x,a) \sim \mu^* \right] = \frac{1}{1-\gamma} \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_i \left[ \phi_{\mathcal{K}}(x_t,a_t) \, \big| \, x_0 \sim \xi_0, \pi_i^* \right] \in \mathbb{R}^s$$

and

$$\Sigma_{\mathcal{K}} := \mathbb{E}_{i} \left[ \phi_{\mathcal{K}}(x,a) \phi_{\mathcal{K}}(x,a)^{\top} \mid (x,a) \sim \bar{\mu} \right]$$
$$= \mathbb{E}_{i} \left[ \frac{1}{L} \sum_{h=0}^{L-1} \phi_{\mathcal{K}}(x_{h}^{(k)}, a_{h}^{(k)}) \phi_{\mathcal{K}}(x_{h}^{(k)}, a_{h}^{(k)})^{\top} \mid x_{0}^{(k)} \sim \bar{\xi}_{0}, \bar{\pi} \right] \in \mathbb{R}^{s \times s}$$

For model  $M_i$ ,  $\overline{x}$  is an absorbing state under the optimal policy  $\pi_i^*$ . Therefore,  $\mu^*(\overline{x}) = 1$  and  $\nu_{\mathcal{K}}^* = \phi_{\mathcal{K}}(\overline{x}, a_i)$ . Under our proposed behavior policy  $\overline{\pi}$ , we have

$$\Sigma_{\mathcal{K}} = \Theta \begin{pmatrix} \Sigma^{\circ} & 0 & \cdots & 0 \\ 0 & \Sigma^{\circ} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Sigma^{\circ} \end{pmatrix} \Theta^{\top}.$$
 (D.11)

It follows that

$$1 + \chi^2_{\mathcal{G}(\mathcal{K})}(\mu^*, \bar{\mu}) = \left(\nu^*_{\mathcal{K}}\right)^\top \Sigma_{\mathcal{K}}^{-1} \nu^*_{\mathcal{K}} = \frac{s}{2} \left( (\Sigma^\circ)^{-1} \right)_{1,1} = \frac{s \Sigma_{22}^\circ}{2 \det(\Sigma^\circ)}$$

To this end, we have

$$\rho = \frac{1}{80} \sqrt{1 + \chi^2_{\mathcal{G}(\mathcal{K})}(\mu^*, \bar{\mu})} \frac{1}{\sqrt{N(1-\gamma)^3}}.$$

This implies there always exists an DMDP instance  $M \in \mathcal{M}_{\phi,s}(\mathcal{X}, \mathcal{A}, \gamma)$  with feature  $\phi \in (\mathbb{R}^d)^{\mathcal{X} \times \mathcal{A}}$  satisfying  $\|\phi(x, a)\|_{\infty} \leq 1$  for all  $(x, a) \in \mathcal{X} \times \mathcal{A}$ , then

$$\mathbb{P}_{M}\left(v_{\xi_{0}}^{*}-v_{\xi_{0}}^{\widehat{\pi}(\mathcal{D})}\gtrsim\frac{1}{(1-\gamma)^{\frac{3}{2}}}\sqrt{1+\chi_{\mathcal{G}(\mathcal{K})}^{2}(\mu^{*},\bar{\mu})}\sqrt{\frac{1}{N}}\right)\geq\frac{1}{6},$$

where  $\mu^*$  is the discounted state-action occupancy measure of  $\pi^*$ .

