E-companion to "Nonsparse Learning with Latent Variables"

This e-companion consists of two parts. Section EC.1 lists the key lemmas and presents the proofs for main results. Additional technical proofs for the lemmas are provided in Section EC.2.

EC.1. Proofs of main results

EC.1.1. Lemmas

The following lemmas are used in the proofs of main results.

LEMMA EC.1 **(Consistency of spiked sample eigenvalues)**. *Under Conditions 1 and 2, with asymptotic probability one, the eigenvalues of the sample covariance matrix* $\mathbf{S}$ *satisfy that for any* l, $1 \le l \le m$, *uniformly over* $i \in J_l$,

$$q^{-\alpha_l}\widehat{\lambda}_i \to c_i \ \text{ as } q \to \infty.$$

LEMMA EC.2. *Denote by* $\mathbf{X}_0$ *and* $\widehat{\mathbf{F}}_0$ *the submatrices of* $\mathbf{X}$ *and* $\widehat{\mathbf{F}}$ *consisting of columns in* $\mathrm{supp}(\boldsymbol{\beta}_0)$ *and* $\mathrm{supp}(\boldsymbol{\gamma}_0)$, *respectively, and* $\widetilde{\boldsymbol{\varepsilon}} = (\mathbf{F} - \widehat{\mathbf{F}})\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$. *For the following two events*

$$\widetilde{\mathcal{E}} = \left\{ \|n^{-1}(\mathbf{X}, \widehat{\mathbf{F}})^T \widetilde{\boldsymbol{\varepsilon}}\|_\infty \le c_2 \sqrt{(\log p)/n} \right\} \quad \text{and}$$
$$\widetilde{\mathcal{E}}_0 = \left\{ \|n^{-1}(\mathbf{X}_0, \widehat{\mathbf{F}}_0)^T \widetilde{\boldsymbol{\varepsilon}}\|_\infty \le c_2 \sqrt{(\log n)/n} \right\}$$

with constant $c_2 > 2\sqrt{2}\sigma$, *when the estimation error bound of* $\widehat{\mathbf{F}}$ *in Condition 3 holds and the columns of* $\mathbf{X}$ *adopt a common scale of* L_2*-norm* $n^{1/2}$, *we have*

$$P(\widetilde{\mathcal{E}} \cap \widetilde{\mathcal{E}}_0) \ge 1 - \frac{4\sqrt{2}\sigma}{c_2\sqrt{\pi \log p}} p^{1 - \frac{c_2^2}{8\sigma^2}} - \frac{2\sqrt{2}\sigma s}{c_2\sqrt{\pi \log n}} n^{-\frac{c_2^2}{8\sigma^2}},$$

which converges to one as $n \to \infty$.

EC.1.2. Proof of Theorem 1

Proof of part (a). In this part, we will focus on the convergence rates of the sample eigenvectors. The key ingredient of this proof is to link the angle between the sample eigenvector and the

space spanned by population eigenvectors with the sum of inner products between the sample and population eigenvectors by the $\cos(\cdot)$ function. In this way, it suffices to show that the sum of inner products converges to one for subspace consistency, and at the same time, deriving the convergence rates by induction. To ease readability, we will finish the proof in four steps.

Step 1: Analysis of the subspace consistency. We first show that for any $i \in J_l$, $1 \le l \le m$, the subspace consistency of the sample eigenvector $\widehat{\mathbf{u}}_i$ is equivalent to

$$\sum_{j \in J_l} p_{ji}^2 \to 1, \tag{EC.1}$$

where $p_{ji} = \mathbf{u}_j^T \widehat{\mathbf{u}}_i$ is the inner product between the population eigenvector $\mathbf{u}_j$ (the jth column of $\mathbf{U}$) and $\widehat{\mathbf{u}}_i$ (the ith column of $\widehat{\mathbf{U}}$).

Since $\mathbf{U}$ and $\widehat{\mathbf{U}}$ are obtained through eigen-decomposition, we know that $\|\mathbf{u}_j\|_2 = 1$ and $\|\widehat{\mathbf{u}}_i\|_2 = 1$ for any i and j, $1 \le i, j \le q$. Note that $\sum_{j \in J_l} (\mathbf{u}_j^T \widehat{\mathbf{u}}_i)\mathbf{u}_j$ is the projection of $\widehat{\mathbf{u}}_i$ on the space $span\{\mathbf{u}_j : j \in J_l\}$. It gives

$$Angle(\widehat{\mathbf{u}}_i, span\{\mathbf{u}_j : j \in J_l\}) = \arccos\Big\{ \frac{\widehat{\mathbf{u}}_i^T [\sum_{j \in J_l} (\mathbf{u}_j^T \widehat{\mathbf{u}}_i)\mathbf{u}_j]}{\|\widehat{\mathbf{u}}_i\|_2 \cdot \|\sum_{j \in J_l} (\mathbf{u}_j^T \widehat{\mathbf{u}}_i)\mathbf{u}_j\|_2} \Big\} =$$
$$\arccos\Big\{ \frac{\sum_{j \in J_l} (\mathbf{u}_j^T \widehat{\mathbf{u}}_i)^2}{[\sum_{j \in J_l} (\mathbf{u}_j^T \widehat{\mathbf{u}}_i)^2]^{1/2}} \Big\} = \arccos\Big\{ \sqrt{\sum_{j \in J_l} (\mathbf{u}_j^T \widehat{\mathbf{u}}_i)^2} \Big\} = \arccos\Big\{ \Big(\sum_{j \in J_l} p_{ji}^2 \Big)^{1/2} \Big\}.$$

Thus, $Angle(\widehat{\mathbf{u}}_i, span\{\mathbf{u}_j : j \in J_l\}) \to 0$ is equivalent to $\sum_{j \in J_l} p_{ji}^2 \to 1$ as $q \to \infty$ for any $i \in J_l$, $1 \le l \le m$. Moreover, the convergence rate of $\sum_{j \in J_l} p_{ji}^2$ indeed provides the convergence rate of the sample eigenvector $\widehat{\mathbf{u}}_i$ to the corresponding space of population eigenvectors.

We will then prove the convergence rates by induction. Hereafter our analysis will be conditional on the event $\mathcal{E}$, which is defined in the proof of Lemma EC.1 for the consistency of the spiked sample eigenvalues and enjoys asymptotic probability one.

Step 2: Convergence rates of sample eigenvectors with indices in J_1. This step aims at proving that uniformly over $i \in J_1$, the convergence rate of $\sum_{j \in J_1} p_{ji}^2$ is given by

$$\sum_{j \in J_1} p_{ji}^2 \ge 1 - O\{ \big(\sum_{l=2}^{m} k_l q^{\alpha_l} + k_{m+1} \big) K^{-1} q^{\alpha - \alpha_1} \} = 1 - O\{A(1)\}, \tag{EC.2}$$

where $A(t)=\big(\sum_{l=t+1}^{m}k_lq^{\alpha_l}+k_{m+1}\big)K^{-1}q^{\alpha-\alpha_t}$ is defined in Theorem 1. It is also the first part of induction. Let $\mathbf{P}=\mathbf{U}^T\widehat{\mathbf{U}}=\{p_{ij}\}_{1\le i,j\le q}$. We have $\sum_{j=1}^{q}p_{ji}^2=1$ for any i since $\mathbf{P}$ is a unitary matrix. To prove (EC.2), it suffices to show

$$\sum_{j\in J_2\cup\cdots\cup J_{m+1}}p_{ji}^2\le O\{A(1)\}.$$

Recall that $\mathbf{Z}=\mathbf{\Lambda}^{-1/2}\mathbf{U}^T\mathbf{W}^T$, $\mathbf{S}=n^{-1}\mathbf{W}^T\mathbf{W}=\widehat{\mathbf{U}}\widehat{\mathbf{\Lambda}}\widehat{\mathbf{U}}^T$. Therefore, we get a connection between $\mathbf{Z}$ and $\mathbf{P}$ that

$$n^{-1}\mathbf{Z}\mathbf{Z}^T=n^{-1}\mathbf{\Lambda}^{-1/2}\mathbf{U}^T\mathbf{W}^T\mathbf{W}\mathbf{U}\mathbf{\Lambda}^{-1/2}=\mathbf{\Lambda}^{-1/2}\mathbf{P}\widehat{\mathbf{\Lambda}}\mathbf{P}^T\mathbf{\Lambda}^{-1/2}.$$

For any j, $1\le j\le q$, in view of the (j,j)th entry, the above equality gives

$$\lambda_j^{-1}\sum_{i=1}^{q}\widehat{\lambda}_ip_{ji}^2=n^{-1}\mathbf{z}_j^T\mathbf{z}_j, \tag{EC.3}$$

where $\mathbf{z}_j$ is the jth column vector of $\mathbf{Z}^T$. It implies for any i, $1\le i\le q$, $\lambda_j^{-1}\widehat{\lambda}_ip_{ji}^2\le n^{-1}\mathbf{z}_j^T\mathbf{z}_j$. Based on this fact, we have

$$\sum_{j\in J_2\cup\cdots\cup J_{m+1}}p_{ji}^2\le\sum_{j\in J_2\cup\cdots\cup J_{m+1}}n^{-1}\mathbf{z}_j^T\mathbf{z}_j\lambda_j/\widehat{\lambda}_i=\sum_{t=1}^{n}\sum_{j\in J_2\cup\cdots\cup J_{m+1}}z_{jt}^2\lambda_j/(n\widehat{\lambda}_i), \tag{EC.4}$$

where z_{jt} is the (j,t)th entry of $\mathbf{Z}$. Conditional on the event $\mathcal{E}$, by Lemma EC.1, Conditions 1 and 2, we have

$$\begin{aligned}\sum_{t=1}^{n}\sum_{j\in J_2\cup\cdots\cup J_{m+1}}z_{jt}^2\lambda_j/(n\widehat{\lambda}_i)&\le\sum_{j\in J_2\cup\cdots\cup J_{m+1}}K^{-1}q^{\alpha}\lambda_j/\widehat{\lambda}_i\\&=O\{K^{-1}q^{\alpha}C\big(\sum_{l=2}^{m}k_lq^{\alpha_l}+k_{m+1}\big)/q^{\alpha_1}\}=O\{A(1)\}.\end{aligned} \tag{EC.5}$$

Since the convergences of $\widehat{\lambda}_i$ are uniform over $i\in J_1$ by Lemma EC.1, the above inequality holds uniformly over $i\in J_1$. Inequalities (EC.4) and (EC.5) together entail $\sum_{j\in J_2\cup\cdots\cup J_{m+1}}p_{ji}^2\le O\{A(1)\}$ uniformly over $i\in J_1$, which implies the convergence rate in (EC.2) for the sample eigenvectors with indices in J_1. It shows that when $s=1$, the convergence rate coincides with our claim that uniformly over $i\in J_l$, $1\le l\le s$,

$$\sum_{j\in J_l}p_{ji}^2\ge 1-\sum_{t=1}^{l-1}\big[\prod_{i=t+1}^{l-1}(1+k_i)\big]O\{k_tA(t)\}-O\{A(l)\}. \tag{EC.6}$$

Note that we define $\sum_{t=a}^{b} s_t = 0$ and $\prod_{t=a}^{b} s_t = 1$ if $b < a$ for any positive sequence $\{s_t\}$.

