Supplement to "Culling the Herd of Moments with Penalized Empirical Likelihood"

Jinyuan Chang, Zhentao Shi, and Jia Zhang

This supplement consists of three parts. Part A provides the proofs and technical details about the method developed in the present article. Part B reports additional simulation results concerning the liner IV model in the main text and an additional dynamic panel data model, respectively. Part C checks the robustness of PEL in the empirical application.

We use "C" to denote a generic positive finite constant that may be different in different uses.

A Theoretical details and technical proofs

A.1 Proposition A.1

Let $\mathbf{V}^{(\mathcal{I})}(\boldsymbol{\theta}) = \mathbb{E}\{\mathbf{g}_i^{(\mathcal{I})}(\boldsymbol{\theta})^{\otimes 2}\}$ and define $\mathbf{J}^{(\mathcal{I})} = ([\mathbb{E}\{\nabla_{\boldsymbol{\theta}}\mathbf{g}_i^{(\mathcal{I})}(\boldsymbol{\theta}_0)\}]^{\mathrm{T}}\{\mathbf{V}^{(\mathcal{I})}(\boldsymbol{\theta}_0)\}^{-1/2})^{\otimes 2}$. Proposition A.1 gives the asymptotic normality of the standard EL estimator $\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{I})}$.

Proposition A.1. *Assume that:* (i) *There exists a universal constant $C_1 > 0$ such that*

$$\inf_{\boldsymbol{\theta}\in\{\boldsymbol{\theta}\in\boldsymbol{\Theta}:\,|\boldsymbol{\theta}-\boldsymbol{\theta}_0|_\infty>\varepsilon\}} |\mathbb{E}\{\mathbf{g}_i^{(\mathcal{I})}(\boldsymbol{\theta})\}|_\infty \geq C_1\varepsilon \tag{A.1}$$

for any $\varepsilon > 0$. (ii) *There exist universal constants $C_2 > 0$, $C_3 > 1$ and $\gamma > 4$ such that*

$$\max_{j\in\mathcal{I}} \mathbb{E}\left\{ \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |g_{i,j}^{(\mathcal{I})}(\boldsymbol{\theta})|^\gamma \right\} \leq C_2\,, \tag{A.2}$$

$$\mathbb{P}\left[C_3^{-1} \leq \inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \lambda_{\min}\{\widehat{\mathbf{V}}^{(\mathcal{I})}(\boldsymbol{\theta})\} \leq \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \lambda_{\max}\{\widehat{\mathbf{V}}^{(\mathcal{I})}(\boldsymbol{\theta})\} \leq C_3 \right] \to 1 \tag{A.3}$$

with $\widehat{\mathbf{V}}^{(\mathcal{I})}(\boldsymbol{\theta}) = \mathbb{E}_n\{\mathbf{g}_i^{(\mathcal{I})}(\boldsymbol{\theta})^{\otimes 2}\}$. (iii) *Each element of $\mathbf{g}^{(\mathcal{I})}(\mathbf{X};\boldsymbol{\theta})$ is twice continuously differentiable with respect to $\boldsymbol{\theta}$ for any $\mathbf{X}$, and*

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left(|\mathbb{E}_n[\{\nabla_{\boldsymbol{\theta}}\mathbf{g}_i^{(\mathcal{I})}(\boldsymbol{\theta})\}^{\circ 2}]|_\infty + \max_{j\in\mathcal{I}} |\mathbb{E}_n[\{\nabla_{\boldsymbol{\theta}}^2 g_{i,j}^{(\mathcal{I})}(\boldsymbol{\theta})\}^{\circ 2}]|_\infty + |\mathbb{E}_n[\{\mathbf{g}_i^{(\mathcal{I})}(\boldsymbol{\theta})\}^{\circ\gamma}]|_\infty \right) = O_{\mathrm{p}}(1)\,, \tag{A.4}$$

where γ is specified in (A.2). (iv) *There exists a universal constant $C_4 > 1$ such that*

$$C_4^{-1} < \lambda_{\min}\{\mathbf{Q}^{(\mathcal{I})}\} \leq \lambda_{\max}\{\mathbf{Q}^{(\mathcal{I})}\} < C_4\,, \tag{A.5}$$

where $\mathbf{Q}^{(\mathcal{I})} = ([\mathbb{E}\{\nabla_{\boldsymbol{\theta}}\mathbf{g}_i^{(\mathcal{I})}(\boldsymbol{\theta}_0)\}]^{\mathrm{T}})^{\otimes 2}$. (v) *$r_1^3 n^{-1+2/\gamma} = o(1)$ and $r_1^3 p^2 n^{-1} = o(1)$. Then*

$$\sqrt{n}\boldsymbol{\alpha}^{\mathrm{T}}\{\mathbf{J}^{(\mathcal{I})}\}^{1/2}\{\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{I})} - \boldsymbol{\theta}_0\} \xrightarrow{d} \mathcal{N}(0,1)$$

as $n \to \infty$ for any $\boldsymbol{\alpha} \in \mathbb{R}^p$ with $|\boldsymbol{\alpha}|_2 = 1$.

Remark A.1. To understand the relative magnitude of r_1, p and n, consider the special case with $\gamma = \infty$ and fixed p, under which $r_1 = o(n^{1/3})$ satisfies the condition. This is consistent with the literature of GMM involving a diverging number of moments under a fixed p (Koenker and Machado 1999); when p diverges, $p \leq r_1 = o(n^{1/5})$ is sufficient for the asymptotic normality in Proposition A.1.

Remark A.2. Strong identification of the parameter of interest is assumed in (A.1). It is possible to generalize ε on the right-hand side of (A.1) to ε^β for some universal constant $\beta > 0$ at the cost of much complicated expressions for the admissible range of r_1 and p. (A.2) restricts the population moments uniformly over the parameter space. (A.3) and (A.5) bound away from zero and infinity the eigenvalues of $\mathbb{E}_n\{\mathbf{g}_i^{(\mathcal{I})}(\boldsymbol{\theta})^{\otimes 2}\}$ and $([\mathbb{E}\{\nabla_{\boldsymbol{\theta}}\mathbf{g}_i^{(\mathcal{I})}(\boldsymbol{\theta}_0)\}]^{\mathrm{T}})^{\otimes 2}$, respectively. The first and second derivatives of the estimating functions are further regularized by (A.4).

Remark A.3. If there are some envelope functions $\{B_{n,j}(\cdot)\}_{j\in\mathcal{I}}$ such that $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |g_j^{(\mathcal{I})}(\mathbf{X};\boldsymbol{\theta})|^\gamma \leq B_{n,j}(\mathbf{X})$ for any j, and $\max_{j\in\mathcal{I}} \mathbb{E}\{B_{n,j}^m(\mathbf{X}_i)\} \leq Km!H^{m-2}$ for any integer $m \geq 2$, where K and H are two universal positive constants independent of j, Petrov (1995)'s Theorem 2.8 implies $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |\mathbb{E}_n[\{\mathbf{g}_i^{(\mathcal{I})}(\boldsymbol{\theta})\}^{\circ\gamma}]|_\infty = O_{\mathrm{p}}(1)$ provided $\log r_1 = o(n)$; the other two requirements in (A.4) can be satisfied in the same manner. The stochastic order $O_{\mathrm{p}}(1)$ in (A.4) can be replaced by $O_{\mathrm{p}}(\varphi_n)$ for some diverging φ_n. Our theoretical results still hold in this broader situation at the expense of more complicated restrictions among r_1, p and n.

Proof. Define $A_n(\boldsymbol{\theta},\boldsymbol{\lambda}) = n^{-1}\sum_{i=1}^n \log\{1 + \boldsymbol{\lambda}^{\mathrm{T}}\mathbf{g}_i^{(\mathcal{I})}(\boldsymbol{\theta})\}$ for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $\boldsymbol{\lambda} \in \hat{\Lambda}_n^{(\mathcal{I})}(\boldsymbol{\theta})$. Then $\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{I})}$ and its associated Lagrange multiplier $\hat{\boldsymbol{\lambda}}$ satisfy the score equation $\nabla_{\boldsymbol{\lambda}} A_n\{\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{I})}, \hat{\boldsymbol{\lambda}}\} = \mathbf{0}$, i.e.

$$\mathbf{0} = \frac{1}{n}\sum_{i=1}^n \frac{\mathbf{g}_i^{(\mathcal{I})}\{\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{I})}\}}{1 + \hat{\boldsymbol{\lambda}}^{\mathrm{T}}\mathbf{g}_i^{(\mathcal{I})}\{\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{I})}\}}\,.$$

By the Taylor expansion, we have

$$\mathbf{0} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{g}_i^{(\mathcal{I})}\{\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{I})}\} - \left(\frac{1}{n}\sum_{i=1}^{n} \frac{\mathbf{g}_i^{(\mathcal{I})}\{\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{I})}\}^{\otimes 2}}{[1 + c\hat{\boldsymbol{\lambda}}^{\mathrm{T}}\mathbf{g}_i^{(\mathcal{I})}\{\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{I})}\}]^2}\right)\hat{\boldsymbol{\lambda}}$$

for some $|c| < 1$, which implies

$$\hat{\boldsymbol{\lambda}} = \left(\frac{1}{n}\sum_{i=1}^{n} \frac{\mathbf{g}_i^{(\mathcal{I})}\{\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{I})}\}^{\otimes 2}}{[1 + c\hat{\boldsymbol{\lambda}}^{\mathrm{T}}\mathbf{g}_i^{(\mathcal{I})}\{\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{I})}\}]^2}\right)^{-1} \bar{\mathbf{g}}^{(\mathcal{I})}\{\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{I})}\}\,.$$