### D.6.2. RESTRICTED MINIMUM EIGENVALUE

The uncentered covariance matrix  $\Sigma \in \mathbb{R}^{d \times d}$  is given by

$$\Sigma := \mathbb{E}_{i} \left[ \frac{1}{L} \sum_{h=0}^{L-1} \phi(x_{h}^{(k)}, a_{h}^{(k)}) \phi(x_{h}^{(k)}, a_{h}^{(k)})^{\top} \, \middle| \, x_{0}^{(k)} \sim \bar{\xi}_{0}, \bar{\pi} \right] \in \mathbb{R}^{s \times s}$$

In the following, we specify the choice of  $\phi_{\mathcal{K}^c}(\overline{x}, \overline{a}_{i,k})$  and  $\phi_{\mathcal{K}^c}(\underline{x}, \overline{a}_{i,k})$  and show that if

$$\bar{\xi}(\bar{x})\varsigma_1^2 + \bar{\xi}(\underline{x})(1-\varsigma_2)^2 \ge \bar{\xi}(\bar{x})(1-\varsigma_1)^2 + \bar{\xi}(\underline{x})\varsigma_2^2, \tag{D.12}$$

then

$$\frac{\Sigma_{22}^{\circ}}{\det(\Sigma^{\circ})} \ge \frac{1}{2C_{\min}(\Sigma, s)}.$$
(D.13)

Under condition (D.12), it holds that  $\Sigma_{22}^{\circ} \geq \Sigma_{11}^{\circ}$ , therefore,  $\frac{\Sigma_{22}^{\circ}}{\det(\Sigma^{\circ})} \geq \frac{\operatorname{Tr}(\Sigma^{\circ})}{2\det(\Sigma^{\circ})}$ . In addition, for the 2-by-2 matrix  $\Sigma^{\circ}$ , we have  $\lambda_{\min}(\Sigma^{\circ}) + \lambda_{\max}(\Sigma^{\circ}) = \operatorname{Tr}(\Sigma^{\circ})$  and  $\lambda_{\min}(\Sigma^{\circ})\lambda_{\max}(\Sigma^{\circ}) = \det(\Sigma^{\circ})$ . It follows that

$$\frac{\Sigma_{22}^{\circ}}{\det(\Sigma^{\circ})} \geq \frac{\operatorname{Tr}(\Sigma^{\circ})}{2\det(\Sigma^{\circ})} \geq \frac{\lambda_{\max}(\Sigma^{\circ})}{2\lambda_{\max}(\Sigma^{\circ})\lambda_{\min}(\Sigma^{\circ})} = \frac{1}{2\lambda_{\min}(\Sigma^{\circ})}$$

We next relate  $\lambda_{\min}(\Sigma^{\circ})$  to  $C_{\min}(\Sigma, s)$ .

Let  $\bar{\Theta} \in \mathbb{R}^{(d-s) \times (d-s)}$  be an orthogonal matrix given by Lemma D.2. We take

$$\phi_{\mathcal{K}^c}(\overline{x}, \overline{a}_{i,k}) = \phi_{\mathcal{K}^c}(\underline{x}, \overline{a}_{i,k}) := \operatorname{sign}(k) \sqrt{\frac{d-s}{2}} \cdot \operatorname{col}_k(\overline{\Theta}),$$

for  $i = 1, 2, \ldots, \frac{s}{2}$ ,  $k = \pm 1, \pm 2, \ldots, \pm (d - s)$ . It holds that  $\|\phi_{\mathcal{K}^c}(\overline{x}, \overline{a}_{i,k})\|_{\infty} \leq 1$  and  $\|\phi_{\mathcal{K}^c}(\underline{x}, \overline{a}_{i,k})\|_{\infty} \leq 1$ . For notational simplicity, let  $\mathcal{K} = [s]$ . Under our proposed behavior policy  $\overline{\pi}(\overline{a}_{i,k} \mid \overline{x}) = \overline{\pi}(\overline{a}_{i,k} \mid \underline{x}) = \frac{1}{s(d-s)}$ , we have

$$\Sigma = \begin{pmatrix} \Sigma_{\mathcal{K}} & 0\\ 0 & \frac{1}{2}I_{d-s} \end{pmatrix}.$$
 (D.14)