Step 3: Convergence rates of sample eigenvectors with indices in J_2. Before formally completing the proof by induction, we would like to derive the convergence rates of $\sum_{j\in J_2} p_{ji}^2$ directly for $i \in J_2$ to get the basic idea of induction.

Since we already proved the convergence rate in (EC.2) uniformly over $i \in J_1$ in **Step 2**, summing over $i \in J_1$ gives

$$\sum_{i\in J_1}\sum_{j\in J_1} p_{ji}^2 \geq k_1(1 - O\{A(1)\}) = k_1 - O\{k_1 A(1)\}. \tag{EC.7}$$

Along with the fact that $\sum_{i=1}^{q} p_{ji}^2 = 1$, we get

$$\begin{aligned}\sum_{i\in J_2\cup\cdots\cup J_{m+1}}\sum_{j\in J_1} p_{ji}^2 &= \sum_{i=1}^{q}\sum_{j\in J_1} p_{ji}^2 - \sum_{i\in J_1}\sum_{j\in J_1} p_{ji}^2 = \sum_{j\in J_1}\sum_{i=1}^{q} p_{ji}^2 - \sum_{i\in J_1}\sum_{j\in J_1} p_{ji}^2 \\ &= k_1 - \sum_{i\in J_1}\sum_{j\in J_1} p_{ji}^2 \leq k_1 - (k_1 - O\{k_1 A(1)\}) = O\{k_1 A(1)\}.\end{aligned} \tag{EC.8}$$

The above result is important as it also implies that uniformly over $i \in J_2$,

$$\sum_{j\in J_1} p_{ji}^2 \leq O\{k_1 A(1)\}. \tag{EC.9}$$

For the sample eigenvector $\widehat{u}_i$ with index $i \in J_2$, in order to find a lower bound for $\sum_{j\in J_2} p_{ji}^2$, we write it as

$$\sum_{j\in J_2} p_{ji}^2 = 1 \; - \sum_{j\in J_1} p_{ji}^2 \; - \sum_{j\in J_3\cup\cdots\cup J_{m+1}} p_{ji}^2. \tag{EC.10}$$

The upper bound of $\sum_{j\in J_1} p_{ji}^2$ was provided in (EC.9). For the second term $\sum_{j\in J_3\cup\cdots\cup J_{m+1}} p_{ji}^2$, similar to (EC.4) and (EC.5) in **Step 2**, by Lemma EC.1, Conditions 1 and 2, we have uniformly over $i \in J_2$,

$$\sum_{j\in J_3\cup\cdots\cup J_{m+1}} p_{ji}^2 \leq O\{K^{-1}q^{\alpha}C\big(\sum_{l=3}^{m} k_l q^{\alpha_l} + k_{m+1}\big)/q^{\alpha_2}\} = O\{A(2)\}.$$

Plugging the above two bounds into (EC.10) gives

$$\sum_{j\in J_2} p_{ji}^2 \geq 1 - O\{k_1 A(1)\} - O\{A(2)\},$$

which shows that the uniform convergence rate of the sample eigenvectors $\widehat{u}_i$ over $i \in J_2$. Together with the uniform convergence rate over $i \in J_1$ established in **Step 2**, our claim in (EC.6) gives the uniform convergence rates of the sample eigenvectors $\widehat{u}_i$ over $i \in J_1 \cup J_2$.

Step 4: Convergence rates of sample eigenvectors with indices in J_3 to J_m. In this step, we will complete the proof by induction. Specifically, we show that the claim in (EC.6) holds for any fixed s, $3 \le s \le m$, based on the induction assumption that the claim holds for $s-1$.

By the induction assumption, we have uniformly over $i \in J_l$, $1 \le l \le s-1$,

$$\sum_{j \in J_l} p_{ji}^2 \ge 1 - \sum_{t=1}^{l-1} \Big[\prod_{i=t+1}^{l-1} (1+k_i) \Big] O\{k_t A(t)\} - O\{A(l)\}.$$

By a similar argument as in (EC.7) and (EC.8), it follows that

$$\sum_{i \in J_{l+1} \cup \cdots \cup J_{m+1}} \sum_{j \in J_l} p_{ji}^2 \le k_l \Big(\sum_{t=1}^{l-1} \Big[\prod_{i=t+1}^{l-1} (1+k_i) \Big] O\{k_t A(t)\} + O\{A(l)\} \Big). \tag{EC.11}$$

Similarly as in **Step 3**, for any $i \in J_s$, to get the convergence rate of $\sum_{j \in J_s} p_{ji}^2$, we write it as

$$\sum_{j \in J_s} p_{ji}^2 = 1 \quad - \sum_{j \in J_1 \cup \cdots \cup J_{s-1}} p_{ji}^2 \quad - \sum_{j \in J_{s+1} \cup \cdots \cup J_{m+1}} p_{ji}^2.$$

We will first derive the convergence rate of $\sum_{j \in J_1 \cup \cdots \cup J_{s-1}} p_{ji}^2$. When $1 \le l \le s-1$, we have $i \in J_s \subset J_{l+1} \cup \cdots \cup J_{m+1}$. In view of (EC.11), it gives that uniformly over $i \in J_s$,

$$\sum_{j \in J_l} p_{ji}^2 \le k_l \Big(\sum_{t=1}^{l-1} \Big[\prod_{i=t+1}^{l-1} (1+k_i) \Big] O\{k_t A(t)\} + O\{A(l)\} \Big).$$

Summing over $l = 1, \ldots, s-1$, we get

$$\sum_{j \in J_1 \cup \cdots \cup J_{s-1}} p_{ji}^2 \le \sum_{l=1}^{s-1} k_l \Big(\sum_{t=1}^{l-1} \Big[\prod_{i=t+1}^{l-1} (1+k_i) \Big] O\{k_t A(t)\} + O\{A(l)\} \Big).$$

To simplify the above expression, exchanging the summation order with respect to l and t gives

$$\begin{aligned} \sum_{j \in J_1 \cup \cdots \cup J_{s-1}} p_{ji}^2 &\le \sum_{l=1}^{s-1} \sum_{t=1}^{l-1} k_l \Big[\prod_{i=t+1}^{l-1} (1+k_i) \Big] O\{k_t A(t)\} + \sum_{l=1}^{s-1} O\{k_l A(l)\} \\ &= \sum_{t=1}^{s-2} \sum_{l=t+1}^{s-1} k_l \Big[\prod_{i=t+1}^{l-1} (1+k_i) \Big] O\{k_t A(t)\} + \sum_{t=1}^{s-1} O\{k_t A(t)\}. \end{aligned}$$

Then we combine the coefficients of $k_t A(t)$ to get

$$\sum_{j\in J_1\cup\cdots\cup J_{s-1}} p_{ji}^2 \le \sum_{t=1}^{s-2} O\{k_t A(t)\}\big(1+\sum_{l=t+1}^{s-1} k_l\big[\prod_{i=t+1}^{l-1}(1+k_i)\big]\big)+\sum_{t=s-1} O\{k_t A(t)\}.$$

Since it is immediate to conclude by induction that

$$\begin{aligned}1+\sum_{l=t+1}^{s-1} k_l\big[\prod_{i=t+1}^{l-1}(1+k_i)\big] &= 1+k_{t+1}+k_{t+2}(1+k_{t+1})+\cdots\\ &+k_{s-1}(1+k_{s-2})(1+k_{s-3})\cdots(1+k_{t+2})(1+k_{t+1})=\prod_{i=t+1}^{s-1}(1+k_i),\end{aligned}$$

we then have

$$\begin{aligned}\sum_{j\in J_1\cup\cdots\cup J_{s-1}} p_{ji}^2 &\le \sum_{t=1}^{s-2}\big[\prod_{i=t+1}^{s-1}(1+k_i)\big]O\{k_t A(t)\}+\sum_{t=s-1} O\{k_t A(t)\}\\ &=\sum_{t=1}^{s-1}\big[\prod_{i=t+1}^{s-1}(1+k_i)\big]O\{k_t A(t)\}.\end{aligned}$$

On the other hand, similar to (EC.4) and (EC.5) in **Step 2**, we have uniformly over $i\in J_s$,

$$\sum_{j\in J_{s+1}\cup\cdots\cup J_{m+1}} p_{ji}^2 \le O\{A(s)\}.$$

Combining the above two bounds gives the convergence rate of $\sum_{j\in J_s} p_{ji}^2$ uniformly over $i\in J_s$ as

$$\sum_{j\in J_s} p_{ji}^2 \ge 1-\sum_{t=1}^{s-1}\big[\prod_{i=t+1}^{s-1}(1+k_i)\big]O\{k_t A(t)\}-O\{A(s)\}.$$

Together with the induction assumption that our claim in (EC.6) holds uniformly over $i\in J_l$, $1\le l\le s-1$, we know that the claim also holds uniformly over $i\in J_l$, $1\le l\le s$. Therefore, by induction, the results in part (a) of Theorem 1 hold uniformly over $i\in J_l$, $1\le l\le m$.