By the implicit function theorem [Theorem 9.28 of Rudin (1976)], for all $\boldsymbol{\theta}$ in a $|\cdot|_2$-neighborhood of $\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{I})}$, there is a $\hat{\boldsymbol{\lambda}}(\boldsymbol{\theta})$ such that $\nabla_{\boldsymbol{\lambda}} A_n\{\boldsymbol{\theta}, \hat{\boldsymbol{\lambda}}(\boldsymbol{\theta})\} = \mathbf{0}$ and $\hat{\boldsymbol{\lambda}}(\boldsymbol{\theta})$ is continuously differentiable in $\boldsymbol{\theta}$. By the concavity of $A_n(\boldsymbol{\theta}, \boldsymbol{\lambda})$ with respect to (w.r.t) $\boldsymbol{\lambda}$, $A_n\{\boldsymbol{\theta}, \hat{\boldsymbol{\lambda}}(\boldsymbol{\theta})\} = \max_{\boldsymbol{\lambda}\in\hat{\Lambda}_n^{(\mathcal{I})}(\boldsymbol{\theta})} A_n(\boldsymbol{\theta}, \boldsymbol{\lambda})$. It follows from the envelope theorem that

$$\mathbf{0} = \nabla_{\boldsymbol{\theta}} A_n\{\boldsymbol{\theta}, \hat{\boldsymbol{\lambda}}(\boldsymbol{\theta})\}\big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{I})}} = \left[\frac{1}{n}\sum_{i=1}^{n} \frac{\nabla_{\boldsymbol{\theta}}\mathbf{g}_i^{(\mathcal{I})}\{\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{I})}\}}{1 + \hat{\boldsymbol{\lambda}}^{\mathrm{T}}\mathbf{g}_i^{(\mathcal{I})}\{\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{I})}\}}\right]^{\mathrm{T}}\hat{\boldsymbol{\lambda}}\,.$$

Therefore, we have

$$\mathbf{0} = \left[\frac{1}{n}\sum_{i=1}^{n} \frac{\nabla_{\boldsymbol{\theta}}\mathbf{g}_i^{(\mathcal{I})}\{\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{I})}\}}{1 + \hat{\boldsymbol{\lambda}}^{\mathrm{T}}\mathbf{g}_i^{(\mathcal{I})}\{\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{I})}\}}\right]^{\mathrm{T}}\left(\frac{1}{n}\sum_{i=1}^{n} \frac{\mathbf{g}_i^{(\mathcal{I})}\{\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{I})}\}^{\otimes 2}}{[1 + c\hat{\boldsymbol{\lambda}}^{\mathrm{T}}\mathbf{g}_i^{(\mathcal{I})}\{\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{I})}\}]^2}\right)^{-1} \bar{\mathbf{g}}^{(\mathcal{I})}\{\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{I})}\}\,. \tag{A.6}$$

Define $F_n(\boldsymbol{\theta}) = \max_{\boldsymbol{\lambda}\in\hat{\Lambda}_n^{(\mathcal{I})}(\boldsymbol{\theta})} A_n(\boldsymbol{\theta}, \boldsymbol{\lambda})$ and let $b_n = r_1 n^{-1}$. As shown in the proof of Proposition 1 of Chang et al. (2018), we have $\max_{\boldsymbol{\lambda}\in\hat{\Lambda}_n^{(\mathcal{I})}(\boldsymbol{\theta}_0)} A_n(\boldsymbol{\theta}_0, \boldsymbol{\lambda}) = O_{\mathrm{p}}(r_1 n^{-1})$ which implies $F_n(\boldsymbol{\theta}_0) = O_{\mathrm{p}}(b_n)$. As $F_n\{\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{I})}\} \le F_n(\boldsymbol{\theta}_0)$, we have $F_n\{\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{I})}\} = O_{\mathrm{p}}(b_n)$. We will first show that for any $\epsilon_n \to \infty$ satisfying $b_n\epsilon_n^2 n^{2/\gamma} = o(1)$, there exists a universal constant $K > 0$ independent of $\boldsymbol{\theta}$ such that $\mathbb{P}\{F_n(\boldsymbol{\theta}) > Kb_n\epsilon_n^2\} \to 1$ as $n \to \infty$ for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ satisfying $|\boldsymbol{\theta} - \boldsymbol{\theta}_0|_\infty > \epsilon_n b_n^{1/2}$. Thus $|\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{I})} - \boldsymbol{\theta}_0|_\infty = O_{\mathrm{p}}(\epsilon_n b_n^{1/2})$. Notice that we can select an arbitrary slowly diverging ϵ_n, to ensure $|\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{I})} - \boldsymbol{\theta}_0|_\infty = O_{\mathrm{p}}(b_n^{1/2})$, following a standard result from probability theory. To do this, we will use the technique developed for the proof of Theorem 1 in Chang et al. (2013). For any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ satisfying $|\boldsymbol{\theta} - \boldsymbol{\theta}_0|_\infty > \epsilon_n b_n^{1/2}$, let $j_0 = \arg\max_{j\in\mathcal{I}} |\mathbb{E}\{g_{i,j}^{(\mathcal{I})}(\boldsymbol{\theta})\}|$. Define $\mu_{j_0} = \mathbb{E}\{g_{i,j_0}^{(\mathcal{I})}(\boldsymbol{\theta})\}$ and $\tilde{\boldsymbol{\lambda}} = \delta b_n^{1/2}\epsilon_n \mathbf{e}_{j_0}$ where $\delta > 0$ is a constant to be determined later, and $\mathbf{e}_{j_0}$ is an r_1-dimensional vector with the j_0-th component being 1 and other components being 0. Without loss of generality, we assume $\mu_{j_0} > 0$. (A.2) and the Markov inequality yield that $\max_{i\in[n]} |g_{i,j_0}^{(\mathcal{I})}(\boldsymbol{\theta})| = O_{\mathrm{p}}(n^{1/\gamma})$, which implies $\max_{i\in[n]} |\tilde{\boldsymbol{\lambda}}^{\mathrm{T}}\mathbf{g}_i^{(\mathcal{I})}(\boldsymbol{\theta})| = O_{\mathrm{p}}(b_n^{1/2}\epsilon_n n^{1/\gamma}) = o_{\mathrm{p}}(1)$. Then $\tilde{\boldsymbol{\lambda}} \in \hat{\Lambda}_n^{(\mathcal{I})}(\boldsymbol{\theta})$ with probability approaching one

(w.p.a.1). Write $\tilde{\boldsymbol{\lambda}} = (\tilde{\lambda}_1, \ldots, \tilde{\lambda}_r)^{\mathrm{T}}$. By the definition of $F_n(\boldsymbol{\theta})$, it holds w.p.a.1 that

$$\begin{aligned} F_n(\boldsymbol{\theta}) \geq \frac{1}{n}\sum_{i=1}^{n} \log\{1 + \tilde{\boldsymbol{\lambda}}^{\mathrm{T}} \mathbf{g}_i^{(\mathcal{I})}(\boldsymbol{\theta})\} &= \frac{1}{n}\sum_{i=1}^{n} \tilde{\lambda}_{j_0} g_{i,j_0}^{(\mathcal{I})}(\boldsymbol{\theta}) - \frac{1}{2n}\sum_{i=1}^{n} \frac{\{\tilde{\lambda}_{j_0} g_{i,j_0}^{(\mathcal{I})}(\boldsymbol{\theta})\}^2}{\{1 + c\tilde{\lambda}_{j_0} g_{i,j_0}^{(\mathcal{I})}(\boldsymbol{\theta})\}^2} \\ &\geq \frac{1}{n}\sum_{i=1}^{n} \tilde{\lambda}_{j_0} g_{i,j_0}^{(\mathcal{I})}(\boldsymbol{\theta}) - \frac{1}{n}\sum_{i=1}^{n} \{\tilde{\lambda}_{j_0} g_{i,j_0}^{(\mathcal{I})}(\boldsymbol{\theta})\}^2 \end{aligned}$$

for some $|c| < 1$ and $\tilde{\lambda}_{j_0} = \delta b_n^{1/2} \epsilon_n$. Therefore, it holds that

$$\mathbb{P}\big\{F_n(\boldsymbol{\theta}) \leq K b_n \epsilon_n^2\big\} \leq \mathbb{P}\bigg[\frac{1}{n}\sum_{i=1}^{n} \{g_{i,j_0}^{(\mathcal{I})}(\boldsymbol{\theta}) - \mu_{j_0}\} \leq b_n^{1/2}\epsilon_n \bigg\{\frac{K}{\delta} + \frac{\delta}{n}\sum_{i=1}^{n} |g_{i,j_0}^{(\mathcal{I})}(\boldsymbol{\theta})|^2\bigg\} - \mu_{j_0}\bigg] + o(1)\,.$$

From (A.2) and the Markov inequality, there exists a universal positive constant L independent of $\boldsymbol{\theta}$ such that $\mathbb{P}\{n^{-1}\sum_{i=1}^{n} |g_{i,j_0}^{(\mathcal{I})}(\boldsymbol{\theta})|^2 > L\} \to 0$ as $n \to \infty$. Thus, with $\delta = (K/L)^{1/2}$, we have

$$\mathbb{P}\big\{F_n(\boldsymbol{\theta}) \leq K b_n \epsilon_n^2\big\} \leq \mathbb{P}\bigg[\frac{1}{n}\sum_{i=1}^{n} \{g_{i,j_0}^{(\mathcal{I})}(\boldsymbol{\theta}) - \mu_{j_0}\} \leq 2b_n^{1/2}\epsilon_n (KL)^{1/2} - \mu_{j_0}\bigg] + o(1)\,.$$

From (A.1), we know that $\mu_{j_0} \geq C_1 \epsilon_n b_n^{1/2}$ with C_1 specified in (A.1). For sufficiently small K independent of $\boldsymbol{\theta}$, we have $2b_n^{1/2}\epsilon_n(KL)^{1/2} - \mu_{j_0} \leq -c\mu_{j_0}$ for some $c \in (0,1)$, which implies that $n^{1/2}\{2b_n^{1/2}\epsilon_n(KL)^{1/2} - \mu_{j_0}\} \leq -cn^{1/2}\mu_{j_0} \lesssim -\epsilon_n b_n^{1/2} n^{1/2} \to -\infty$. Since $n^{-1/2}\sum_{i=1}^{n}\{g_{i,j_0}^{(\mathcal{I})}(\boldsymbol{\theta}) - \mu_{j_0}\} \xrightarrow{d} \mathcal{N}(0, \sigma^2)$ for some $\sigma > 0$, it holds that $\mathbb{P}\{F_n(\boldsymbol{\theta}) \leq K b_n \epsilon_n^2\} \to 0$. Therefore, we have $|\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{I})} - \boldsymbol{\theta}_0|_\infty = O_{\mathrm{p}}(b_n^{1/2})$.