By (D.11),  $\lambda_{\min}(\Sigma_{\mathcal{K}}) = \lambda_{\min}(\Sigma^{\circ})$ . We also note that  $\operatorname{Tr}(\Sigma^{\circ}) = \overline{\xi}(\overline{x}) ||(1 - \varsigma_1, \varsigma_1)||_2^2 + \overline{\xi}(\underline{x}) ||(\varsigma_2, 1 - \varsigma_2)||_2^2 \leq 1$ , and therefore

$$\lambda_{\min}(\Sigma^{\circ}) \leq \frac{\operatorname{Tr}(\Sigma^{\circ})}{2} \leq \frac{1}{2}.$$

It follows that  $\lambda_{\min}(\Sigma) = \lambda_{\min}(\Sigma^{\circ})$ , which further implies  $C_{\min}(\Sigma, s) \ge \lambda_{\min}(\Sigma) = \lambda_{\min}(\Sigma^{\circ})$ . On the other hand, the eigenvector of  $\Sigma$  corresponding to  $\lambda_{\min}(\Sigma^{\circ})$  has support set  $\mathcal{K}$  and is *s*-sparse. Hence,  $\lambda_{\min}(\Sigma^{\circ}) \ge C_{\min}(\Sigma, s)$ . In this way, we have proved  $C_{\min}(\Sigma, s) = \lambda_{\min}(\Sigma^{\circ})$  for  $\Sigma$  defined in (D.14).

In the special case where  $\varsigma_1 = \varsigma_2 = \frac{1-\gamma}{2\gamma}$  and  $\bar{\xi}(\bar{x}) = \bar{\xi}(\underline{x}) = \frac{1}{2}$ , condition (D.12) holds. Plugging (D.13) into (D.10), we finish our proof of Theorem 5.2.

# **E.** Proofs of Auxiliary Results

## E.1. Proof of Lemma B.3

We prove if the population covariance matrix satisfies the restricted eigenvalue condition, the empirical covariance matrix satisfies it as well with high probability. Recall that

$$\widehat{\Sigma} = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{L} \sum_{h=0}^{L-1} \phi(x_h^{(k)}, a_h^{(k)}) \phi(x_h^{(k)}, a_h^{(k)})^{\top},$$

and

$$\widehat{\Psi} = \begin{pmatrix} \widehat{\Sigma} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \widehat{\Sigma} . \end{pmatrix}.$$

For any  $i, j \in [d]$ , define

$$v_{ij}^{(k)} = \frac{1}{L} \sum_{h=0}^{L-1} \phi_i(x_h^{(k)}, a_h^{(k)}) \phi_j(x_h^{(k)}, a_h^{(k)}) - \Sigma_{ij}.$$

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

$$\mathbb{P}\Big(\Big|\sum_{k=1}^{K} v_{ij}^{(k)}\Big| \ge \delta\Big) \le 3\exp\Big(-\frac{C_0\delta^2}{K}\Big)$$

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

$$\mathbb{P}\Big(\max_{i,j}\Big|\sum_{k=1}^{K} v_{ij}^{(k)}\Big| \ge \delta\Big) \le 3d^2 \exp\Big(-\frac{C_0\delta^2}{K}\Big)$$
$$\Rightarrow \mathbb{P}\Big(\big\|\widehat{\Sigma} - \Sigma\big\|_{\infty} \ge \delta\Big) \le 3d^2 \exp\Big(-\frac{C_0\delta^2}{K}\Big).$$

Since the blocks of  $\Psi$  are the same, the following holds holds with probability  $1 - \delta$ .

$$\left\|\widehat{\Psi} - \Psi\right\|_{\infty} \le \sqrt{\frac{\log(3d^2/\delta)}{K}}$$

Therefore, when the number of episodes  $K \ge 32^2 \log(3d^2/\delta)s^2/\widetilde{C}_{\min}(\Psi, s)^2$ , the following holds with probability at least  $1 - \delta$ ,

$$\left\|\widehat{\Psi} - \Psi\right\|_{\infty} \le \frac{\widetilde{C}_{\min}(\Psi, s)}{32s}.$$

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

Lemma E.1 (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  $\widetilde{C}_{\min}(\Sigma_0, s) > 0$  and  $\|\Sigma_1 - \Sigma_0\|_{\infty} \leq \widetilde{C}_{\min}(\Sigma_0, s)/(32s)$ . Then the restricted eigenvalue of  $\Sigma_1$  satisfies  $\widetilde{C}_{\min}(\Sigma_1, s) > \widetilde{C}_{\min}(\Sigma_0, s)/2$ .