Proof of part (b). In this part, we will show that when each group of spiked eigenvalues has size one (that is, $k_l=1$ for any l, $1\le l\le m$), the convergence rates of the angles between the sample score vectors $\mathbf{W}\widehat{\mathbf{u}}_i$ and the population score vectors $\mathbf{W}\mathbf{u}_i$, $1\le i\le K$, are at least as fast as those of the angles between the corresponding sample and population eigenvectors established in part (a) of Theorem 1. The key idea is to conduct delicate analysis on the $\cos(\cdot)$ function of the

angles between the sample score vectors and population score vectors, where some results about the sample eigenvalues derived in the proof of Lemma EC.1 will be used.

When each group has size one, we have $K = m$ and the convergence rates of $\widehat{\mathbf{u}}_i$ $(i \in J_l)$ to the space $span\{\mathbf{u}_j : j \in J_l\}$ become the convergence rates of $\widehat{\mathbf{u}}_i$ to $\mathbf{u}_i$, $1 \le i \le K$. Denote by $\theta_{ii} = Angle(\widehat{\mathbf{u}}_i, \mathbf{u}_i)$ and $\omega_{ii} = Angle(\mathbf{W}\widehat{\mathbf{u}}_i, \mathbf{W}\mathbf{u}_i)$. Then the results in part (a) give that uniformly over $1 \le i \le K$,

$$\cos^2(\theta_{ii}) = p_{ii}^2 \ge 1 - \sum_{t=1}^{i-1} 2^{i-t-1} O\{A(t)\} - O\{A(i)\}. \tag{EC.12}$$

Since $\mathbf{S} = n^{-1}\mathbf{W}^T\mathbf{W} = \widehat{\mathbf{U}}\widehat{\mathbf{\Lambda}}\widehat{\mathbf{U}}^T$, $\widehat{\mathbf{u}}_i$ would be the eigenvector of $\mathbf{W}^T\mathbf{W}$ corresponding to the eigenvalue $n\widehat{\lambda}_i$ with L_2-norm 1. It follows that

$$\cos(\omega_{ii}) = \frac{(\mathbf{W}\mathbf{u}_i)^T\mathbf{W}\widehat{\mathbf{u}}_i}{\|\mathbf{W}\mathbf{u}_i\|_2\|\mathbf{W}\widehat{\mathbf{u}}_i\|_2} = \frac{n\widehat{\lambda}_i\mathbf{u}_i^T\widehat{\mathbf{u}}_i}{\sqrt{n\widehat{\lambda}_i}\|\mathbf{W}\mathbf{u}_i\|_2} = \frac{\sqrt{n\widehat{\lambda}_i}\cos(\theta_{ii})}{\|\mathbf{W}\mathbf{u}_i\|_2}.$$

Squaring both sides above gives

$$\cos^2(\omega_{ii}) = \frac{n\widehat{\lambda}_i\cos^2(\theta_{ii})}{\|\mathbf{W}\mathbf{u}_i\|_2^2}. \tag{EC.13}$$

Therefore, it suffices to show $\|\mathbf{W}\mathbf{u}_i\|_2^2 \le n\widehat{\lambda}_i$.

For the term $\|\mathbf{W}\mathbf{u}_i\|_2^2$, it follows from $\mathbf{W}^T\mathbf{W} = n\widehat{\mathbf{U}}\widehat{\mathbf{\Lambda}}\widehat{\mathbf{U}}^T$ that

$$\|\mathbf{W}\mathbf{u}_i\|_2^2 = \mathbf{u}_i^T\mathbf{W}^T\mathbf{W}\mathbf{u}_i = n\mathbf{u}_i^T\widehat{\mathbf{U}}\widehat{\mathbf{\Lambda}}\widehat{\mathbf{U}}^T\mathbf{u}_i = n\sum_{j=1}^{q}\widehat{\lambda}_j(\mathbf{u}_i^T\widehat{\mathbf{u}}_j)^2 = n\sum_{j=1}^{q}\widehat{\lambda}_j p_{ij}^2.$$

By further making use of equality (EC.3), we have

$$\|\mathbf{W}\mathbf{u}_i\|_2^2 = n\sum_{j=1}^{q}\widehat{\lambda}_j p_{ij}^2 = \lambda_i\mathbf{z}_i^T\mathbf{z}_i,$$

where $\mathbf{z}_i$ is the ith column vector of $\mathbf{Z}^T$. On the other hand, inequality (EC.33) in the proof of Lemma EC.1 gives a lower bound for the sample eigenvalues $\widehat{\lambda}_i$, $1 \le i \le K$. Under the current setting that each group has size one, it gives

$$\widehat{\lambda}_i \ge \varphi_1(n^{-1}\lambda_i\mathbf{z}_i\mathbf{z}_i^T) = \varphi_1(n^{-1}\lambda_i\mathbf{z}_i^T\mathbf{z}_i) = n^{-1}\lambda_i\mathbf{z}_i^T\mathbf{z}_i,$$

where $\varphi_1(\cdot)$ denotes the largest eigenvalue of a given matrix. It follows that

$$n\widehat{\lambda}_i \ge \lambda_i\mathbf{z}_i^T\mathbf{z}_i = \|\mathbf{W}\mathbf{u}_i\|_2^2.$$

Therefore, in view of (EC.13), we get

$$\cos^2(\omega_{ii}) \geq \cos^2(\theta_{ii}),$$

which means that the convergence rate of the sample score vector is at least as good as that of the corresponding sample eigenvector. Then it follows from (EC.12) that uniformly over $1 \leq i \leq K$,

$$\cos^2(\omega_{ii}) \geq 1 - \sum_{t=1}^{i-1} 2^{i-t-1} O\{A(t)\} - O\{A(i)\},$$

which completes the proof of part (b) of Theorem 1.

EC.1.3. Proof of Proposition 1

By Condition 4, the inequality $\|n^{-1/2}(\mathbf{X}, \mathbf{F})\boldsymbol{\delta}\|_2 \geq c\|\boldsymbol{\delta}\|_2$ holds for any $\boldsymbol{\delta}$ satisfying $\|\boldsymbol{\delta}\|_0 < M$ with significant probability $1 - \theta_{n,p}$. We now derive a similar result for $(\mathbf{X}, \widehat{\mathbf{F}})$ by analyzing the estimation errors of confounding factors $\mathbf{F}$.

By the estimation error bound in Condition 3, we have for any $1 \leq j \leq K$,

$$\begin{aligned}\|\mathbf{f}_j - \widehat{\mathbf{f}}_j\|_2^2 &\leq \|\mathbf{f}_j\|_2^2 + \|\widehat{\mathbf{f}}_j\|_2^2 - 2\ \mathbf{f}_j'\widehat{\mathbf{f}}_j = n + n - 2\|\mathbf{f}_j\|_2\|\widehat{\mathbf{f}}_j\|_2 \cos(\omega_{jj}) \\ &= 2n - 2n\cos(\omega_{jj}) = 2n\{1 - \cos(\omega_{jj})\} \leq \frac{c_2^2 \log n}{4K^2\rho^2}.\end{aligned}$$

Since the above bound does not vary with the index j, it gives the uniform confounding factor estimation error bound

$$\max_{1\leq j\leq K} \|\mathbf{f}_j - \widehat{\mathbf{f}}_j\|_2 \leq \frac{c_2}{2K\rho}\sqrt{\log n}. \tag{EC.14}$$

Now we proceed to prove the inequality for $(\mathbf{X}, \widehat{\mathbf{F}})$. First of all, it follows from Condition 4 and the triangular inequality that

$$\begin{aligned}\|n^{-1/2}(\mathbf{X}, \widehat{\mathbf{F}})\boldsymbol{\delta}\|_2 &\geq \|n^{-1/2}(\mathbf{X}, \mathbf{F})\boldsymbol{\delta}\|_2 - \|n^{-1/2}(\mathbf{X}, \mathbf{F})\boldsymbol{\delta} - n^{-1/2}(\mathbf{X}, \widehat{\mathbf{F}})\boldsymbol{\delta}\|_2 \\ &\geq c\|\boldsymbol{\delta}\|_2 - n^{-1/2}\|(\mathbf{F} - \widehat{\mathbf{F}})\boldsymbol{\delta}_1\|_2 \geq c\|\boldsymbol{\delta}\|_2 - n^{-1/2} \max_{1\leq j\leq K} \|(\mathbf{f}_j - \widehat{\mathbf{f}}_j)\|_2 \|\boldsymbol{\delta}_1\|_1,\end{aligned}$$

where $\boldsymbol{\delta}_1$ is a subvector of $\boldsymbol{\delta}$ consisting of the last K components. Note that $\|\boldsymbol{\delta}_1\|_1 \leq \sqrt{K}\|\boldsymbol{\delta}_1\|_2 \leq \sqrt{K}\|\boldsymbol{\delta}\|_2$. Further applying inequality (EC.14) yields

$$\|n^{-1/2}(\mathbf{X}, \widehat{\mathbf{F}})\boldsymbol{\delta}\|_2 \geq c\|\boldsymbol{\delta}\|_2 - n^{-1/2} \cdot \frac{c_2}{2K\rho}\sqrt{\log n} \cdot \sqrt{K}\|\boldsymbol{\delta}\|_2 \geq c_1\|\boldsymbol{\delta}\|_2,$$

where c_1 is some positive constant no larger than $c - \frac{c_2}{2\rho}\sqrt{\frac{\log n}{nK}}$. It is clear that c_1 is smaller than but close to c when n is relatively large. In view of the tail probabilities in Conditions 3 and 4, the above inequality holds with probability at least $1-\theta_1-\theta_2$. Thus, we finish the proof of Proposition 1.