Under (A.2) and (A.3), Proposition 1 of Chang et al. (2018) implies that $|\bar{\mathbf{g}}^{(\mathcal{I})}\{\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{I})}\}|_2 = O_{\mathrm{p}}(r_1^{1/2} n^{-1/2})$. It follows from the Taylor expansion that $\bar{\mathbf{g}}^{(\mathcal{I})}\{\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{I})}\} - \bar{\mathbf{g}}^{(\mathcal{I})}(\boldsymbol{\theta}_0) = \{\nabla_{\boldsymbol{\theta}} \bar{\mathbf{g}}^{(\mathcal{I})}(\dot{\boldsymbol{\theta}})\}\{\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{I})} - \boldsymbol{\theta}_0\}$ for some $\dot{\boldsymbol{\theta}}$ between $\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{I})}$ and $\boldsymbol{\theta}_0$. Similar to Lemma 3 of Chang et al. (2018), we know $|[\nabla_{\boldsymbol{\theta}} \bar{\mathbf{g}}^{(\mathcal{I})}(\dot{\boldsymbol{\theta}}) - \mathbb{E}\{\nabla_{\boldsymbol{\theta}} \mathbf{g}_i^{(\mathcal{I})}(\boldsymbol{\theta}_0)\}]\mathbf{z}|_2 = |\mathbf{z}|_2 \cdot [O_{\mathrm{p}}(r_1^{1/2} p^{3/2} b_n^{1/2}) + O_{\mathrm{p}}\{(n^{-1} r_1 p \log r_1)^{1/2}\}]$ holds uniformly over $\mathbf{z} \in \mathbb{R}^p$. If $r_1 p^3 b_n = o(1)$ and $n^{-1} r_1 p \log r_1 = o(1)$, (A.5) implies that $\lambda_{\min}([\{\nabla_{\boldsymbol{\theta}} \bar{\mathbf{g}}^{(\mathcal{I})}(\dot{\boldsymbol{\theta}})\}^{\mathrm{T}}]^{\otimes 2})$ is uniformly bounded away from zero w.p.a.1. Recall $|\bar{\mathbf{g}}^{(\mathcal{I})}(\boldsymbol{\theta}_0)|_2 = O_{\mathrm{p}}(r_1^{1/2} n^{-1/2})$. Then $O_{\mathrm{p}}(r_1 n^{-1}) \geq \lambda_{\min}([\{\nabla_{\boldsymbol{\theta}} \bar{\mathbf{g}}^{(\mathcal{I})}(\dot{\boldsymbol{\theta}})\}^{\mathrm{T}}]^{\otimes 2}) |\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{I})} - \boldsymbol{\theta}_0|_2^2$, which implies $|\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{I})} - \boldsymbol{\theta}_0|_2 = O_{\mathrm{p}}(r_1^{1/2} n^{-1/2})$. Repeating the proof of Lemma 3 of Chang et al. (2018), we can improve the convergence rate of $|[\nabla_{\boldsymbol{\theta}} \bar{\mathbf{g}}^{(\mathcal{I})}(\dot{\boldsymbol{\theta}}) - \mathbb{E}\{\nabla_{\boldsymbol{\theta}} \mathbf{g}_i^{(\mathcal{I})}(\boldsymbol{\theta}_0)\}]\mathbf{z}|_2$. More specifically, $|[\nabla_{\boldsymbol{\theta}} \bar{\mathbf{g}}^{(\mathcal{I})}(\dot{\boldsymbol{\theta}}) - \mathbb{E}\{\nabla_{\boldsymbol{\theta}} \mathbf{g}_i^{(\mathcal{I})}(\boldsymbol{\theta}_0)\}]\mathbf{z}|_2 = |\mathbf{z}|_2 \cdot O_{\mathrm{p}}(r_1 p n^{-1/2})$ holds uniformly over $\mathbf{z} \in \mathbb{R}^p$. Identical to the proof of Proposition 1 of Chang et al. (2018), we have $|\hat{\boldsymbol{\lambda}}|_2 = O_{\mathrm{p}}(r_1^{1/2} n^{-1/2})$. Under (A.3) and (A.4), similar to Lemmas 1–3 of Chang et al. (2018), we

have the following two results:

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \frac{\mathbf{g}_i^{(\mathcal{I})}\{\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{I})}\}^{\otimes 2}}{[1 + c\hat{\boldsymbol{\lambda}}^{\mathrm{T}}\mathbf{g}_i^{(\mathcal{I})}\{\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{I})}\}]^2} - \mathbf{V}^{(\mathcal{I})}(\boldsymbol{\theta}_0) \right\|_2 = O_{\mathrm{p}}(r_1 n^{-1/2+1/\gamma}) + O_{\mathrm{p}}(r_1 p^{1/2} n^{-1/2}) + O_{\mathrm{p}}\{r_1 (n^{-1}\log r_1)^{1/2}\}$$

and

$$\left| \left[\frac{1}{n} \sum_{i=1}^{n} \frac{\nabla_{\boldsymbol{\theta}} \mathbf{g}_i^{(\mathcal{I})}\{\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{I})}\}}{1 + \hat{\boldsymbol{\lambda}}^{\mathrm{T}}\mathbf{g}_i^{(\mathcal{I})}\{\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{I})}\}} - \mathbb{E}\{\nabla_{\boldsymbol{\theta}} \mathbf{g}_i^{(\mathcal{I})}(\boldsymbol{\theta}_0)\} \right] \mathbf{z} \right|_2 = |\mathbf{z}|_2 \cdot O_{\mathrm{p}}(r_1 p n^{-1/2})$$

holds uniformly over $\mathbf{z} \in \mathbb{R}^p$. Therefore, by (A.6), for any $\boldsymbol{\delta} \in \mathbb{R}^p$ with finite L_2-norm, we have

$$\begin{aligned} & n^{1/2}\boldsymbol{\delta}^{\mathrm{T}}[\mathbb{E}\{\nabla_{\boldsymbol{\theta}} \mathbf{g}_i^{(\mathcal{I})}(\boldsymbol{\theta}_0)\}]^{\mathrm{T}}\{\mathbf{V}^{(\mathcal{I})}(\boldsymbol{\theta}_0)\}^{-1}[\bar{\mathbf{g}}^{(\mathcal{I})}\{\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{I})}\} - \bar{\mathbf{g}}^{(\mathcal{I})}(\boldsymbol{\theta}_0)] \\ = & - n^{1/2}\boldsymbol{\delta}^{\mathrm{T}}[\mathbb{E}\{\nabla_{\boldsymbol{\theta}} \mathbf{g}_i^{(\mathcal{I})}(\boldsymbol{\theta}_0)\}]^{\mathrm{T}}\{\mathbf{V}^{(\mathcal{I})}(\boldsymbol{\theta}_0)\}^{-1}\bar{\mathbf{g}}^{(\mathcal{I})}(\boldsymbol{\theta}_0) + O_{\mathrm{p}}(r_1^{3/2} n^{-1/2}\log^{1/2} r_1) \\ & + O_{\mathrm{p}}(r_1^{3/2} n^{-1/2+1/\gamma}) + O_{\mathrm{p}}(r_1^{3/2} p n^{-1/2}) . \end{aligned} \tag{A.7}$$

Recall $\bar{\mathbf{g}}^{(\mathcal{I})}\{\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{I})}\} - \bar{\mathbf{g}}^{(\mathcal{I})}(\boldsymbol{\theta}_0) = \{\nabla_{\boldsymbol{\theta}} \bar{\mathbf{g}}^{(\mathcal{I})}(\dot{\boldsymbol{\theta}})\}\{\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{I})} - \boldsymbol{\theta}_0\}$, $\mathbf{J}^{(\mathcal{I})} = ([\mathbb{E}\{\nabla_{\boldsymbol{\theta}} \mathbf{g}_i^{(\mathcal{I})}(\boldsymbol{\theta}_0)\}]^{\mathrm{T}}\{\mathbf{V}^{(\mathcal{I})}(\boldsymbol{\theta}_0)\}^{-1/2})^{\otimes 2}$ and $|[\nabla_{\boldsymbol{\theta}} \bar{\mathbf{g}}^{(\mathcal{I})}(\dot{\boldsymbol{\theta}}) - \mathbb{E}\{\nabla_{\boldsymbol{\theta}} \mathbf{g}_i^{(\mathcal{I})}(\boldsymbol{\theta}_0)\}]\mathbf{z}|_2 = |\mathbf{z}|_2 \cdot O_{\mathrm{p}}(r_1 p n^{-1/2})$ holds uniformly over $\mathbf{z} \in \mathbb{R}^p$. Thus, (A.7) implies

$$\begin{aligned} n^{1/2}\boldsymbol{\delta}^{\mathrm{T}}\mathbf{J}^{(\mathcal{I})}\{\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{I})} - \boldsymbol{\theta}_0\} = & - n^{1/2}\boldsymbol{\delta}^{\mathrm{T}}[\mathbb{E}\{\nabla_{\boldsymbol{\theta}} \mathbf{g}_i^{(\mathcal{I})}(\boldsymbol{\theta}_0)\}]^{\mathrm{T}}\{\mathbf{V}^{(\mathcal{I})}(\boldsymbol{\theta}_0)\}^{-1}\bar{\mathbf{g}}^{(\mathcal{I})}(\boldsymbol{\theta}_0) \\ & + O_{\mathrm{p}}(r_1^{3/2} n^{-1/2}\log^{1/2} r_1) + O_{\mathrm{p}}(r_1^{3/2} n^{-1/2+1/\gamma}) + O_{\mathrm{p}}(r_1^{3/2} p n^{-1/2}) . \end{aligned} \tag{A.8}$$