Applying Lemma E.1 with  $\widehat{\Psi}$  and  $\Psi$ , we have the restricted eigenvalue of  $\widehat{\Phi}$  satisfies  $\widetilde{C}_{\min}(\widehat{\Psi}, s) > \widetilde{C}_{\min}(\Psi, s)/2$  with probability at least  $1 - \delta$ , as long as the sample size  $N \ge 32^2 L \log(3d^2/\delta)s^2/\widetilde{C}_{\min}(\Psi, s)^2$ . This ends the proof.

### E.2. Proof of Lemma B.1

We prove the  $\ell_1$ -norm bound of estimating  $\bar{w}_t$  using a fresh fold of batch data. We overload the notation  $\hat{\Sigma}$  to denote

$$\widehat{\Sigma} = \frac{1}{RL} \sum_{i=1}^{RL} \phi(x_i, a_i) \phi(x_i, a_i)^{\top} = \frac{1}{R} \sum_{r=1}^{R} \frac{1}{L} \sum_{h=1}^{L} \phi(x_h^{(r)}, a_h^{(r)}) \phi(x_h^{(r)}, a_h^{(r)})^{\top}$$

Similar to the proof of Lemma B.3 in Appendix E.1, we can have with probability at least  $1 - \delta$ ,

$$\left\|\widehat{\Sigma} - \Sigma\right\|_{\infty} \le \sqrt{\frac{C_1}{R}\log\left(\frac{3d^2}{\delta}\right)},$$

where  $C_1$  is an absolute constant. When  $R \ge C_1 32^2 \log(3d^2/\delta)s^2/C_{\min}(\Sigma, s)$ , we have

$$\left\|\widehat{\Sigma} - \Sigma\right\|_{\infty} \le \frac{C_{\min}(\Sigma, s)}{32s}.$$

Applying Lemma E.1, we have  $C_{\min}(\widehat{\Sigma}, s) > C_{\min}(\Sigma, s)/2$  with probability at least  $1 - \delta$ . Note that  $\{\varepsilon_i \phi_j(x_i, a_i)\}_{i=1}^{RL}$  is a martingale difference sequence and  $|\varepsilon_i \phi_j(x_i, a_i)| \le 1/(1 - \gamma)$ . Similar to the proof of Eq. (B.16) by Azuma-Hoeffding inequality,

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

Denote event  $\ensuremath{\mathcal{E}}$  as

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

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

$$\left\|\widehat{w}_t - \bar{w}_t\right\|_1 \le \frac{16\sqrt{2}s\lambda_1}{C_{\min}(\Sigma, s)},$$

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

### E.2.1. PROOF OF LEMMA D.3

*Proof of Lemma D.3.* It is easy to see that

$$\ln \frac{\mathcal{L}_j(\mathcal{D})}{\mathcal{L}_i(\mathcal{D})} = \sum_{k=1}^K \sum_{l=0}^{L-1} \ln \frac{p_j(x_{l+1}^{(k)} | x_l^{(k)}, a_l^{(k)})}{p_i(x_{l+1}^{(k)} | x_l^{(k)}, a_l^{(k)})} = \sum_{k=1}^K \sum_{l=0}^{L-1} \ln(1 - \Lambda_l^{(k)}),$$