EC.1.4. Proof of Theorem 2

With Proposition 1, we will apply a similar idea as in Zheng et al. (2014) to prove the global properties. The proof consists of two parts. The first part shows the model selection consistency property with the range of λ given in Theorem 2. Based on the first part, several oracle inequalities will then be induced. We will first prove the properties when the columns of design matrix $\mathbf{X}$ have a common scale of L_2-norm $n^{1/2}$ as a benchmark, meaning that $\boldsymbol{\beta}_* = \boldsymbol{\beta}$ and $L = 1$, and then illustrate the results in general cases.

Part 1: Model selection consistency. This part contains two steps. In the first step, it will be shown that when $c_1^{-1}c_2\sqrt{(2s+1)(\log p)/n} < \lambda < b_0$, the number of nonzero elements in $(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T$ is no larger than s conditioning on the event $\widetilde{\mathcal{E}}$ defined in Lemma EC.2. We prove this by using the global optimality of $(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T$.

By the hard-thresholding property (Zheng et al. 2014, Lemma 1) and $\lambda < b_0$, any nonzero component of the true regression coefficient vector $(\boldsymbol{\beta}_0^T, \boldsymbol{\gamma}_0^T)^T$ or of the global minimizer $(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T$ is greater than λ, which ensures $\|p_\lambda\{(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T\}\|_1 = \lambda^2\|(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T\|_0/2$ and $\|p_\lambda\{(\boldsymbol{\beta}_0^T, \boldsymbol{\gamma}_0^T)^T\}\|_1$ $= s\lambda^2/2$. Thus, we get

$$\left\|p_\lambda\left\{(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T\right\}\right\|_1 - \left\|p_\lambda\left\{(\boldsymbol{\beta}_0^T, \boldsymbol{\gamma}_0^T)^T\right\}\right\|_1 = \left\{\|(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T\|_0 - s\right\}\lambda^2/2.$$

Denote by $\boldsymbol{\delta} = (\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T - (\boldsymbol{\beta}_0^T, \boldsymbol{\gamma}_0^T)^T$. Direct calculation yields

$$\begin{aligned} Q\left\{(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T\right\} - Q\left\{(\boldsymbol{\beta}_0^T, \boldsymbol{\gamma}_0^T)^T\right\} &= 2^{-1}\left\|n^{-\frac{1}{2}}(\mathbf{X}, \widehat{\mathbf{F}})\boldsymbol{\delta}\right\|_2^2 - n^{-1}\widetilde{\boldsymbol{\varepsilon}}^T(\mathbf{X}, \widehat{\mathbf{F}})\boldsymbol{\delta} \\ &\quad + \left\{\|(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T\|_0 - s\right\}\lambda^2/2, \end{aligned} \tag{EC.15}$$

where $\widetilde{\boldsymbol{\varepsilon}} = (\mathbf{F} - \widehat{\mathbf{F}})\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$, the sum of the random error vector $\boldsymbol{\varepsilon}$ and estimation errors $(\mathbf{F} - \widehat{\mathbf{F}})\boldsymbol{\gamma}$.

On the other hand, conditional on event $\widetilde{\mathcal{E}}$, we have

$$\begin{aligned} |n^{-1}\widetilde{\boldsymbol{\varepsilon}}^T(\mathbf{X}, \widehat{\mathbf{F}})\boldsymbol{\delta}| \;&\le\; \|n^{-1}\widetilde{\boldsymbol{\varepsilon}}^T(\mathbf{X}, \widehat{\mathbf{F}})\|_\infty \|\boldsymbol{\delta}\|_1 \\ &\le\; c_2\sqrt{(\log p)/n}\|\boldsymbol{\delta}\|_1 \;\le\; c_2\sqrt{(\log p)/n}\|\boldsymbol{\delta}\|_0^{\frac{1}{2}}\|\boldsymbol{\delta}\|_2. \end{aligned} \tag{EC.16}$$

In addition, by Condition 6 and the definition of $\mathbb{S}_{M/2}$, we obtain $\|\boldsymbol{\delta}\|_0 \le \|(\boldsymbol{\beta}_0^T, \boldsymbol{\gamma}_0^T)^T\|_0 + \|(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T\|_0 < M$, where M is the robust spark of $(\mathbf{X}, \widehat{\mathbf{F}})$ with bound c_1 by Proposition 1. Thus, we have

$$\|n^{-\frac{1}{2}}(\mathbf{X}, \widehat{\mathbf{F}})\boldsymbol{\delta}\|_2 \ge c_1\|\boldsymbol{\delta}\|_2. \tag{EC.17}$$

Plugging inequalities (EC.16) and (EC.17) into (EC.15) gives that

$$\begin{aligned} Q\left\{(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T\right\} - Q\left\{(\boldsymbol{\beta}_0^T, \boldsymbol{\gamma}_0^T)^T\right\} \ge 2^{-1}c_1^2\|\boldsymbol{\delta}\|_2^2 - \; & c_2\sqrt{(\log p)/n}\|\boldsymbol{\delta}\|_0^{\frac{1}{2}}\|\boldsymbol{\delta}\|_2 \\ & + \left\{\|(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T\|_0 - s\right\}\lambda^2/2. \end{aligned} \tag{EC.18}$$

Thus, the global optimality of $(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T$ ensures that

$$2^{-1}c_1^2\|\boldsymbol{\delta}\|_2^2 - c_2\sqrt{\frac{\log p}{n}}\|\boldsymbol{\delta}\|_0^{\frac{1}{2}}\|\boldsymbol{\delta}\|_2 + \left\{\|(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T\|_0 - s\right\}\lambda^2/2 \le 0.$$

After completing the squares in the above inequality, we get

$$\left[c_1\|\boldsymbol{\delta}\|_2 - \frac{c_2}{c_1}\sqrt{\frac{\log p}{n}}\|\boldsymbol{\delta}\|_0^{\frac{1}{2}}\right]^2 - \left(\frac{c_2}{c_1}\right)^2\frac{\log p}{n}\|\boldsymbol{\delta}\|_0 + \left\{\|(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T\|_0 - s\right\}\lambda^2 \le 0.$$

Since $\left[c_1\|\boldsymbol{\delta}\|_2 - \frac{c_2}{c_1}\sqrt{\frac{\log p}{n}}\|\boldsymbol{\delta}\|_0^{\frac{1}{2}}\right]^2 \ge 0$, it gives

$$\left\{\|(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T\|_0 - s\right\}\lambda^2 \le \left(\frac{c_2}{c_1}\right)^2\frac{\log p}{n}\|\boldsymbol{\delta}\|_0. \tag{EC.19}$$

We continue to bound the value of $\|(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T\|_0$ by the above inequality. Let $k = \|(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T\|_0$. Then $\|\boldsymbol{\delta}\|_0 = \|(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T - (\boldsymbol{\beta}_0^T, \boldsymbol{\gamma}_0^T)^T\|_0 \le k + s$. Thus, it follows from (EC.19) that

$$(k - s)\lambda^2 \le \left(\frac{c_2}{c_1}\right)^2\frac{\log p}{n}(k + s).$$

Organizing it in terms of k and s, we get

$$k\left(\lambda^2-\left(\frac{c_2}{c_1}\right)^2\frac{\log p}{n}\right)\le s\left(\lambda^2+\left(\frac{c_2}{c_1}\right)^2\frac{\log p}{n}\right). \tag{EC.20}$$

Since $\lambda > c_1^{-1}c_2\sqrt{(2s+1)\log p/n}$, we have $\lambda^2-(c_1^{-1}c_2)^2(2s+1)\frac{\log p}{n}>0$ and $\lambda^2c_1^2n-c_2^2\log p > 2c_2^2 s\log p$. Thus we have $\frac{2c_2^2\log p}{\lambda^2c_1^2n-c_2^2\log p}<1/s$. Then it follows from inequality (EC.20) that

$$k\le s\frac{(\lambda^2+(\frac{c_2}{c_1})^2\frac{\log p}{n})}{(\lambda^2-(\frac{c_2}{c_1})^2\frac{\log p}{n})}=s\left(1+\frac{2c_2^2\log p}{\lambda^2c_1^2n-c_2^2\log p}\right)<s+1.$$

Therefore, the number of nonzero elements in $(\widehat{\boldsymbol{\beta}}^T,\widehat{\boldsymbol{\gamma}}^T)^T$ satisfies

$$\left\|(\widehat{\boldsymbol{\beta}}^T,\widehat{\boldsymbol{\gamma}}^T)^T\right\|_0\le s.$$

The second step is based on the first step, where we will use proof by contradiction to show that $\mathrm{supp}((\boldsymbol{\beta}_0^T,\boldsymbol{\gamma}_0^T)^T)\subset\mathrm{supp}((\widehat{\boldsymbol{\beta}}^T,\widehat{\boldsymbol{\gamma}}^T)^T)$ with the additional assumption $\lambda<b_0c_1/\sqrt{2}$ in the theorem. Suppose that $\mathrm{supp}((\boldsymbol{\beta}_0^T,\boldsymbol{\gamma}_0^T)^T)\not\subset\mathrm{supp}(\widehat{\boldsymbol{\beta}}^T,\widehat{\boldsymbol{\gamma}}^T)^T$, and we denote the number of missed true coefficients as

$$k=\left|\mathrm{supp}\{(\boldsymbol{\beta}_0^T,\boldsymbol{\gamma}_0^T)^T\}\backslash\mathrm{supp}\{(\widehat{\boldsymbol{\beta}}^T,\widehat{\boldsymbol{\gamma}}^T)^T\}\right|\ge 1.$$

Then we have $\|(\widehat{\boldsymbol{\beta}}^T,\widehat{\boldsymbol{\gamma}}^T)^T\|_0\ge s-k$ and $\|\boldsymbol{\delta}\|_0\le\|(\widehat{\boldsymbol{\beta}}^T,\widehat{\boldsymbol{\gamma}}^T)^T\|_0+\|(\boldsymbol{\beta}_0^T,\boldsymbol{\gamma}_0^T)^T\|_0\le 2s$ by the first step. Combining these two results with inequality (EC.18) yields

$$Q\left\{(\widehat{\boldsymbol{\beta}}^T,\widehat{\boldsymbol{\gamma}}^T)^T\right\}-Q\left\{(\boldsymbol{\beta}_0^T,\boldsymbol{\gamma}_0^T)^T\right\}\ge\left(2^{-1}c_1^2\|\boldsymbol{\delta}\|_2-c_2\sqrt{\frac{2s\log p}{n}}\right)\|\boldsymbol{\delta}\|_2-k\lambda^2/2. \tag{EC.21}$$