For any $\boldsymbol{\alpha} \in \mathbb{R}^p$ with unit L_2-norm, let $\boldsymbol{\delta} = \{\mathbf{J}^{(\mathcal{I})}\}^{-1/2}\boldsymbol{\alpha}$. Write $\mathbf{U} = \{\mathbf{V}^{(\mathcal{I})}(\boldsymbol{\theta}_0)\}^{-1/2}\mathbb{E}\{\nabla_{\boldsymbol{\theta}} \mathbf{g}_i^{(\mathcal{I})}(\boldsymbol{\theta}_0)\}$ and $\mathbf{J}^{(\mathcal{I})} = \mathbf{U}^{\mathrm{T}}\mathbf{U}$. Notice that $\mathbf{U}^{\mathrm{T}}\mathbf{V}^{(\mathcal{I})}(\boldsymbol{\theta}_0)\mathbf{U} = ([\mathbb{E}\{\nabla_{\boldsymbol{\theta}} \mathbf{g}_i^{(\mathcal{I})}(\boldsymbol{\theta}_0)\}]^{\mathrm{T}})^{\otimes 2}$. Then,

$$\begin{aligned} |\mathbb{E}\{\nabla_{\boldsymbol{\theta}} \mathbf{g}_i^{(\mathcal{I})}(\boldsymbol{\theta}_0)\}\boldsymbol{\delta}|_2^2 &= \boldsymbol{\alpha}^{\mathrm{T}}(\mathbf{U}^{\mathrm{T}}\mathbf{U})^{-1/2}\mathbf{U}^{\mathrm{T}}\mathbf{V}^{(\mathcal{I})}(\boldsymbol{\theta}_0)\mathbf{U}(\mathbf{U}^{\mathrm{T}}\mathbf{U})^{-1/2}\boldsymbol{\alpha} \\ &\le \lambda_{\max}\{\mathbf{V}^{(\mathcal{I})}(\boldsymbol{\theta}_0)\}|\mathbf{U}(\mathbf{U}^{\mathrm{T}}\mathbf{U})^{-1/2}\boldsymbol{\alpha}|_2^2 = \lambda_{\max}\{\mathbf{V}^{(\mathcal{I})}(\boldsymbol{\theta}_0)\} . \end{aligned}$$

Then, it follows from (A.3) and (A.5) that $|\boldsymbol{\delta}|_2^2 \le \lambda_{\max}\{\mathbf{V}^{(\mathcal{I})}(\boldsymbol{\theta}_0)\}\lambda_{\min}^{-1}\{([\mathbb{E}\{\nabla_{\boldsymbol{\theta}} \mathbf{g}_i^{(\mathcal{I})}(\boldsymbol{\theta}_0)\}]^{\mathrm{T}})^{\otimes 2}\} = O(1)$. From (A.8), the Central Limit Theorem implies that

$$\begin{aligned} n^{1/2}\boldsymbol{\alpha}^{\mathrm{T}}\{\mathbf{J}^{(\mathcal{I})}\}^{1/2}\{\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{I})} - \boldsymbol{\theta}_0\} = & - n^{1/2}\boldsymbol{\alpha}^{\mathrm{T}}\{\mathbf{J}^{(\mathcal{I})}\}^{-1/2}[\mathbb{E}\{\nabla_{\boldsymbol{\theta}} \mathbf{g}_i^{(\mathcal{I})}(\boldsymbol{\theta}_0)\}]^{\mathrm{T}}\{\mathbf{V}^{(\mathcal{I})}(\boldsymbol{\theta}_0)\}^{-1}\bar{\mathbf{g}}^{(\mathcal{I})}(\boldsymbol{\theta}_0) \\ & + O_{\mathrm{p}}(r_1^{3/2} n^{-1/2}\log^{1/2} r_1) + O_{\mathrm{p}}(r_1^{3/2} n^{-1/2+1/\gamma}) + O_{\mathrm{p}}(r_1^{3/2} p n^{-1/2}) \\ & \xrightarrow{d} \mathcal{N}(0, 1) \end{aligned}$$

provided that $r_1^3 n^{-1+2/\gamma} = o(1)$ and $r_1^3 p^2 n^{-1} = o(1)$. □

A.2 Proof of Proposition 2.2

Define $L^{(\mathcal{I})}(\boldsymbol{\theta}) = \max\big\{\prod_{i=1}^n \pi_i : \pi_i > 0, \sum_{i=1}^n \pi_i = 1, \sum_{i=1}^n \pi_i \mathbf{g}_i^{(\mathcal{I})}(\boldsymbol{\theta}) = \mathbf{0}\big\}$ and $L^{(\mathcal{T})}(\boldsymbol{\theta}, \boldsymbol{\xi}) = \max\big\{\prod_{i=1}^n \pi_i : \pi_i > 0, \sum_{i=1}^n \pi_i = 1, \sum_{i=1}^n \pi_i \mathbf{g}_i^{(\mathcal{I})}(\boldsymbol{\theta}) = \mathbf{0}, \sum_{i=1}^n \pi_i \mathbf{g}_i^{(\mathcal{D})}(\boldsymbol{\theta}) - \boldsymbol{\xi} = \mathbf{0}\big\}$. Let $\hat{\boldsymbol{\pi}}^{(\mathcal{I})} = \{\hat{\pi}_1^{(\mathcal{I})}, \ldots, \hat{\pi}_n^{(\mathcal{I})}\}$ and $\hat{\boldsymbol{\pi}}^{(\mathcal{T})} = \{\hat{\pi}_1^{(\mathcal{T})}, \ldots, \hat{\pi}_n^{(\mathcal{T})}\}$ be the associated $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)$'s such that $L^{(\mathcal{I})}\{\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{I})}\} = \prod_{i=1}^n \hat{\pi}_i^{(\mathcal{I})}$ and $L^{(\mathcal{T})}\{\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{T})}, \hat{\boldsymbol{\xi}}_{\mathrm{EL}}^{(\mathcal{T})}\} = \prod_{i=1}^n \hat{\pi}_i^{(\mathcal{T})}$. Due to $\hat{\pi}_i^{(\mathcal{T})} > 0$, $\sum_{i=1}^n \hat{\pi}_i^{(\mathcal{T})} = 1$ and $\sum_{i=1}^n \hat{\pi}_i^{(\mathcal{T})} \mathbf{g}_i^{(\mathcal{I})}\{\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{T})}\} = \mathbf{0}$, we have $L^{(\mathcal{I})}\{\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{T})}\} \geq \prod_{i=1}^n \hat{\pi}_i^{(\mathcal{T})}$. Due to $\hat{\pi}_i^{(\mathcal{I})} > 0$, $\sum_{i=1}^n \hat{\pi}_i^{(\mathcal{I})} = 1$ and $\sum_{i=1}^n \hat{\pi}_i^{(\mathcal{I})} \mathbf{g}_i^{(\mathcal{I})}\{\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{I})}\} = \mathbf{0}$, letting $\hat{\boldsymbol{\xi}} = \sum_{i=1}^n \hat{\pi}_i^{(\mathcal{I})} \mathbf{g}_i^{(\mathcal{D})}\{\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{I})}\}$, we have $L^{(\mathcal{T})}\{\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{I})}, \hat{\boldsymbol{\xi}}\} \geq \prod_{i=1}^n \hat{\pi}_i^{(\mathcal{I})}$. Since $\{\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{T}),\mathrm{T}}, \hat{\boldsymbol{\xi}}_{\mathrm{EL}}^{(\mathcal{T}),\mathrm{T}}\}^{\mathrm{T}} = \arg\max_{(\boldsymbol{\theta}^{\mathrm{T}}, \boldsymbol{\xi}^{\mathrm{T}})^{\mathrm{T}} \in \boldsymbol{\Psi}} L^{(\mathcal{T})}(\boldsymbol{\theta}, \boldsymbol{\xi})$, then $\prod_{i=1}^n \hat{\pi}_i^{(\mathcal{T})} = L^{(\mathcal{T})}\{\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{T})}, \hat{\boldsymbol{\xi}}_{\mathrm{EL}}^{(\mathcal{T})}\} \geq L^{(\mathcal{T})}\{\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{I})}, \hat{\boldsymbol{\xi}}\} \geq \prod_{i=1}^n \hat{\pi}_i^{(\mathcal{I})}$. Hence, $L^{(\mathcal{I})}\{\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{T})}\} \geq \prod_{i=1}^n \hat{\pi}_i^{(\mathcal{T})} \geq \prod_{i=1}^n \hat{\pi}_i^{(\mathcal{I})} = L^{(\mathcal{I})}\{\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{I})}\}$. Notice that $\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{I})} = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} L^{(\mathcal{I})}(\boldsymbol{\theta})$. Then $L^{(\mathcal{I})}\{\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{I})}\} = L^{(\mathcal{I})}\{\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{T})}\}$. Since (2.2) has a unique solution, we have $\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{I})} = \hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{T})}$. □

A.3 Proof of Proposition 3.1

As we have defined in Section 3, $\mathcal{M}_{\boldsymbol{\psi}}^* = \mathcal{I} \cup \mathcal{D}_{\boldsymbol{\psi}}^*$ with $\mathcal{D}_{\boldsymbol{\psi}}^* = \{j \in \mathcal{D} : |\bar{g}_j^{(\mathcal{T})}(\boldsymbol{\psi})| \geq C_* \nu \rho_2'(0^+)\}$ for any $\boldsymbol{\psi} \in \boldsymbol{\Psi}$, where $C_* \in (0,1)$ is a prescribed constant. Define $\mathcal{D}_{\boldsymbol{\psi}}(c) = \{j \in \mathcal{D} : |\bar{g}_j^{(\mathcal{T})}(\boldsymbol{\psi})| \geq c\nu \rho_2'(0^+)\}$ for any $c \in (C_*, 1)$ and $\mathcal{M}_{\boldsymbol{\psi}}(c) = \mathcal{I} \cup \mathcal{D}_{\boldsymbol{\psi}}(c)$. For any index set $\mathcal{F} \subset \mathcal{T}$ and $\boldsymbol{\psi} \in \boldsymbol{\Psi}$, we write $\widehat{\mathbf{V}}_{\mathcal{F}}^{(\mathcal{T})}(\boldsymbol{\psi}) = n^{-1} \sum_{i=1}^n \mathbf{g}_{i,\mathcal{F}}^{(\mathcal{T})}(\boldsymbol{\psi})^{\otimes 2}$. For any $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_r)^{\mathrm{T}}$ and $\boldsymbol{\psi} = (\theta_1, \ldots, \theta_p, \xi_1, \ldots, \xi_{r_2})^{\mathrm{T}} \in \boldsymbol{\Psi}$, we define

$$f(\boldsymbol{\lambda}; \boldsymbol{\psi}) = \frac{1}{n} \sum_{i=1}^n \log\{1 + \boldsymbol{\lambda}^{\mathrm{T}} \mathbf{g}_i^{(\mathcal{T})}(\boldsymbol{\psi})\} - \sum_{j \in \mathcal{D}} P_{2,\nu}(|\lambda_j|), \tag{A.9}$$

$$S_n(\boldsymbol{\psi}) = \max_{\boldsymbol{\lambda} \in \hat{\Lambda}_n^{(\mathcal{T})}(\boldsymbol{\psi})} f(\boldsymbol{\lambda}; \boldsymbol{\psi}) + \sum_{k \in \mathcal{D}} P_{1,\pi}(|\xi_k|).$$