where

$$\Lambda_{l}^{(k)} := \frac{p_{i}(x_{l+1}^{(k)} \mid x_{l}^{(k)}, a_{l}^{(k)}) - p_{j}(s_{l+1}^{(k)} \mid s_{l}^{(k)}, a_{l}^{(k)})}{p_{i}(s_{l+1}^{(k)} \mid s_{l}^{(k)}, a_{l}^{(k)})} = \frac{\phi(s_{l}^{(k)}, a_{l}^{(k)})^{\top} \left(\psi_{i}(s_{l+1}^{(k)}) - \psi_{j}(s_{l+1}^{(k)})\right)}{p_{i}(s_{l+1}^{(k)} \mid s_{l}^{(k)}, a_{l}^{(k)})}$$

If we take  $\delta_1 \vee \delta_2 \leq rac{p_{\min}}{2},$  then  $|\Lambda_l^{(k)}| \leq rac{1}{2}$  and

$$\ln \frac{\mathcal{L}_{j}(\mathcal{D})}{\mathcal{L}_{i}(\mathcal{D})} \ge -\sum_{k=1}^{K} \sum_{l=0}^{L-1} \Lambda_{l}^{(k)} - \sum_{k=1}^{K} \sum_{l=0}^{L-1} \left(\Lambda_{l}^{(k)}\right)^{2}.$$
(E.1)

Since  $\mathbb{E}_{i}[\Lambda_{l}^{(k)} | s_{l}^{(k)}, a_{l}^{(k)}] = 0$ , we apply Freedman's inequality to analyze  $E_{1}$ . The conditional variances satisfy

$$\begin{split} & \mathbb{E}_{i} \left[ \left( \Lambda_{l}^{(k)} \right)^{2} \middle| s_{l}^{(k)}, a_{l}^{(k)} \right] \\ &= p_{i}(\overline{x} \middle| s_{l}^{(k)}, a_{l}^{(k)}) \left( \frac{\phi(s_{l}^{(k)}, a_{l}^{(k)})^{\top} \left( \psi_{i}(\overline{x}) - \psi_{j}(\overline{x}) \right)}{p_{i}(\overline{x} \middle| s_{l}^{(k)}, a_{l}^{(k)})} \right)^{2} \\ &+ p_{i}(\underline{x} \middle| s_{l}^{(k)}, a_{l}^{(k)}) \left( \frac{\phi(s_{l}^{(k)}, a_{l}^{(k)})^{\top} \left( \psi_{i}(\underline{x}) - \psi_{j}(\underline{x}) \right)}{p_{i}(\underline{x} \middle| s_{l}^{(k)}, a_{l}^{(k)})} \right)^{2} \\ &= \frac{\left( \phi(s_{l}^{(k)}, a_{l}^{(k)})^{\top} \left( \psi_{i}(\overline{x}) - \psi_{j}(\overline{x}) \right) \right)^{2}}{p_{i}(\overline{x} \middle| s_{l}^{(k)}, a_{l}^{(k)})} \leq \frac{1}{p_{\min}(1 - p_{\min})} \left( \phi(s_{l}^{(k)}, a_{l}^{(k)})^{\top} \left( \psi_{i}(\overline{x}) - \psi_{j}(\overline{x}) \right) \right)^{2}. \end{split}$$

Denote  $\Xi_k := \frac{1}{L} \sum_{l=0}^{L-1} \left( \phi(s_l^{(k)}, a_l^{(k)})^\top (\psi_i(\overline{x}) - \psi_j(\overline{x})) \right)^2$ . Note that

$$\psi_i(\overline{x}) - \psi_j(\overline{x}) = \sqrt{\frac{2}{s}} \cdot \Theta \begin{pmatrix} 0 & 0 & \cdots & \delta_1 & -\delta_2 & \cdots & -\delta_1 & \delta_2 & \cdots & 0 & 0 \end{pmatrix}^\top \in \mathbb{R}^s,$$

therefore,

$$\mathbb{E}[\Xi_k] = \frac{4}{s} \begin{pmatrix} \delta_1 & -\delta_2 \end{pmatrix} \Sigma^{\circ} \begin{pmatrix} \delta_1 \\ -\delta_2 \end{pmatrix} \quad \text{and} \quad |\Xi_k| \le (\delta_1 \lor \delta_2)^2$$