Note that for each $j\in\mathrm{supp}((\boldsymbol{\beta}_0^T,\boldsymbol{\gamma}_0^T)^T)\setminus\mathrm{supp}((\widehat{\boldsymbol{\beta}}^T,\widehat{\boldsymbol{\gamma}}^T)^T)$, we have $|\delta_j|\ge b_0$ with b_0 the lowest signal strength defined in Condition 6. Thus, $\|\boldsymbol{\delta}\|_2\ge\sqrt{k}b_0$, which together with Condition 6 entails

$$4^{-1}c_1^2\|\boldsymbol{\delta}\|_2\ge 4^{-1}c_1^2\sqrt{k}b_0\ge 4^{-1}c_1^2b_0>c_2\sqrt{(2s\log p)/n}.$$

Thus, it follows from (EC.21) that

$$Q\left\{(\widehat{\boldsymbol{\beta}}^T,\widehat{\boldsymbol{\gamma}}^T)^T\right\}-Q\left\{(\boldsymbol{\beta}_0^T,\boldsymbol{\gamma}_0^T)^T\right\}\ge 4^{-1}c_1^2\|\boldsymbol{\delta}\|_2^2-k\lambda^2/2\ge 4^{-1}c_1^2kb_0^2-k\lambda^2/2>0,$$

where the last step is because of the additional assumption $\lambda < b_0 c_1/\sqrt{2}$. The above inequality contradicts with the global optimality of $(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T$. Thus, we have $\mathrm{supp}((\boldsymbol{\beta}_0^T, \boldsymbol{\gamma}_0^T)^T) \subset \mathrm{supp}((\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T)$. Combining this with $\|(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T\|_0 \le s$ from the first step, we know that $\mathrm{supp}\{(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T\} = \mathrm{supp}\{(\boldsymbol{\beta}_0^T, \boldsymbol{\gamma}_0^T)^T\}$.

Part 2: Prediction and estimation losses. In this part, we will bound the prediction and estimation losses. The idea is to get the L_2-estimation loss bound by the global optimality of $(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T$, conditional on the event $\widetilde{\mathcal{E}} \cap \widetilde{\mathcal{E}}_0$ defined in Lemma EC.2. Then by similar techniques as in the first part, we would derive bounds for the prediction and estimation losses.

Recall that $\mathbf{X}_0$, $\widehat{\mathbf{F}}_0$ are the submatrices of $\mathbf{X}$ and $\widehat{\mathbf{F}}$ consisting of columns in $\mathrm{supp}(\boldsymbol{\beta}_0)$ and $\mathrm{supp}(\boldsymbol{\gamma}_0)$, respectively. Conditioning on $\widetilde{\mathcal{E}} \cap \widetilde{\mathcal{E}}_0$, we have $\|\boldsymbol{\delta}\|_0 \le s$ by the model selection consistency established before. Thus, applying the Cauchy-Schwarz inequality and the definition of $\widetilde{\mathcal{E}}_0$ gives

$$\begin{aligned} |n^{-1}\widetilde{\boldsymbol{\varepsilon}}^T(\mathbf{X}_0, \widehat{\mathbf{F}}_0)\boldsymbol{\delta}| \;&\le\; \|n^{-1}\widetilde{\boldsymbol{\varepsilon}}^T(\mathbf{X}_0, \widehat{\mathbf{F}}_0)\|_\infty \|\boldsymbol{\delta}\|_1 \\ &\le\; c_2\sqrt{\frac{\log n}{n}}\|\boldsymbol{\delta}\|_1 \;\le\; c_2\sqrt{\frac{s\log n}{n}}\|\boldsymbol{\delta}\|_2. \end{aligned} \tag{EC.22}$$

In views of (EC.15) and (EC.17), it follows from inequality (EC.22) and the model selection consistency property $\|(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T\|_0 = s$ that

$$\begin{aligned} &Q\left\{(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T\right\} - Q\left\{(\boldsymbol{\beta}_0^T, \boldsymbol{\gamma}_0^T)^T\right\} \\ &= 2^{-1}\|n^{-1}(\mathbf{X}, \widehat{\mathbf{F}})\boldsymbol{\delta}\|_2^2 - n^{-1}\widetilde{\boldsymbol{\varepsilon}}^T(\mathbf{X}, \widehat{\mathbf{F}})\boldsymbol{\delta} + \left\{\|(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T\|_0 - s\right\}\lambda^2/2 \\ &\ge 2^{-1}c_1^2\|\boldsymbol{\delta}\|_2^2 - n^{-1}\widetilde{\boldsymbol{\varepsilon}}^T(\mathbf{X}_0, \widehat{\mathbf{F}}_0)\boldsymbol{\delta} \ge \left(2^{-1}c_1^2\|\boldsymbol{\delta}\|_2 - c_2\sqrt{\frac{s\log n}{n}}\right)\|\boldsymbol{\delta}\|_2. \end{aligned}$$

Since $(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T$ is the global optimizer of Q, we have

$$2^{-1}c_1^2\|\boldsymbol{\delta}\|_2 - c_2\sqrt{\frac{s\log n}{n}} \le 0,$$

which gives the L_2 and L_∞ estimation loss bounds as

$$\begin{aligned} &\|(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T - (\boldsymbol{\beta}_0^T, \boldsymbol{\gamma}_0^T)^T\|_2 = \|\boldsymbol{\delta}\|_2 \le\; 2c_1^{-2}c_2\sqrt{(s\log n)/n}, \\ &\|(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T - (\boldsymbol{\beta}_0^T, \boldsymbol{\gamma}_0^T)^T\|_\infty \le \|(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T - (\boldsymbol{\beta}_0^T, \boldsymbol{\gamma}_0^T)^T\|_2 \le 2c_1^{-2}c_2\sqrt{(s\log n)/n}. \end{aligned}$$

For L_q-estimation losses with $1 \le q < 2$, applying Hölder's inequality gives

$$\begin{aligned}\|(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T - (\boldsymbol{\beta}_0^T, \boldsymbol{\gamma}_0^T)^T\|_q &= \Big(\sum_j |\delta_j|^q\Big)^{1/q} \le \Big(\sum_j |\delta_j|^2\Big)^{\frac{1}{2}} \Big(\sum_{\delta_j \ne 0} 1^{\frac{2}{2-q}}\Big)^{\frac{1}{q}-\frac{1}{2}} \\ &= \|\boldsymbol{\delta}\|_2 \|\boldsymbol{\delta}\|_0^{\frac{1}{q}-\frac{1}{2}} \le 2c_1^{-2} c_2 s^{\frac{1}{q}} \sqrt{(\log n)/n}.\end{aligned}$$

Next we prove the bound for oracle prediction loss. Since $(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T$ is the global minimizer, it follows from (EC.15) and the model selection consistency property that

$$\begin{aligned}&n^{-1/2} \|(\mathbf{X}, \widehat{\mathbf{F}})\{(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T - (\boldsymbol{\beta}_0^T, \boldsymbol{\gamma}_0^T)^T\}\|_2 \\ \le &\Big\{2n^{-1} \widetilde{\boldsymbol{\varepsilon}}^T (\mathbf{X}, \widehat{\mathbf{F}}) \boldsymbol{\delta}\Big\}^{1/2} \le \Big\{2\|n^{-1}(\mathbf{X}_0, \widehat{\mathbf{F}}_0)^T \tilde{\boldsymbol{\varepsilon}}\|_\infty \|\boldsymbol{\delta}\|_1\Big\}^{1/2} \le 2c_2 c_1^{-1} \sqrt{s(\log n)/n},\end{aligned}$$

where the last step is because of the L_1 estimation loss bound proved before. Then for the oracle prediction loss, together with (EC.34) in the proof of Lemma EC.2, it follows that

$$\begin{aligned}&n^{-1/2} \|(\mathbf{X}, \widehat{\mathbf{F}})(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T - (\mathbf{X}, \mathbf{F})(\boldsymbol{\beta}_0^T, \boldsymbol{\gamma}_0^T)^T\|_2 \\ \le\ & 2c_2 c_1^{-1} \sqrt{s(\log n)/n} + n^{-1/2} \|(\mathbf{F} - \widehat{\mathbf{F}})\boldsymbol{\gamma}_0\|_2 \le (2c_2 c_1^{-1} \sqrt{s} + c_2/2) \sqrt{(\log n)/n}.\end{aligned}$$

Last we will derive our results for general cases when the L_2-norms of columns of $\mathbf{X}$ are not of the common scale $n^{1/2}$. Note that the penalized least squares in (3) can be rewritten as

$$Q\big\{(\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T)^T\big\} = (2n)^{-1} \|\mathbf{y} - \widetilde{\mathbf{X}} \boldsymbol{\beta}_* - \widehat{\mathbf{F}} \boldsymbol{\gamma}\|_2^2 + \|p_\lambda\big\{(\boldsymbol{\beta}_*^T, \boldsymbol{\gamma}^T)^T\big\}\|_1,$$

where $\widetilde{\mathbf{X}}$ is the matrix with the L_2-norm of each column rescaled to $n^{1/2}$ and

$$\boldsymbol{\beta}_* = n^{-1/2}(\beta_1 \|\mathbf{x}_1\|_2, \ldots, \beta_p \|\mathbf{x}_p\|_2)^T$$

is the corresponding coefficient vector defined in (3). By Conditions 5 and 6, the same argument applies to derive the model selection consistency property and the bounds on oracle prediction and estimation losses for $(\widehat{\boldsymbol{\beta}}_*^T, \widehat{\boldsymbol{\gamma}}^T)^T$ since the relationship between λ and signal strength keeps the same even if $L \ne 1$. Based on Condition 5, it is clear that the model selection consistency of $\widehat{\boldsymbol{\beta}}_*$ implies

that of $\widehat{\boldsymbol{\beta}}$. And the bound on prediction loss does not change since $\widetilde{\mathbf{X}}\widehat{\boldsymbol{\beta}}_* = \mathbf{X}\widehat{\boldsymbol{\beta}}$. As for the bounds of estimation losses on $\widehat{\boldsymbol{\beta}}$, they can be deduced as

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 \leq 2c_1^{-2}c_2 L\sqrt{(s\log n)/n}, \quad \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_q \leq 2c_1^{-2}c_2 L s^{\frac{1}{q}}\sqrt{(\log n)/n},$$
$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_\infty \leq \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 \leq 2L^{-1}c_1^{-2}c_2\sqrt{(s\log n)/n}.$$

The tail probability for these results to hold is at most the sum of the tail probabilities in Conditions 3-5 and Lemma EC.2. Thus, we know that these properties hold simultaneously with probability at least

$$1 - \frac{4\sqrt{2}\sigma}{c_2\sqrt{\pi\log p}}p^{1-\frac{c_2^2}{8\sigma^2}} + \frac{2\sqrt{2}\sigma s}{c_2\sqrt{\pi\log n}}n^{-\frac{c_2^2}{8\sigma^2}} - \theta_1 - \theta_2 - \theta_3,$$

which concludes the proof of Theorem 2.