Write $\boldsymbol{\xi}_0 = (\xi_{0,1}, \ldots, \xi_{0,r_2})^{\mathrm{T}}$. Recall that $\aleph_n = (n^{-1} \log r)^{1/2}$, $\mathcal{S} = \mathcal{P} \cup \mathcal{A}^{\mathrm{c}}$ with $s = |\mathcal{S}|$, $\boldsymbol{\psi}_{0,\mathcal{S}^{\mathrm{c}}} = \mathbf{0}$ and $\boldsymbol{\Psi}_* = \{\boldsymbol{\psi} = (\boldsymbol{\psi}_{\mathcal{S}}^{\mathrm{T}}, \boldsymbol{\psi}_{\mathcal{S}^{\mathrm{c}}}^{\mathrm{T}})^{\mathrm{T}} : |\boldsymbol{\psi}_{\mathcal{S}} - \boldsymbol{\psi}_{0,\mathcal{S}}|_\infty \leq \varepsilon, |\boldsymbol{\psi}_{\mathcal{S}^{\mathrm{c}}}|_1 \leq \aleph_n\}$ for some fixed $\varepsilon > 0$. Then

$$\hat{\boldsymbol{\psi}} = \arg\min_{\boldsymbol{\psi} \in \boldsymbol{\Psi}_*} S_n(\boldsymbol{\psi}).$$

The proof of Proposition 3.1 requires the following lemmas. The proof of Lemma A.1 is similar to that of Lemma 1 in Chang et al. (2018) and we omit it here. Lemma A.2 presents general properties of the Lagrange multiplier $\hat{\boldsymbol{\lambda}}(\boldsymbol{\psi})$ when $\boldsymbol{\psi}$ is in a small neighborhood of $\boldsymbol{\psi}_0$, whose proof is given in Section A.7.1. If we just focus on the properties of $\hat{\boldsymbol{\lambda}}(\boldsymbol{\psi}_0)$, Lemma A.3 states a refined version of Lemma A.2 with proof given in Section A.7.2.

Lemma A.1. *Let* $\mathscr{F} = \{\mathcal{F} \subset \mathcal{T} : |\mathcal{F}| \leq \ell_n\}$ *and* $\boldsymbol{\Psi}_n = \{\boldsymbol{\psi} \in \boldsymbol{\Psi} : |\boldsymbol{\psi}_{\mathcal{S}} - \boldsymbol{\psi}_{0,\mathcal{S}}|_\infty = O_{\mathrm{p}}(\zeta_{1,n}), |\boldsymbol{\psi}_{\mathcal{S}^{\mathrm{c}}}|_1 \leq$

$\zeta_{2,n}\}$ *for some* $\zeta_{1,n}$, $\zeta_{2,n} \to 0$ *as* $n \to \infty$. *If Conditions* 3 *and* 4 *hold,* $\log r = o(n^{1/3})$, $\ell_n(s^2\zeta_{1,n}^2 + \zeta_{2,n}^2) = o(1)$ *and* $\ell_n \aleph_n = o(1)$, *then* $\sup_{\boldsymbol{\psi}\in\boldsymbol{\Psi}_n}\sup_{\mathcal{F}\in\mathscr{F}} \|\widehat{\mathbf{V}}_{\mathcal{F}}^{(\mathcal{T})}(\boldsymbol{\psi}) - \mathbf{V}_{\mathcal{F}}^{(\mathcal{T})}(\boldsymbol{\psi}_0)\|_2 = O_{\mathrm{p}}\{\ell_n^{1/2}(s\zeta_{1,n} + \zeta_{2,n})\} + O_{\mathrm{p}}(\ell_n\aleph_n)$.

Lemma A.2. *Let* $\{\boldsymbol{\psi}_n\}$ *be a sequence in* $\boldsymbol{\Psi}$ *and* $P_{2,\nu}(\cdot) \in \mathscr{P}$ *be a convex function for* $\mathscr{P}$ *defined as* (3.2). *For some* $c \in (C_*, 1)$, *assume that all the eigenvalues of* $\widehat{\mathbf{V}}^{(\mathcal{T})}_{\mathcal{M}_{\boldsymbol{\psi}_n}(c)}(\boldsymbol{\psi}_n)$ *are uniformly bounded away form zero and infinity w.p.a.1. Let* $|\bar{\mathbf{g}}^{(\mathcal{I})}(\boldsymbol{\psi}_n)|_2^2 + |\bar{\mathbf{g}}^{(\mathcal{T})}_{\mathcal{D}_{\boldsymbol{\psi}_n}(c)}(\boldsymbol{\psi}_n) - \nu\rho_2'(0^+)\mathrm{sgn}\{\bar{\mathbf{g}}^{(\mathcal{T})}_{\mathcal{D}_{\boldsymbol{\psi}_n}(c)}(\boldsymbol{\psi}_n)\}|_2^2 = O_{\mathrm{p}}(u_n^2)$ *for some* $u_n \to 0$, *and* $\max_{j\in\mathcal{T}} n^{-1}\sum_{i=1}^n |g_{i,j}^{(\mathcal{T})}(\boldsymbol{\psi}_n)|^\gamma = O_{\mathrm{p}}(1)$. *For some non-random sequence* $\{m_n\}$ *such that* $\mathbb{P}(|\mathcal{M}^*_{\boldsymbol{\psi}_n}| \le m_n) \to 1$ *as* $n \to \infty$, *if* $m_n^{1/2}u_n = o(\nu)$ *and* $m_n^{1/2}u_n n^{1/\gamma} = o(1)$, *then w.p.a.1 there is a sparse global maximizer* $\hat{\boldsymbol{\lambda}}(\boldsymbol{\psi}_n)$ *for* $f(\boldsymbol{\lambda};\boldsymbol{\psi}_n)$ *satisfying the following three results:* (i) $|\hat{\boldsymbol{\lambda}}(\boldsymbol{\psi}_n)|_2 = O_{\mathrm{p}}(u_n)$, (ii) $\mathrm{supp}\{\hat{\boldsymbol{\lambda}}_{\mathcal{D}}(\boldsymbol{\psi}_n)\} \subset \mathcal{D}_{\boldsymbol{\psi}_n}(c)$, *and* (iii) $\mathrm{sgn}(\hat{\lambda}_{n,j}) = \mathrm{sgn}\{\bar{g}_j^{(\mathcal{T})}(\boldsymbol{\psi}_n)\}$ *for any* $j \in \mathcal{D}_{\boldsymbol{\psi}_n}(c)$ *with* $\hat{\lambda}_{n,j} \neq 0$, *where* $\hat{\boldsymbol{\lambda}}(\boldsymbol{\psi}_n) = (\hat{\lambda}_{n,1}, \ldots, \hat{\lambda}_{n,r})^{\mathrm{T}}$.

Lemma A.3. *Let* $P_{2,\nu}(\cdot) \in \mathscr{P}$ *be a convex function for* $\mathscr{P}$ *defined as* (3.2). *Assume that all the eigenvalues of* $\widehat{\mathbf{V}}^{(\mathcal{T})}_{\mathcal{M}_{\boldsymbol{\psi}_0}(c)}(\boldsymbol{\psi}_0)$ *are uniformly bounded away from zero and infinity w.p.a.1 for some* $c \in (C_*, 1)$, *and* $\max_{j\in\mathcal{T}} n^{-1}\sum_{i=1}^n |g_{i,j}^{(\mathcal{T})}(\boldsymbol{\psi}_0)|^\gamma = O_{\mathrm{p}}(1)$. *If* $\log r = o(n^{1/3})$ *and* $r_1\aleph_n = o[\min\{n^{-1/\gamma}, \nu\}]$, *then w.p.a.1 there is a sparse global maximizer* $\hat{\boldsymbol{\lambda}}(\boldsymbol{\psi}_0)$ *for* $f(\boldsymbol{\lambda};\boldsymbol{\psi}_0)$ *satisfying* $\mathrm{supp}\{\hat{\boldsymbol{\lambda}}_{\mathcal{D}}(\boldsymbol{\psi}_0)\} \subset \mathcal{D}_{\boldsymbol{\psi}_0}(c)$.

Now we begin to prove Proposition 3.1. Recall $S_n(\boldsymbol{\psi}) = \max_{\boldsymbol{\lambda}\in\hat{\Lambda}_n^{(\mathcal{T})}(\boldsymbol{\psi})} f(\boldsymbol{\lambda};\boldsymbol{\psi}) + \sum_{k\in\mathcal{D}} P_{1,\pi}(|\xi_k|)$ and $a_n = \sum_{k\in\mathcal{D}} P_{1,\pi}(|\xi_{0,k}|)$. Then $S_n(\boldsymbol{\psi}_0) = f\{\hat{\boldsymbol{\lambda}}(\boldsymbol{\psi}_0);\boldsymbol{\psi}_0\} + a_n$, where $\hat{\boldsymbol{\lambda}}(\boldsymbol{\psi}_0) = \arg\max_{\boldsymbol{\lambda}\in\hat{\Lambda}_n^{(\mathcal{T})}(\boldsymbol{\psi}_0)} f(\boldsymbol{\lambda};\boldsymbol{\psi}_0)$. Let $\mathcal{G} = \mathcal{I} \cup \mathrm{supp}\{\hat{\boldsymbol{\lambda}}_{\mathcal{D}}(\boldsymbol{\psi}_0)\}$ and write $\hat{\boldsymbol{\lambda}}(\boldsymbol{\psi}_0) = (\hat{\lambda}_1, \ldots, \hat{\lambda}_r)^{\mathrm{T}}$. It holds that

$$
\begin{aligned}
f\{\hat{\boldsymbol{\lambda}}(\boldsymbol{\psi}_0);\boldsymbol{\psi}_0\} &= \frac{1}{n}\sum_{i=1}^n \log\{1 + \hat{\boldsymbol{\lambda}}_{\mathcal{G}}(\boldsymbol{\psi}_0)^{\mathrm{T}}\mathbf{g}_{i,\mathcal{G}}^{(\mathcal{T})}(\boldsymbol{\psi}_0)\} - \sum_{j\in\mathcal{D}:\,\hat{\lambda}_j\neq 0} P_{2,\nu}(|\hat{\lambda}_j|) \\
&\le \max_{\boldsymbol{\eta}\in\hat{\Lambda}_n^\dagger(\boldsymbol{\psi}_0)} \frac{1}{n}\sum_{i=1}^n \log\{1 + \boldsymbol{\eta}^{\mathrm{T}}\mathbf{g}_{i,\mathcal{G}}^{(\mathcal{T})}(\boldsymbol{\psi}_0)\}\,, \qquad \text{(A.10)}
\end{aligned}
$$

where $\hat{\Lambda}_n^\dagger(\boldsymbol{\psi}_0) = \{\boldsymbol{\eta} = (\eta_1, \ldots, \eta_{|\mathcal{G}|})^{\mathrm{T}} \in \mathbb{R}^{|\mathcal{G}|} : \boldsymbol{\eta}^{\mathrm{T}}\mathbf{g}_{i,\mathcal{G}}^{(\mathcal{T})}(\boldsymbol{\psi}_0) \in \mathcal{V} \text{ for any } i \in [n]\}$ for some open interval $\mathcal{V}$ containing zero. We will first prove that

$$
\max_{\boldsymbol{\eta}\in\hat{\Lambda}_n^\dagger(\boldsymbol{\psi}_0)} \frac{1}{n}\sum_{i=1}^n \log\{1 + \boldsymbol{\eta}^{\mathrm{T}}\mathbf{g}_{i,\mathcal{G}}^{(\mathcal{T})}(\boldsymbol{\psi}_0)\} = O_{\mathrm{p}}(r_1\aleph_n^2)\,. \qquad \text{(A.11)}
$$