By Bernstein inequality and the independence of trajectories  $\tau_1, \tau_2, \ldots, \tau_K$ , we have with  $\mathbb{P}_i$ -probability at least  $\frac{5}{6}$ ,

$$\frac{1}{K}\sum_{k=1}^{K}\Xi_k \le \left(\sqrt{\mathbb{E}[\Xi_k]} + (\delta_1 \lor \delta_2)\sqrt{\frac{2\ln 6}{3K}}\right)^2 =: \sigma^2.$$
(E.2)

Since  $|\Lambda_l^{(k)}| \le p_{\min}^{-1}(\delta_1 \vee \delta_2)$ , by Freedman's inequality, with  $\mathbb{P}_i$ -probability at least  $\frac{5}{6}$ ,

$$\frac{1}{N}\sum_{k=1}^{K}\sum_{l=0}^{L-1}\Lambda_{l}^{(k)} \le \frac{\sigma}{\sqrt{p_{\min}(1-p_{\min})}}\sqrt{\frac{2\ln 6}{N}} + p_{\min}^{-1}(\delta_{1} \lor \delta_{2})\frac{2\ln 6}{3N} \quad \text{and} \quad (E.2) \text{ holds.}$$
(E.3)

Combining (E.2) and (E.3), we use union bound and derive that with  $\mathbb{P}_i$ -probabliity at least  $\frac{2}{3}$ ,

$$\sum_{k=1}^{K} \sum_{l=0}^{L-1} \Lambda_l^{(k)} \le 4\sqrt{\ln 6} \sqrt{\left(\delta_1 - \delta_2\right) \Sigma^{\circ} \left(\frac{\delta_1}{-\delta_2}\right)} \sqrt{\frac{N}{p_{\min}s}} + \frac{2\ln 6}{3} (\delta_1 \lor \delta_2) \left(\sqrt{\frac{6L}{p_{\min}}} + \frac{1}{p_{\min}}\right). \tag{E.4}$$

As for  $E_2$ , since

$$\mathbb{E}_{i}\left[\frac{1}{L}\sum_{l=0}^{L-1} \left(\Lambda_{l}^{(k)}\right)^{2}\right] \leq \frac{4}{sp_{\min}(1-p_{\min})} \left(\delta_{1} - \delta_{2}\right) \Sigma^{\circ} \left(\frac{\delta_{1}}{-\delta_{2}}\right), \quad \left|\frac{1}{L}\sum_{l=0}^{L-1} \left(\Lambda_{l}^{(k)}\right)^{2}\right| \leq \frac{\left(\delta_{1} \vee \delta_{2}\right)^{2}}{p_{\min}^{2}},$$

by Bernstein's inequality, with  $\mathbb{P}_i$ -probability at least  $\frac{5}{6}$ ,

$$\sum_{k=1}^{K} \sum_{l=0}^{L-1} \left( \Lambda_{l}^{(k)} \right)^{2} \leq \left( 2\sqrt{2} \sqrt{\left( \delta_{1} - \delta_{2} \right) \Sigma^{\circ} \left( \frac{\delta_{1}}{-\delta_{2}} \right)} \sqrt{\frac{N}{sp_{\min}}} + \frac{\delta_{1} \vee \delta_{2}}{p_{\min}} \sqrt{\frac{2L \ln 6}{3}} \right)^{2}.$$
(E.5)

Plugging (E.4) and (E.5) into (E.1) and applying condition (D.4), we obtain (D.5).