EC.2. Additional technical details

The following lemma is needed in proving Lemma EC.1.

LEMMA EC.3 **(Weyl's inequality (Horn and Johnson 1990))**. *If* $\mathbf{A}$ *and* $\mathbf{B}$ *are* $m \times m$ *real symmetric matrices, then for all* $k = 1, \ldots, m$,

$$\left.\begin{array}{c} \varphi_k(\mathbf{A}) + \varphi_m(\mathbf{B}) \\ \varphi_{k+1}(\mathbf{A}) + \varphi_{m-1}(\mathbf{B}) \\ \vdots \\ \varphi_m(\mathbf{A}) + \varphi_k(\mathbf{B}) \end{array}\right\} \leq \varphi_k(\mathbf{A}+\mathbf{B}) \leq \left\{\begin{array}{c} \varphi_k(\mathbf{A}) + \varphi_1(\mathbf{B}) \\ \varphi_{k-1}(\mathbf{A}) + \varphi_2(\mathbf{B}) \\ \vdots \\ \varphi_1(\mathbf{A}) + \varphi_k(\mathbf{B}) \end{array}\right.,$$

where $\varphi_i(\cdot)$ *is the function that takes the* i*th largest eigenvalue of a given matrix.*

EC.2.1. Proof of Lemma EC.1

The main idea of proving Lemma EC.1 is to use induction to show that the sample eigenvalues divided by their corresponding orders of q will be convergent in an event with asymptotic probability one. To ease readability, the proof is divided into three steps.

Step 1: Large probability event $\mathcal{E}$. In this step, we will define an event $\mathcal{E}$ and show that its probability approaches one when q increases to infinity. Our later discussion will be conditional on this event. Denote a series of events by $\mathcal{E}_{jt}$, $1 \le j \le q$, $1 \le t \le n$, such that

$$\mathcal{E}_{jt} = \{z_{jt}^2 \le K^{-1} q^{\alpha}\},$$

where z_{jt} is the (j,t)th entry of $\mathbf{Z}$. By Condition 2, the events $\mathcal{E}_{jt}$ satisfy a uniform tail probability bound $P(\mathcal{E}_{jt}^c) = o(q^{-1}n^{-1})$. Let $\mathcal{E} = \cap_{t=1}^{n} \cap_{j=1}^{q} \mathcal{E}_{jt}$ be the intersection of all events in the series. Then the probability of event $\mathcal{E}$ converges to one since

$$P(\mathcal{E}^c) = P(\cup_{t=1}^{n} \cup_{j=1}^{q} \mathcal{E}_{jt}^c) \le \sum_{t=1}^{n} \sum_{j=1}^{q} P(\mathcal{E}_{jt}^c) = nq \cdot o(q^{-1}n^{-1}) \to 0, \ as \ q \to \infty.$$

Step 2: Convergence of eigenvalues with indices in J_1. This is the first part of induction. We will show that conditional on event $\mathcal{E}$, uniformly over $i \in J_1$, $q^{-\alpha_1}\widehat{\lambda}_i \to c_i$, as $q \to \infty$.

Denote by $\mathbf{C}$ the $q \times q$ diagonal matrix with the first K diagonal components equaling to c_j, $1 \le j \le K$, and the rest diagonal components 1. We decompose $\mathbf{Z}, \mathbf{C}$ and $\mathbf{\Lambda}$ into block matrices according to the index sets $J_1, J_2, \ldots, J_{m+1}$ such that

$$\mathbf{Z} = \begin{pmatrix} \mathbf{Z}_1 \\ \mathbf{Z}_2 \\ \vdots \\ \mathbf{Z}_{m+1} \end{pmatrix}, \mathbf{C} = \begin{pmatrix} \mathbf{C}_1 & \mathbf{O} & \cdots & \mathbf{O} \\ \mathbf{O} & \mathbf{C}_2 & \cdots & \mathbf{O} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{O} & \mathbf{O} & \cdots & \mathbf{C}_{m+1} \end{pmatrix}, \mathbf{\Lambda} = \begin{pmatrix} \mathbf{\Lambda}_1 & \mathbf{O} & \cdots & \mathbf{O} \\ \mathbf{O} & \mathbf{\Lambda}_2 & \cdots & \mathbf{O} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{O} & \mathbf{O} & \cdots & \mathbf{\Lambda}_{m+1} \end{pmatrix}. \tag{EC.23}$$

Then for the dual matrix S_D, we have

$$\mathbf{S}_D = n^{-1}\mathbf{Z}^T\mathbf{\Lambda}\mathbf{Z} = n^{-1}\sum_{l=1}^{m+1} \mathbf{Z}_l^T \mathbf{\Lambda}_l \mathbf{Z}_l. \tag{EC.24}$$

Divided by q^{α_1} on both sides of (EC.24) gives

$$q^{-\alpha_1}\mathbf{S}_D = n^{-1}q^{-\alpha_1}\mathbf{Z}_1^T\mathbf{\Lambda}_1\mathbf{Z}_1 + n^{-1}q^{-\alpha_1}\sum_{l=2}^{m} \mathbf{Z}_l^T\mathbf{\Lambda}_l\mathbf{Z}_l + n^{-1}q^{-\alpha_1}\mathbf{Z}_{m+1}^T\mathbf{\Lambda}_{m+1}\mathbf{Z}_{m+1}. \tag{EC.25}$$

We will show the sum of the last two terms above converges to the zero matrix in Frobenius norm, where the Frobenius norm is defined as $\|\mathbf{A}\|_F = \{\mathrm{tr}(\mathbf{A}\mathbf{A}^T)\}^{1/2}$ for a given matrix $\mathbf{A}$.

For any l, $1 \le l \le m$, let $\lambda_t^{(l)}$ and $c_t^{(l)}$ be the tth diagonal elements of $\mathbf{\Lambda}_l$ and $\mathbf{C}_l$, respectively. Conditional on event $\mathcal{E}$, for any j and k, $1 \le j,k \le n$, the absolute value of the (j,k)th element in $\sum_{l=2}^m \mathbf{Z}_l^T \mathbf{\Lambda}_l \mathbf{Z}_l$ is

$$|\sum_{l=2}^{m}\sum_{t=1}^{k_l} \lambda_t^{(l)} z_{tj}^{(l)} z_{tk}^{(l)}| \le K^{-1} q^{\alpha} \sum_{l=2}^{m}\sum_{t=1}^{k_l} \lambda_t^{(l)},$$

where $z_{tj}^{(l)}$ and $z_{tk}^{(l)}$ are the (t,j)th and (t,k)th elements in $\mathbf{Z}_l$, respectively. By Condition 1, uniformly over $1 \le l \le m$ and $1 \le t \le k_l$, $\lambda_t^{(l)} = O(q^{\alpha_l} c_t^{(l)})$. Then it follows that

$$\begin{aligned}
&\|n^{-1}q^{-\alpha_1}\sum_{l=2}^{m} \mathbf{Z}_l^T \mathbf{\Lambda}_l \mathbf{Z}_l\|_F \le n^{-1}q^{-\alpha_1}(nK^{-1}q^{\alpha}\sum_{l=2}^{m}\sum_{t=1}^{k_l}\lambda_t^{(l)}) \\
&= O\{q^{-\alpha_1}K^{-1}q^{\alpha}\sum_{l=2}^{m}\sum_{t=1}^{k_l} q^{\alpha_l} c_t^{(l)}\} = O\{K^{-1}q^{\alpha}\sum_{l=2}^{m} k_l C q^{\alpha_l}/q^{\alpha_1}\}.
\end{aligned}$$

Similarly we would get

$$\|n^{-1}q^{-\alpha_1}\mathbf{Z}_{m+1}^T \mathbf{\Lambda}_{m+1}\mathbf{Z}_{m+1}\|_F \le K^{-1}q^{\alpha}k_{m+1}C/q^{\alpha_1}.$$

Together with $\alpha < \min\{\Delta, \alpha_m - 1\}$ by Condition 2 and $k_{m+1} < q$, we have

$$\begin{aligned}
&\|n^{-1}q^{-\alpha_1}\sum_{l=2}^{m}\mathbf{Z}_l^T\mathbf{\Lambda}_l\mathbf{Z}_l + n^{-1}q^{-\alpha_1}\mathbf{Z}_{m+1}^T\mathbf{\Lambda}_{m+1}\mathbf{Z}_{m+1}\|_F \\
&\le O\{(\sum_{l=2}^{m} k_l q^{\alpha_l} + k_{m+1})K^{-1}q^{\alpha}C/q^{\alpha_1}\} \to 0, \ as\ q\to\infty. \qquad \text{(EC.26)}
\end{aligned}$$