Based on (A.11), we have $f\{\hat{\boldsymbol{\lambda}}(\boldsymbol{\psi}_0);\boldsymbol{\psi}_0\} = O_{\mathrm{p}}(r_1\aleph_n^2)$. Let $A_n(\boldsymbol{\psi},\boldsymbol{\eta}) = n^{-1}\sum_{i=1}^n \log\{1 + \boldsymbol{\eta}^{\mathrm{T}}\mathbf{g}_{i,\mathcal{G}}^{(\mathcal{T})}(\boldsymbol{\psi})\}$ and $\tilde{\boldsymbol{\eta}} = \arg\max_{\boldsymbol{\eta}\in\hat{\Lambda}_n^\dagger(\boldsymbol{\psi}_0)} A_n(\boldsymbol{\psi}_0,\boldsymbol{\eta})$. As we have shown in the proof of Lemma A.3 that $|\mathcal{M}_{\boldsymbol{\psi}_0}(c)| \le 2r_1$ w.p.a.1, it then follows from Lemma A.3 that $|\mathcal{G}| \le |\mathcal{M}_{\boldsymbol{\psi}_0}(c)| \le 2r_1$ w.p.a.1. Pick $\delta_n =$

$o(r_1^{-1/2}n^{-1/\gamma})$ and $r_1^{1/2}\aleph_n = o(\delta_n)$, which can be guaranteed by $r_1\aleph_n = o(n^{-1/\gamma})$. Define $\Lambda_n = \{\boldsymbol{\eta} \in \mathbb{R}^{|\mathcal{G}|} : |\boldsymbol{\eta}|_2 \le \delta_n\}$ and let $\bar{\boldsymbol{\eta}} = \arg\max_{\boldsymbol{\eta}\in\Lambda_n} A_n(\boldsymbol{\psi}_0, \boldsymbol{\eta})$. It follows from the last requirement of Condition 3 that $\max_{i\in[n]} |\mathbf{g}_{i,\mathcal{G}}^{(\mathcal{T})}(\boldsymbol{\psi}_0)|_2 = O_{\mathrm{p}}(r_1^{1/2}n^{1/\gamma})$, which implies that $\max_{i\in[n]} \sup_{\boldsymbol{\eta}\in\Lambda_n} |\boldsymbol{\eta}^{\mathrm{T}}\mathbf{g}_{i,\mathcal{G}}^{(\mathcal{T})}(\boldsymbol{\psi}_0)| = o_{\mathrm{p}}(1)$. By the Taylor expansion, it holds w.p.a.1 that

$$\begin{aligned} 0 = A_n(\boldsymbol{\psi}_0, \mathbf{0}) \le A_n(\boldsymbol{\psi}_0, \bar{\boldsymbol{\eta}}) &= \bar{\boldsymbol{\eta}}^{\mathrm{T}}\bar{\mathbf{g}}_{\mathcal{G}}^{(\mathcal{T})}(\boldsymbol{\psi}_0) - \frac{1}{2n}\sum_{i=1}^{n} \frac{\bar{\boldsymbol{\eta}}^{\mathrm{T}}\mathbf{g}_{i,\mathcal{G}}^{(\mathcal{T})}(\boldsymbol{\psi}_0)^{\otimes 2}\bar{\boldsymbol{\eta}}}{\{1 + \bar{c}\bar{\boldsymbol{\eta}}^{\mathrm{T}}\mathbf{g}_{i,\mathcal{G}}^{(\mathcal{T})}(\boldsymbol{\psi}_0)\}^2} \\ &\le |\bar{\boldsymbol{\eta}}|_2 |\bar{\mathbf{g}}_{\mathcal{G}}^{(\mathcal{T})}(\boldsymbol{\psi}_0)|_2 - C|\bar{\boldsymbol{\eta}}|_2^2\{1 + o_{\mathrm{p}}(1)\} \end{aligned} \tag{A.12}$$

for some $\bar{c} \in (0, 1)$, where the last inequality is implied by Condition 4 and Lemma A.1. As we have shown in the proof of Lemma A.3 that $|\bar{\mathbf{g}}^{(\mathcal{T})}(\boldsymbol{\psi}_0)|_\infty = O_{\mathrm{p}}(\aleph_n)$, then $|\bar{\mathbf{g}}_{\mathcal{G}}^{(\mathcal{T})}(\boldsymbol{\psi}_0)|_2 = O_{\mathrm{p}}(r_1^{1/2}\aleph_n)$. It follows from (A.12) that $|\bar{\boldsymbol{\eta}}|_2 = O_{\mathrm{p}}(r_1^{1/2}\aleph_n) = o_{\mathrm{p}}(\delta_n)$. Hence, $\bar{\boldsymbol{\eta}} \in \mathrm{int}(\Lambda_n)$ w.p.a.1. Since $\Lambda_n \subset \hat{\Lambda}_n^{\dagger}(\boldsymbol{\psi}_0)$ w.p.a.1, by the concavity of $A_n(\boldsymbol{\psi}_0, \boldsymbol{\eta})$ and the convexity of $\hat{\Lambda}_n^{\dagger}(\boldsymbol{\psi}_0)$, we have $\tilde{\boldsymbol{\eta}} = \bar{\boldsymbol{\eta}}$ w.p.a.1. Then we can obtain (A.11) from (A.12).

Recall that $b_{1,n} = \max\{a_n, r_1\aleph_n^2\}$. Then $S_n(\boldsymbol{\psi}_0) = O_{\mathrm{p}}(r_1\aleph_n^2) + a_n = O_{\mathrm{p}}(b_{1,n})$. Notice that $\hat{\boldsymbol{\psi}} = \arg\min_{\boldsymbol{\psi}\in\boldsymbol{\Psi}_*} S_n(\boldsymbol{\psi})$ with $\boldsymbol{\Psi}_* = \{\boldsymbol{\psi} = (\boldsymbol{\psi}_{\mathcal{S}}^{\mathrm{T}}, \boldsymbol{\psi}_{\mathcal{S}^{\mathrm{c}}}^{\mathrm{T}})^{\mathrm{T}} : |\boldsymbol{\psi}_{\mathcal{S}} - \boldsymbol{\psi}_{0,\mathcal{S}}|_\infty \le \varepsilon, |\boldsymbol{\psi}_{\mathcal{S}^{\mathrm{c}}}|_1 \le \aleph_n\}$, and $\boldsymbol{\psi}_{0,\mathcal{S}^{\mathrm{c}}} = \mathbf{0}$. We then have $\boldsymbol{\psi}_0 \in \boldsymbol{\Psi}_*$ which implies that $S_n(\hat{\boldsymbol{\psi}}) \le S_n(\boldsymbol{\psi}_0) = O_{\mathrm{p}}(b_{1,n})$. We need to show $\hat{\boldsymbol{\psi}} \in \mathrm{int}(\boldsymbol{\Psi}_*)$ w.p.a.1, which indicates that $\hat{\boldsymbol{\psi}}$ is a local minimizer of $S_n(\boldsymbol{\psi})$. Our proof includes three parts: (i) to show that for any $\epsilon_n \to \infty$ satisfying $b_{1,n}\epsilon_n^2 n^{2/\gamma} = o(1)$ and any $\boldsymbol{\psi} = (\boldsymbol{\theta}^{\mathrm{T}}, \boldsymbol{\xi}^{\mathrm{T}})^{\mathrm{T}} \in \boldsymbol{\Psi}_*$ satisfying $|\boldsymbol{\theta} - \boldsymbol{\theta}_0|_\infty > \epsilon_n b_{1,n}^{1/2}$, there exists a universal constant $K > 0$ independent of $\boldsymbol{\psi}$ such that $\mathbb{P}\{S_n(\boldsymbol{\psi}) > Kb_{1,n}\epsilon_n^2\} \to 1$ as $n \to \infty$. Due to $b_{1,n} = o(n^{-2/\gamma})$, we can select an arbitrary slowly diverging ϵ_n satisfying $b_{1,n}\epsilon_n^2 n^{2/\gamma} = o(1)$. Thus, we have $|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0|_\infty = O_{\mathrm{p}}(b_{1,n}^{1/2})$; (ii) letting $b_{2,n} = \max\{b_{1,n}, \nu^2\}$ and $\phi_n = \max\{pb_{1,n}^{1/2}, b_{2,n}^{1/2}\}$, to show that for any $\varepsilon_n \to \infty$ satisfying $b_{2,n}\varepsilon_n^2 n^{2/\gamma} = o(1)$ and $\boldsymbol{\psi} = (\boldsymbol{\theta}^{\mathrm{T}}, \boldsymbol{\xi}_{\mathcal{A}}^{\mathrm{T}}, \boldsymbol{\xi}_{\mathcal{A}^{\mathrm{c}}}^{\mathrm{T}})^{\mathrm{T}} \in \boldsymbol{\Psi}_*$ satisfying $|\boldsymbol{\theta} - \boldsymbol{\theta}_0|_\infty \le O(\varepsilon_n^{1/2}b_{1,n}^{1/2})$ and $|\boldsymbol{\xi}_{\mathcal{A}^{\mathrm{c}}} - \boldsymbol{\xi}_{0,\mathcal{A}^{\mathrm{c}}}|_\infty > \varepsilon_n\phi_n$, there exists a universal constant $M > 0$ independent of $\boldsymbol{\psi}$ such that $\mathbb{P}\{S_n(\boldsymbol{\psi}) > Mb_{2,n}\varepsilon_n^2\} \to 1$ as $n \to \infty$. Recall $\hat{\boldsymbol{\psi}} = (\hat{\boldsymbol{\theta}}^{\mathrm{T}}, \hat{\boldsymbol{\xi}}_{\mathcal{A}}^{\mathrm{T}}, \hat{\boldsymbol{\xi}}_{\mathcal{A}^{\mathrm{c}}}^{\mathrm{T}})^{\mathrm{T}}$. Due to $|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}|_\infty = O_{\mathrm{p}}(b_{1,n}^{1/2})$, we know $|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0| \le O(\varepsilon_n^{1/2}b_{1,n}^{1/2})$ w.p.a.1. Since we can select an arbitrary slowly diverging ε_n satisfying $b_{2,n}\varepsilon_n^2 n^{2/\gamma} = o(1)$, it holds that $|\hat{\boldsymbol{\xi}}_{\mathcal{A}^{\mathrm{c}}} - \boldsymbol{\xi}_{0,\mathcal{A}^{\mathrm{c}}}|_\infty = O_{\mathrm{p}}(\phi_n)$; (iii) to show that $\hat{\boldsymbol{\psi}}_{\mathcal{S}^{\mathrm{c}}} = \mathbf{0}$ w.p.a.1.