### E.2.2. PROOF OF LEMMA D.4

*Proof of Lemma D.4.* We consider another policy  $\pi'_i$  such that  $\pi'_i(\overline{x}) = a_j$  for some  $a_j \neq a_i$  and  $\pi'_i(\underline{x}) = \pi^*_i(\underline{x})$ . It holds that

$$v_{M_i,\xi_0}^{\pi'_i} = \sup\left\{v_{M_i,\xi_0}^{\pi} \mid \pi(\overline{x}) \neq a_i\right\}$$

For any  $\pi$ , denote its corresponding transition matrix by  $P_i^{\pi} = \begin{pmatrix} p_i^{\pi}(\overline{x} \mid \overline{x}) & p_i^{\pi}(\underline{x} \mid \overline{x}) \\ p_i^{\pi}(\overline{x} \mid \overline{x}) & p_i^{\pi}(\underline{x} \mid \underline{x}) \end{pmatrix} \in \mathbb{R}^{2 \times 2}$ . Define  $v_{M_i}^* := \langle v_{M_i}^* \in (\overline{x}) \rangle$ 

 $\begin{pmatrix} v_{M_i}^*(\overline{x})\\ v_{M_i}^*(\underline{x}) \end{pmatrix} \in \mathbb{R}^2.$  We have the following decomposition,

$$v_{M_{i},\xi_{0}}^{*} - v_{M_{i},\xi_{0}}^{\pi_{i}'} = \sum_{t=0}^{\infty} \gamma^{t+1} \xi_{0}^{\top} \left( P_{i}^{\pi_{i}'} \right)^{t} \left( P_{i}^{\pi_{i}^{*}} - P_{i}^{\pi_{i}'} \right) v_{M_{i}}^{*} = \gamma \xi_{0}^{\top} \left( I - \gamma P_{i}^{\pi_{i}'} \right)^{-1} \left( P_{i}^{\pi_{i}^{*}} - P_{i}^{\pi_{i}'} \right) v_{M_{i}}^{*}.$$
(E.6)

Under model  $M_i$ , when  $\delta_2 \leq \varsigma_2$ ,  $\pi_i^*$  and  $\pi_i'$  satisfy

$$p_i^{\pi_i^*}(\overline{x} \,|\, \overline{x}) = 1, \quad p_i^{\pi_i'}(\overline{x} \,|\, \overline{x}) = 1 - \delta_1, \qquad p_i^{\pi_i^*}(\underline{x} \,|\, \overline{x}) = 0, \quad p_i^{\pi_i'}(\underline{x} \,|\, \overline{x}) = \delta_1$$
$$p_i^{\pi_i^*}(\overline{x} \,|\, \underline{x}) = p_i^{\pi_i'}(\overline{x} \,|\, \underline{x}) \le 2\varsigma_2, \qquad p_i^{\pi_i^*}(\underline{x} \,|\, \underline{x}) = p_i^{\pi_i'}(\underline{x} \,|\, \underline{x}) \ge 1 - 2\varsigma_2,$$

therefore,

$$\left(P_i^{\pi_i^*} - P_i^{\pi_i'}\right)v_{M_i}^* = \begin{pmatrix}\delta_1 & -\delta_1\\0 & 0\end{pmatrix}v_{M_i}^* = \begin{pmatrix}\delta_1\\0\end{pmatrix}\cdot\left(v_{M_i}^*(\overline{x}) - v_{M_i}^*(\underline{x})\right).$$

To this end, we reduce (E.6) into

$$v_{M_{i},\xi_{0}}^{*} - v_{M_{i},\xi_{0}}^{\pi_{i}'} = \gamma \xi_{0}^{\top} \left( I - \gamma P_{i}^{\pi_{i}'} \right)^{-1} \begin{pmatrix} \delta_{1} \\ 0 \end{pmatrix} \cdot \left( v_{M_{i}}^{*}(\overline{x}) - v_{M_{i}}^{*}(\underline{x}) \right).$$
(E.7)