By a similar argument, under Condition 1, we have

$$\begin{aligned}
&\|n^{-1}q^{-\alpha_1}\mathbf{Z}_1^T\mathbf{\Lambda}_1\mathbf{Z}_1 - n^{-1}\mathbf{Z}_1^T\mathbf{C}_1\mathbf{Z}_1\|_F = \|n^{-1}\mathbf{Z}_1^T(q^{-\alpha_1}\mathbf{\Lambda}_1 - \mathbf{C}_1)\mathbf{Z}_1\|_F \\
&\le n^{-1}[nK^{-1}q^{\alpha}\sum_{t=1}^{k_1}(q^{-\alpha_1}\lambda_t^{(1)} - c_t^{(1)})] \le k_1 K^{-1}q^{\alpha}\cdot O(q^{-\Delta}) \to 0, \ as\ q\to\infty. \qquad \text{(EC.27)}
\end{aligned}$$

In view of (EC.25), it is immediate that

$$\begin{aligned}
&\|q^{-\alpha_1}\mathbf{S}_D - n^{-1}\mathbf{Z}_1^T\mathbf{C}_1\mathbf{Z}_1\|_F \le \|n^{-1}q^{-\alpha_1}\mathbf{Z}_1^T\mathbf{\Lambda}_1\mathbf{Z}_1 - n^{-1}\mathbf{Z}_1^T\mathbf{C}_1\mathbf{Z}_1\|_F \\
&+ \|n^{-1}q^{-\alpha_1}\sum_{l=2}^{m}\mathbf{Z}_l^T\mathbf{\Lambda}_l\mathbf{Z}_l + n^{-1}q^{-\alpha_1}\mathbf{Z}_{m+1}^T\mathbf{\Lambda}_{m+1}\mathbf{Z}_{m+1}\|_F \to 0, \ as\ q\to\infty. \qquad \text{(EC.28)}
\end{aligned}$$

Further applying (Horn and Johnson 1990, Corollary 6.3.8) gives as $q \to \infty$,

$$\max_{1\le i\le n} |\varphi_i(q^{-\alpha_1}\mathbf{S}_D) - \varphi_i(n^{-1}\mathbf{Z}_1^T\mathbf{C}_1\mathbf{Z}_1)| \le \|q^{-\alpha_1}\mathbf{S}_D - n^{-1}\mathbf{Z}_1^T\mathbf{C}_1\mathbf{Z}_1\|_F \to 0. \tag{EC.29}$$

Note that $n^{-1}\mathbf{Z}_1^T\mathbf{C}_1\mathbf{Z}_1$ shares the same nonzero eigenvalues with its due matrix, that is, $n^{-1}\mathbf{C}_1^{1/2}\mathbf{Z}_1\mathbf{Z}_1^T\mathbf{C}_1^{1/2}$ of dimensionality k_1. It follows from (EC.29) that

$$\max_{i\in J_1} |\varphi_i(q^{-\alpha_1}\mathbf{S}_D) - \varphi_i(n^{-1}\mathbf{C}_1^{1/2}\mathbf{Z}_1\mathbf{Z}_1^T\mathbf{C}_1^{1/2})| \to 0. \tag{EC.30}$$

Moreover, by part (b) of Condition 2, we have

$$\max_{i\in J_1} |\varphi_i(n^{-1}\mathbf{C}_1^{1/2}\mathbf{Z}_1\mathbf{Z}_1^T\mathbf{C}_1^{1/2}) - \varphi_i(\mathbf{C}_1)| \le \|\mathbf{C}_1^{1/2}(n^{-1}\mathbf{Z}_1\mathbf{Z}_1^T - \mathbf{I}_{k_1})\mathbf{C}_1^{1/2}\|_F \to 0. \tag{EC.31}$$

Therefore, (EC.30) and (EC.31) together yield that uniformly over $i \in J_1$,

$$q^{-\alpha_1}\widehat{\lambda}_i = \varphi_i(q^{-\alpha_1}\mathbf{S}_D) \to \varphi_i(\mathbf{C}_1) = c_i,$$

as $q \to \infty$. This completes the proof of **Step 2**.

Step 3: Convergence of eigenvalues with indices in $J_2, \ldots, J_m$. As the second part of induction, for any fixed t, $2 \le t \le m$, we will show $q^{-\alpha_t}\widehat{\lambda}_i \to c_i$ for any $i \in J_t$, as $q \to \infty$. The basic idea in this step is to use Weyl's inequality (Lemma EC.3) to get both a lower bound and an upper bound of $q^{-\alpha_t}\widehat{\lambda}_i$, and show that they converge to the same limit.

We derive the upper bound first. Divided by q^{α_t} on both sides of (EC.24) gives

$$q^{-\alpha_t}\mathbf{S}_D = n^{-1}q^{-\alpha_t}\sum_{l=1}^{t-1}\mathbf{Z}_l^T\mathbf{\Lambda}_l\mathbf{Z}_l + n^{-1}q^{-\alpha_t}\sum_{l=t}^{m+1}\mathbf{Z}_l^T\mathbf{\Lambda}_l\mathbf{Z}_l.$$

Applying Weyl's inequality, we get

$$\begin{aligned}\varphi_i(q^{-\alpha_t}\mathbf{S}_D) &\le \varphi_{1+\sum_{l=1}^{t-1}k_l}(\sum_{l=1}^{t-1}\mathbf{Z}_l^T\mathbf{\Lambda}_l\mathbf{Z}_l/nq^{\alpha_t}) + \varphi_{i-\sum_{l=1}^{t-1}k_l}(n^{-1}q^{-\alpha_t}\sum_{l=t}^{m+1}\mathbf{Z}_l^T\mathbf{\Lambda}_l\mathbf{Z}_l)\\ &= \varphi_{i-\sum_{l=1}^{t-1}k_l}(n^{-1}q^{-\alpha_t}\sum_{l=t}^{m+1}\mathbf{Z}_l^T\mathbf{\Lambda}_l\mathbf{Z}_l),\end{aligned} \tag{EC.32}$$

where the first term is indeed zero since $n^{-1}q^{-\alpha_t}\sum_{l=1}^{t-1}\mathbf{Z}_l^T\mathbf{\Lambda}_l\mathbf{Z}_l$ has a rank no more than $\sum_{l=1}^{t-1}k_l$. It gives an upper bound of $\varphi_i(q^{-\alpha_t}\mathbf{S}_D)$. By the same argument as (EC.28) in **Step 2**, under Conditions 1 and 2, we have

$$\|n^{-1}q^{-\alpha_t}\sum_{l=t}^{m+1}\mathbf{Z}_l^T\mathbf{\Lambda}_l\mathbf{Z}_l - n^{-1}\mathbf{Z}_t^T\mathbf{C}_t\mathbf{Z}_t\|_F \to 0,\ as\ q\to\infty.$$

Similar to (EC.29), it implies the upper bound of $\varphi_i(q^{-\alpha_t}\mathbf{S}_D)$ in (EC.32) converges to the same limit as $\varphi_{i-\sum_{l=1}^{t-1}k_l}(n^{-1}\mathbf{Z}_t^T\mathbf{C}_t\mathbf{Z}_t)$ uniformly over $i\in J_t$ as $q\to\infty$.

On the other hand, by Weyl's inequality, we also have

$$\begin{aligned}\varphi_i(q^{-\alpha_t}\mathbf{S}_D) &\geq \varphi_i(n^{-1}q^{-\alpha_t}\sum_{l=1}^{t}\mathbf{Z}_l^T\mathbf{\Lambda}_l\mathbf{Z}_l) + \varphi_n(n^{-1}q^{-\alpha_t}\sum_{l=t+1}^{m+1}\mathbf{Z}_l^T\mathbf{\Lambda}_l\mathbf{Z}_l)\\ &\geq \varphi_i(n^{-1}q^{-\alpha_t}\sum_{l=1}^{t}\mathbf{Z}_l^T\mathbf{\Lambda}_l\mathbf{Z}_l),\end{aligned}$$

where the second term vanishes since the eigenvalues of $\sum_{l=t+1}^{m+1}\mathbf{Z}_l^T\mathbf{\Lambda}_l\mathbf{Z}_l$ are non-negative. In fact, $n^{-1}q^{-\alpha_t}\sum_{l=t+1}^{m+1}\mathbf{Z}_l^T\mathbf{\Lambda}_l\mathbf{Z}_l$ would converge to a zero matrix in Frobenius norm under Conditions 1 and 2, similarly as in (EC.26). For the term $\varphi_i(n^{-1}q^{-\alpha_t}\sum_{l=1}^{t}\mathbf{Z}_l^T\mathbf{\Lambda}_l\mathbf{Z}_l)$, we use Weyl's inequality once more to get

$$\begin{aligned}&\varphi_{\sum_{l=1}^{t}k_l}(n^{-1}q^{-\alpha_t}\sum_{l=1}^{t-1}\mathbf{Z}_l^T\mathbf{\Lambda}_l\mathbf{Z}_l)\\ \leq\ &\varphi_i(n^{-1}q^{-\alpha_t}\sum_{l=1}^{t}\mathbf{Z}_l^T\mathbf{\Lambda}_l\mathbf{Z}_l) + \varphi_{1-i+\sum_{l=1}^{t}k_l}(-n^{-1}q^{-\alpha_t}\mathbf{Z}_t^T\mathbf{\Lambda}_t\mathbf{Z}_t).\end{aligned}$$

Note that the term on the left hand side is indeed zero since the inside matrix has a rank no more than $\sum_{l=1}^{t-1}k_l$. It follows that

$$\begin{aligned}&\varphi_i(n^{-1}q^{-\alpha_t}\sum_{l=1}^{t}\mathbf{Z}_t^T\mathbf{\Lambda}_t\mathbf{Z}_t) \geq -\varphi_{1-i+\sum_{l=1}^{t}k_l}(-n^{-1}q^{-\alpha_t}\mathbf{Z}_t^T\mathbf{\Lambda}_t\mathbf{Z}_t)\\ =&\varphi_{k_t-(1-i+\sum_{l=1}^{t}k_l)+1}(n^{-1}q^{-\alpha_t}\mathbf{Z}_t^T\mathbf{\Lambda}_t\mathbf{Z}_t) = \varphi_{i-\sum_{l=1}^{t-1}k_l}(n^{-1}q^{-\alpha_t}\mathbf{Z}_t^T\mathbf{\Lambda}_t\mathbf{Z}_t),\end{aligned}$$

where we make use of the fact that $\varphi_i(\mathbf{A}) = -\varphi_{n-i+1}(-\mathbf{A})$ for any $n\times n$ real symmetric matrix $\mathbf{A}$, and any $1\leq i\leq n$.