Proof of Part (i). The proof is similar to that for Part (i) of Proposition A.1. For any $\boldsymbol{\psi} = (\boldsymbol{\theta}^{\mathrm{T}}, \boldsymbol{\xi}^{\mathrm{T}})^{\mathrm{T}} \in \boldsymbol{\Psi}_*$ satisfying $|\boldsymbol{\theta} - \boldsymbol{\theta}_0|_\infty > \epsilon_n b_{1,n}^{1/2}$, let $j_0 = \arg\max_{j\in\mathcal{I}} |\mathbb{E}\{g_{i,j}^{(\mathcal{I})}(\boldsymbol{\theta})\}|$ and $\mu_{j_0} = \mathbb{E}\{g_{i,j_0}^{(\mathcal{I})}(\boldsymbol{\theta})\}$. Select $\tilde{\boldsymbol{\lambda}} = \delta b_{1,n}^{1/2}\epsilon_n \mathbf{e}_{j_0}$, where $\delta > 0$ is a sufficiently small constant, and $\mathbf{e}_{j_0}$ is an r-dimensional vector with the j_0-th component being 1 and other components being 0. Then $\tilde{\boldsymbol{\lambda}} \in \hat{\Lambda}_n^{(\mathcal{T})}(\boldsymbol{\psi})$ w.p.a.1. Without loss of generality, we assume that $\mu_{j_0} > 0$. Write $\tilde{\boldsymbol{\lambda}} = (\tilde{\lambda}_1, \ldots, \tilde{\lambda}_r)^{\mathrm{T}}$. Notice that $j_0 \notin \mathcal{D}$. By the Taylor expansion, it holds w.p.a.1 that $S_n(\boldsymbol{\psi}) \ge n^{-1}\sum_{i=1}^{n} \log\{1 + \tilde{\boldsymbol{\lambda}}^{\mathrm{T}}\mathbf{g}_i^{(\mathcal{T})}(\boldsymbol{\psi})\} -$

$\sum_{j\in\mathcal{D}} P_{2,\nu}(|\tilde{\lambda}_j|) \geq n^{-1}\sum_{i=1}^n \tilde{\lambda}_{j_0} g_{i,j_0}^{(\mathcal{I})}(\boldsymbol{\theta}) - n^{-1}\sum_{i=1}^n \{\tilde{\lambda}_{j_0} g_{i,j_0}^{(\mathcal{I})}(\boldsymbol{\theta})\}^2$. Thus,

$$\begin{aligned}\mathbb{P}\{S_n(\boldsymbol{\psi}) \leq Kb_{1,n}\epsilon_n^2\} &\leq \mathbb{P}\bigg[\frac{1}{n}\sum_{i=1}^n \tilde{\lambda}_{j_0} g_{i,j_0}^{(\mathcal{I})}(\boldsymbol{\theta}) - \frac{1}{n}\sum_{i=1}^n \{\tilde{\lambda}_{j_0} g_{i,j_0}^{(\mathcal{I})}(\boldsymbol{\theta})\}^2 \leq Kb_{1,n}\epsilon_n^2\bigg] + o(1) \\ &\leq \mathbb{P}\bigg[\bar{g}_{j_0}^{(\mathcal{I})}(\boldsymbol{\theta}) - \mu_{j_0} \leq b_{1,n}^{1/2}\epsilon_n\bigg\{\frac{K}{\delta} + \frac{\delta}{n}\sum_{i=1}^n |g_{i,j_0}^{(\mathcal{I})}(\boldsymbol{\theta})|^2\bigg\} - \mu_{j_0}\bigg] + o(1)\,.\end{aligned}$$

Using the same arguments stated in the proof of Proposition A.1, we have $\mathbb{P}\{S_n(\boldsymbol{\psi}) > Kb_{1,n}\epsilon_n^2\} \to 1$ as $n \to \infty$. We complete the proof of Part (i).

Proof of Part (ii). The proof is also similar to that for Part (i) of Proposition A.1. For any $\boldsymbol{\psi} = (\boldsymbol{\theta}^{\mathrm{T}}, \boldsymbol{\xi}^{\mathrm{T}})^{\mathrm{T}} \in \boldsymbol{\Psi}_*$ with $\boldsymbol{\xi} = (\boldsymbol{\xi}_{\mathcal{A}}^{\mathrm{T}}, \boldsymbol{\xi}_{\mathcal{A}^{\mathrm{c}}}^{\mathrm{T}})^{\mathrm{T}}$ satisfying $|\boldsymbol{\theta}-\boldsymbol{\theta}_0|_\infty \leq O(\varepsilon_n^{1/2} b_{1,n}^{1/2})$ and $|\boldsymbol{\xi}_{\mathcal{A}^{\mathrm{c}}} - \boldsymbol{\xi}_{0,\mathcal{A}^{\mathrm{c}}}|_\infty > \varepsilon_n\phi_n$, let $j_0 = \arg\max_{j\in\mathcal{A}^{\mathrm{c}}} |\xi_j - \xi_{0,j}|$ and $\mu_{j_0} = \mathbb{E}\{g_{i,j_0}^{(\mathcal{T})}(\boldsymbol{\psi})\}$. Without loss of generality, we assume $\xi_{0,j_0} - \xi_{j_0} > 0$. Select $\tilde{\boldsymbol{\lambda}} = \delta b_{2,n}^{1/2}\varepsilon_n \mathbf{e}_{j_0}$, where $\delta > 0$ is a sufficiently small constant, and $\mathbf{e}_{j_0}$ is similarly defined as that in the proof of Part (i). Then $\tilde{\boldsymbol{\lambda}} \in \hat{\Lambda}_n^{(\mathcal{T})}(\boldsymbol{\psi})$ w.p.a.1. By the Taylor expansion, it holds w.p.a.1 that $S_n(\boldsymbol{\psi}) \geq n^{-1}\sum_{i=1}^n \tilde{\lambda}_{j_0} g_{i,j_0}^{(\mathcal{T})}(\boldsymbol{\psi}) - n^{-1}\sum_{i=1}^n \{\tilde{\lambda}_{j_0} g_{i,j_0}^{(\mathcal{T})}(\boldsymbol{\psi})\}^2 - P_{2,\nu}(|\tilde{\lambda}_{j_0}|) \geq n^{-1}\sum_{i=1}^n \tilde{\lambda}_{j_0} g_{i,j_0}^{(\mathcal{T})}(\boldsymbol{\psi}) - n^{-1}\sum_{i=1}^n \{\tilde{\lambda}_{j_0} g_{i,j_0}^{(\mathcal{T})}(\boldsymbol{\psi})\}^2 - C\nu\tilde{\lambda}_{j_0}$. Thus,

$$\begin{aligned}\mathbb{P}\{S_n(\boldsymbol{\psi}) \leq Mb_{2,n}\varepsilon_n^2\} &\leq \mathbb{P}\bigg[\frac{1}{n}\sum_{i=1}^n \tilde{\lambda}_{j_0} g_{i,j_0}^{(\mathcal{T})}(\boldsymbol{\psi}) - \frac{1}{n}\sum_{i=1}^n \{\tilde{\lambda}_{j_0} g_{i,j_0}^{(\mathcal{T})}(\boldsymbol{\psi})\}^2 - C\nu\tilde{\lambda}_{j_0} \leq Mb_{2,n}\varepsilon_n^2\bigg] + o(1) \\ &\leq \mathbb{P}\bigg[\bar{g}_{j_0}^{(\mathcal{T})}(\boldsymbol{\psi}) - \mu_{j_0} \leq b_{2,n}^{1/2}\varepsilon_n\bigg\{\frac{M}{\delta} + \frac{\delta}{n}\sum_{i=1}^n |g_{i,j_0}^{(\mathcal{T})}(\boldsymbol{\psi})|^2\bigg\} + C\nu - \mu_{j_0}\bigg] + o(1)\,.\end{aligned}$$

By the Taylor expansion and Condition 3, $|\mathbb{E}\{g_{i,j_0}^{(\mathcal{D})}(\boldsymbol{\theta})\} - \mathbb{E}\{g_{i,j_0}^{(\mathcal{D})}(\boldsymbol{\theta}_0)\}| \leq |\mathbb{E}\{\nabla_{\boldsymbol{\theta}} g_{i,j_0}^{(\mathcal{D})}(\dot{\boldsymbol{\theta}})\}|_\infty |\boldsymbol{\theta} - \boldsymbol{\theta}_0|_1 \leq O(\varepsilon_n^{1/2} p b_{1,n}^{1/2})$. Recall $b_{2,n} = \max\{b_{1,n}, \nu^2\}$ and $\phi_n = \max\{pb_{1,n}^{1/2}, b_{2,n}^{1/2}\}$. Then

$$\mu_{j_0} = \mathbb{E}\{g_{i,j_0}^{(\mathcal{D})}(\boldsymbol{\theta})\} - \mathbb{E}\{g_{i,j_0}^{(\mathcal{D})}(\boldsymbol{\theta}_0)\} + \xi_{0,j_0} - \xi_{j_0} \geq \varepsilon_n\phi_n - O(\varepsilon_n^{1/2} p b_{1,n}^{1/2}) \geq \varepsilon_n b_{2,n}^{1/2}/2 \tag{A.13}$$

when n is sufficiently large. Using the same arguments stated in the proof of Proposition A.1, we have $\mathbb{P}\{S_n(\boldsymbol{\psi}) > Mb_{2,n}\varepsilon_n^2\} \to 1$ as $n \to \infty$. We complete the proof of Part (ii).