Note that

$$v_{M_i}^*(\overline{x}) = (1-\gamma)^{-1}$$
 and  $v_{M_i}^*(\underline{x}) = rac{\gamma}{1-\gamma} \cdot rac{p_i^{\pi_i}(\underline{x}\,|\,\underline{x})}{1-\gammaig(1-p_i^{\pi_i^*}(\underline{x}\,|\,\underline{x})ig)}.$ 

Therefore,

$$v_{M_{i}}^{*}(\overline{x}) - v_{M_{i}}^{*}(\underline{x}) = \frac{1}{1 - \gamma} - \frac{\gamma}{1 - \gamma} \cdot \frac{p_{i}^{\pi_{i}^{*}}(\underline{x} \,|\, \underline{x})}{1 - \gamma \left(1 - p_{i}^{\pi_{i}^{*}}(\underline{x} \,|\, \underline{x})\right)} = \frac{1}{1 - \gamma p_{i}^{\pi_{i}^{*}}(\underline{x} \,|\, \underline{x})} \ge \frac{1}{1 - \gamma (1 - 2\varsigma_{2})}.$$
 (E.8)

In addition, we have

$$\xi_0^{\top} \left( I - \gamma P_i^{\pi_i'} \right)^{-1} \begin{pmatrix} \delta_1 \\ 0 \end{pmatrix} = \frac{\delta_1}{1 - \gamma} \cdot \frac{1 - \gamma P_i^{\pi_i^-}(\underline{x} \mid \underline{x})}{1 + \gamma \delta_1 - \gamma P_i^{\pi_i^*}(\underline{x} \mid \underline{x})}$$

Under the condition  $\delta_1 \leq \frac{1-\gamma}{\gamma}$ , we have  $\gamma \delta_1 \leq 1 - \gamma \leq 1 - \gamma p_i^{\pi_i^*}(\underline{x} \,|\, \underline{x})$ , therefore,

$$\xi_0^{\top} \left( I - \gamma P_i^{\pi_i'} \right)^{-1} \begin{pmatrix} \delta_1 \\ 0 \end{pmatrix} \ge \frac{\delta_1}{2(1-\gamma)}.$$
(E.9)

Plugging (E.8) and (E.9) into (E.7), we finish our proof.

# **F.** Supporting Lemmas

**Lemma F.1.** Let  $Z_1, \ldots, Z_n$  be random, positive-semidefinite adaptively chosen matrices with dimension d. Suppose  $\lambda_{\max}(Z_i) \leq R^2$  almost surely for all i. Let  $Z^+ = \sum_{i=1}^n Z_i$  and  $W = \sum_{i=1}^n \mathbb{E}[Z_i|Z_1, \ldots, Z_{i-1}]$ . Then for any  $\mu$  and any  $\alpha \in (0, 1)$ ,

$$\mathbb{P}\Big(\lambda_{\min}(Z^+) \le (1-\alpha)\mu \text{ and } \lambda_{\min}(W) \ge \mu\Big) d\Big(\frac{1}{e^{\alpha}(1-\alpha)^{1-\alpha}}\Big)^{\mu/R^2}$$

**Lemma F.2** (Azuma-Hoeffding's inequality). Let  $\mathcal{F}_n = \sigma(x_1, \ldots, x_n)$  be a sequence of  $\sigma$ -fields known as a filtration. Let  $\{(x_n, \mathcal{F}_n)\}_{n=1}^{\infty}$  be a martingale difference sequence for which there are constants  $\{(a_k, b_k)_{k=1}^n\}$  such that  $x_k \in [a_k, b_k]$  almost surely for  $k = 1, \ldots, n$ . Then for all  $t \ge 0$ ,

$$\mathbb{P}\Big(\Big|\sum_{k=1}^n x_k\Big| \ge t\Big) \le 2\exp\Big(-\frac{2t^2}{\sum_{k=1}^n (b_k - a_k)^2}\Big).$$