Therefore, we get a lower bound $\varphi_{i-\sum_{l=1}^{t-1}k_l}(n^{-1}q^{-\alpha_t}\mathbf{Z}_t^T\mathbf{\Lambda}_t\mathbf{Z}_t)$ for $\varphi_i(q^{-\alpha_t}\mathbf{S}_D)$. In terms of sample eigenvalues, the above argument shows that for any $\widehat{\lambda}_i$, $i\in J_t$, $1\le t\le m$,

$$\widehat{\lambda}_i=\varphi_i(\mathbf{S}_D)\ge\varphi_{i-\sum_{l=1}^{t-1}k_l}(n^{-1}\mathbf{Z}_t^T\mathbf{\Lambda}_t\mathbf{Z}_t), \tag{EC.33}$$

which is useful in proving the convergence properties of the sample score vectors.

Now we show that the two bounds converge to the same limit. Similar to (EC.27), as $q\to\infty$, we have

$$\|n^{-1}q^{-\alpha_t}\mathbf{Z}_t^T\mathbf{\Lambda}_t\mathbf{Z}_t-n^{-1}\mathbf{Z}_t^T\mathbf{C}_t\mathbf{Z}_t\|_F\to 0,$$

which gives

$$\max_{i\in J_t}\left|\varphi_{i-\sum_{l=1}^{t-1}k_l}(n^{-1}q^{-\alpha_t}\mathbf{Z}_t^T\mathbf{\Lambda}_t\mathbf{Z}_t)-\varphi_{i-\sum_{l=1}^{t-1}k_l}(n^{-1}\mathbf{Z}_t^T\mathbf{C}_t\mathbf{Z}_t)\right|\to 0.$$

It shows that the lower bound of $\varphi_i(q^{-\alpha_t}\mathbf{S}_D)$ converges to the same limit as the term $\varphi_{i-\sum_{l=1}^{t-1}k_l}(n^{-1}\mathbf{Z}_t^T\mathbf{C}_t\mathbf{Z}_t)$ uniformly over $i\in J_t$, so does the upper bound in (EC.32). It follows that $\varphi_i(q^{-\alpha_t}\mathbf{S}_D)$ would also converge to the same limit as $\varphi_{i-\sum_{l=1}^{t-1}k_l}(n^{-1}\mathbf{Z}_t^T\mathbf{C}_t\mathbf{Z}_t)$ uniformly over $i\in J_t$. That is, as $q\to\infty$,

$$\max_{i\in J_t}|\varphi_i(q^{-\alpha_t}\mathbf{S}_D)-\varphi_{i-\sum_{l=1}^{t-1}k_l}(n^{-1}\mathbf{Z}_t^T\mathbf{C}_t\mathbf{Z}_t)|\to 0.$$

By a similar argument as in (EC.30) and (EC.31), we then have

$$\varphi_i(q^{-\alpha_t}\mathbf{S}_D)\to\varphi_{i-\sum_{l=1}^{t-1}k_l}(\mathbf{C}_t)=c_i,$$

uniformly over $i\in J_t$, as $q\to\infty$. Along with the first step of induction in **Step 2**, we finish the proof of Lemma EC.1.

EC.2.2. Proof of Lemma EC.2

To prove the probability bound in Lemma EC.2, we will apply Bonferroni's inequality and Gaussian tail probability bound. Since $\widetilde{\boldsymbol{\varepsilon}}=(\mathbf{F}-\widehat{\mathbf{F}})\boldsymbol{\gamma}+\boldsymbol{\varepsilon}$, some important bounds are needed before

continuation. First, the inequality $\|\boldsymbol{\gamma}\|_1 \le K\rho$ follows immediately from the fact $\|\boldsymbol{\gamma}\|_\infty \le \rho$. Moreover, based on the estimation error bound of $\widehat{\mathbf{F}}$ in Condition 3, we know that inequality (EC.14) holds. These two inequalities yield

$$\|(\mathbf{F}-\widehat{\mathbf{F}})\boldsymbol{\gamma}\|_2 \le \|\boldsymbol{\gamma}\|_1 \cdot \max_{1\le j\le K}\|\mathbf{f}_j-\widehat{\mathbf{f}}_j\|_2 \le K\rho\cdot\frac{c_2}{2K\rho}\sqrt{\log n}=\frac{c_2}{2}\sqrt{\log n}, \tag{EC.34}$$

which gives

$$n^{-1}|\mathbf{x}_i^T(\mathbf{F}-\widehat{\mathbf{F}})\boldsymbol{\gamma}| \le n^{-1/2}\|(\mathbf{F}-\widehat{\mathbf{F}})\boldsymbol{\gamma}\|_2 \le \frac{c_2}{2}\sqrt{\frac{\log n}{n}} \le \frac{c_2}{2}\sqrt{\frac{\log p}{n}}. \tag{EC.35}$$

Similarly we have $n^{-1}|\mathbf{f}_j^T(\mathbf{F}-\widehat{\mathbf{F}})\boldsymbol{\gamma}| \le 2^{-1}c_2\sqrt{(\log n)/n}$.

Now we proceed to prove the probability bounds of the two events. Recall that both $\mathbf{f}_j$ and $\widehat{\mathbf{f}}_j$ have been rescaled to have L_2-norm $n^{1/2}$ and $\boldsymbol{\varepsilon}\sim N(\mathbf{0},\sigma^2\mathbf{I}_n)$ (Section 2). Given $\mathbf{x}_i$ and $\widehat{\mathbf{f}}_j$, it follows that $n^{-1}\mathbf{x}_i^T\boldsymbol{\varepsilon}\sim N(0,\sigma^2/n)$ and $n^{-1}\widehat{\mathbf{f}}_j^T\boldsymbol{\varepsilon}\sim N(0,\sigma^2/n)$ for any i and j. By Bonferroni's inequality, the tail probability of $\widetilde{\mathcal{E}}$ satisfies

$$P(\widetilde{\mathcal{E}}^c) \le \sum_{i=1}^p P\left(|n^{-1}\mathbf{x}_i^T\widetilde{\boldsymbol{\varepsilon}}|>c_2\sqrt{(\log p)/n}\right)+\sum_{j=1}^K P\left(|n^{-1}\widehat{\mathbf{f}}_j^T\widetilde{\boldsymbol{\varepsilon}}|>c_2\sqrt{(\log p)/n}\right).$$

By inequality (EC.35) and Gaussian tail probability bound, for the first term on the right hand side above, we have

$$\begin{aligned}&\sum_{i=1}^p P\left(\frac{|\mathbf{x}_i^T\widetilde{\boldsymbol{\varepsilon}}|}{n}>c_2\sqrt{\frac{\log p}{n}}\right) \le \sum_{i=1}^p P\left(\frac{|\mathbf{x}_i^T\boldsymbol{\varepsilon}|}{n}>c_2\sqrt{\frac{\log p}{n}}-n^{-1}|\mathbf{x}_i^T(\mathbf{F}-\widehat{\mathbf{F}})\boldsymbol{\gamma}|\right)\\ \le &\sum_{i=1}^p P\left(\frac{|\mathbf{x}_i^T\boldsymbol{\varepsilon}|}{n}>\frac{c_2}{2}\sqrt{\frac{\log p}{n}}\right) \le \sum_{j=1}^p \frac{4\sigma}{c_2\sqrt{\log p}}\frac{1}{\sqrt{2\pi}}e^{-\frac{c_2^2\log p}{8\sigma^2}} \le \frac{2\sqrt{2}\sigma}{c_2\sqrt{\pi\log p}}p^{1-\frac{c_2^2}{8\sigma^2}}.\end{aligned}$$

For the second term, similarly we have

$$\sum_{j=1}^K P\left(|n^{-1}\widehat{\mathbf{f}}_j^T\widetilde{\boldsymbol{\varepsilon}}|>c_2\sqrt{(\log p)/n}\right) \le \frac{2\sqrt{2}\sigma K}{c_2\sqrt{\pi\log p}}p^{-\frac{c_2^2}{8\sigma^2}}.$$

As K is no larger than p, the two bounds above give

$$P(\widetilde{\mathcal{E}}^c) \le \frac{4\sqrt{2}\sigma}{c_2\sqrt{\pi\log p}}p^{1-\frac{c_2^2}{8\sigma^2}}.$$

By a similar argument, the bound on $P(\widetilde{\mathcal{E}}_0^c)$ can be derived as

$$P(\widetilde{\mathcal{E}}_0^c) \leq \frac{2\sqrt{2}\sigma s}{c_2\sqrt{\pi \log n}} n^{-\frac{c_2^2}{8\sigma^2}}.$$

Thus, for the intersection event $\widetilde{\mathcal{E}} \cap \widetilde{\mathcal{E}}_0$, we have

$$P\{(\widetilde{\mathcal{E}} \cap \widetilde{\mathcal{E}}_0)^c\} \leq P(\tilde{\mathcal{E}}^c) + P(\widetilde{\mathcal{E}}_0^c) \leq \frac{4\sqrt{2}\sigma}{c_2\sqrt{\pi \log p}} p^{1-\frac{c_2^2}{8\sigma^2}} + \frac{2\sqrt{2}\sigma s}{c_2\sqrt{\pi \log n}} n^{-\frac{c_2^2}{8\sigma^2}},$$

which converges to zero as $n \to \infty$ for $c_2 > 2\sqrt{2}\sigma$. This completes the proof of Lemma EC.2.