Proof of Part (iii). If $\hat{\boldsymbol{\psi}}_{\mathcal{S}^{\mathrm{c}}} \neq \mathbf{0}$, we define $\hat{\boldsymbol{\psi}}^* = (\hat{\boldsymbol{\psi}}_{\mathcal{S}}^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}}$ and will show $S_n(\hat{\boldsymbol{\psi}}^*) < S_n(\hat{\boldsymbol{\psi}})$ w.p.a.1. This contradicts the definition of $\hat{\boldsymbol{\psi}}$. Then we have $\hat{\boldsymbol{\psi}}_{\mathcal{S}^{\mathrm{c}}} = \mathbf{0}$ w.p.a.1. Write $\hat{\boldsymbol{\psi}} = (\hat{\psi}_1, \ldots, \hat{\psi}_{p+r_2})^{\mathrm{T}} = (\hat{\theta}_1, \ldots, \hat{\theta}_p, \hat{\xi}_1, \ldots, \hat{\xi}_{r_2})^{\mathrm{T}}$ and $\boldsymbol{\psi}_0 = (\theta_{0,1}, \ldots, \theta_{0,p}, \xi_{0,1}, \ldots, \xi_{0,r_2})^{\mathrm{T}}$. Recall $\xi_{0,k} = 0$ for any $k \in \mathcal{A}$ and $\xi_{0,k} \neq 0$ for any $k \in \mathcal{A}^{\mathrm{c}}$. As shown in Part (ii) that $\max_{k\in\mathcal{A}^{\mathrm{c}}} |\hat{\xi}_k - \xi_{0,k}| = O_{\mathrm{p}}(\phi_n)$,

due to $\phi_n = o(\min_{k\in\mathcal{A}^{\mathrm{c}}}|\xi_{0,k}|)$ and (3.5), (A.10) and (A.11) imply that

$$\begin{aligned}\max_{\boldsymbol{\lambda}\in\hat{\Lambda}_n^{(\mathcal{T})}(\hat{\boldsymbol{\psi}})} f(\boldsymbol{\lambda};\hat{\boldsymbol{\psi}}) &\le \max_{\boldsymbol{\lambda}\in\hat{\Lambda}_n^{(\mathcal{T})}(\boldsymbol{\psi}_0)} f(\boldsymbol{\lambda};\boldsymbol{\psi}_0) + \sum_{k\in\mathcal{D}} P_{1,\pi}(|\xi_{0,k}|) - \sum_{k\in\mathcal{D}} P_{1,\pi}(|\hat{\xi}_k|) \\ &\le O_{\mathrm{p}}(r_1\aleph_n^2) + \sum_{k\in\mathcal{A}^c} P_{1,\pi}(|\xi_{0,k}|) - \sum_{k\in\mathcal{A}^c} P_{1,\pi}(|\hat{\xi}_k|) \qquad (A.14) \\ &= O_{\mathrm{p}}(r_1\aleph_n^2)\,.\end{aligned}$$

Notice that $\ell_n\aleph_n = o(n^{-1/\gamma})$. We pick $\delta_n = o(\ell_n^{-1/2}n^{-1/\gamma})$ and $\ell_n^{1/2}\aleph_n = o(\delta_n)$. Recall $\mathcal{M}_{\boldsymbol{\psi}}(c) = \mathcal{I}\cup\mathcal{D}_{\boldsymbol{\psi}}(c)$ with $\mathcal{D}_{\boldsymbol{\psi}}(c) = \{j\in\mathcal{D}: |\bar{g}_j^{(\mathcal{T})}(\boldsymbol{\psi})| \ge c\nu\rho_2'(0^+)\}$ for any $c\in(C_*,1)$. Define

$$\boldsymbol{\beta}_{\mathcal{M}_{\hat{\boldsymbol{\psi}}}(\tilde{c}_1)}(\hat{\boldsymbol{\psi}}) := \begin{pmatrix} \bar{\mathbf{g}}^{(\mathcal{I})}(\hat{\boldsymbol{\psi}}) \\ \bar{\mathbf{g}}^{(\mathcal{T})}_{\mathcal{D}_{\hat{\boldsymbol{\psi}}}(\tilde{c}_1)}(\hat{\boldsymbol{\psi}}) - \nu\rho_2'(0^+)\mathrm{sgn}\{\bar{\mathbf{g}}^{(\mathcal{T})}_{\mathcal{D}_{\hat{\boldsymbol{\psi}}}(\tilde{c}_1)}(\hat{\boldsymbol{\psi}})\} \end{pmatrix}$$

for some $\tilde{c}_1\in(C_*,1)$. Select $\boldsymbol{\lambda}^*$ satisfying $\boldsymbol{\lambda}^*_{\mathcal{M}_{\hat{\boldsymbol{\psi}}}(\tilde{c}_1)} = \delta_n\boldsymbol{\beta}_{\mathcal{M}_{\hat{\boldsymbol{\psi}}}(\tilde{c}_1)}(\hat{\boldsymbol{\psi}})/|\boldsymbol{\beta}_{\mathcal{M}_{\hat{\boldsymbol{\psi}}}(\tilde{c}_1)}(\hat{\boldsymbol{\psi}})|_2$ and $\boldsymbol{\lambda}^*_{\mathcal{M}^{\mathrm{c}}_{\hat{\boldsymbol{\psi}}}(\tilde{c}_1)} = \mathbf{0}$. Since $|\mathcal{M}_{\hat{\boldsymbol{\psi}}}(\tilde{c}_1)| \le \ell_n$ w.p.a.1, it holds that $\max_{i\in[n]}|\boldsymbol{\lambda}^{*,\mathrm{T}}\mathbf{g}_i^{(\mathcal{T})}(\hat{\boldsymbol{\psi}})| \le |\boldsymbol{\lambda}^*_{\mathcal{M}_{\hat{\boldsymbol{\psi}}}(\tilde{c}_1)}|_2 \max_{i\in[n]}|\mathbf{g}^{(\mathcal{T})}_{i,\mathcal{M}_{\hat{\boldsymbol{\psi}}}(\tilde{c}_1)}(\hat{\boldsymbol{\psi}})|_2 = o(\ell_n^{-1/2}n^{-1/\gamma})\cdot O_{\mathrm{p}}(\ell_n^{1/2}n^{1/\gamma}) = o_{\mathrm{p}}(1)$, which indicates that $\boldsymbol{\lambda}^*\in\hat{\Lambda}_n^{(\mathcal{T})}(\hat{\boldsymbol{\psi}})$ w.p.a.1. Write $\boldsymbol{\lambda}^* = (\lambda_1^*,\ldots,\lambda_r^*)^{\mathrm{T}}$.

Recall $P_{2,\nu}(t) = \nu\rho_2(t;\nu)$ for any $t\ge 0$. Notice that $\mathbb{P}[\cup_{j\in\mathcal{T}}\{|\bar{g}_j^{(\mathcal{T})}(\hat{\boldsymbol{\psi}})|\in[\tilde{c}\nu\rho_2'(0^+),\nu\rho_2'(0^+))\}] \to 0$ for some constant $\tilde{c}\in(C_*,1)$. Then $\{j\in\mathcal{T}:\ \tilde{c}_1\nu\rho_2'(0^+)\le|\bar{g}_j^{(\mathcal{T})}(\hat{\boldsymbol{\psi}})|<\nu\rho_2'(0^+)\} = \emptyset$ w.p.a.1 by letting $\tilde{c}_1 = \tilde{c}$. Notice that $r_1\lesssim\ell_n$. By the Taylor expansion, it holds w.p.a.1 that

$$\begin{aligned} o_{\mathrm{p}}(\delta_n^2) &= \max_{\boldsymbol{\lambda}\in\hat{\Lambda}_n^{(\mathcal{T})}(\hat{\boldsymbol{\psi}})} f(\boldsymbol{\lambda};\hat{\boldsymbol{\psi}}) \\ &\ge \frac{1}{n}\sum_{i=1}^n \log\{1+\boldsymbol{\lambda}^{*,\mathrm{T}}_{\mathcal{M}_{\hat{\boldsymbol{\psi}}}(\tilde{c}_1)}\mathbf{g}^{(\mathcal{T})}_{i,\mathcal{M}_{\hat{\boldsymbol{\psi}}}(\tilde{c}_1)}(\hat{\boldsymbol{\psi}})\} - \sum_{j\in\mathcal{D}_{\hat{\boldsymbol{\psi}}}(\tilde{c}_1)} P_{2,\nu}(|\lambda_j^*|) \\ &= \boldsymbol{\lambda}^{*,\mathrm{T}}_{\mathcal{M}_{\hat{\boldsymbol{\psi}}}(\tilde{c}_1)}\bar{\mathbf{g}}^{(\mathcal{T})}_{\mathcal{M}_{\hat{\boldsymbol{\psi}}}(\tilde{c}_1)}(\hat{\boldsymbol{\psi}}) - \frac{1}{2n}\sum_{i=1}^n \frac{\boldsymbol{\lambda}^{*,\mathrm{T}}_{\mathcal{M}_{\hat{\boldsymbol{\psi}}}(\tilde{c}_1)}\mathbf{g}^{(\mathcal{T})}_{i,\mathcal{M}_{\hat{\boldsymbol{\psi}}}(\tilde{c}_1)}(\hat{\boldsymbol{\psi}})^{\otimes 2}\boldsymbol{\lambda}^*_{\mathcal{M}_{\hat{\boldsymbol{\psi}}}(\tilde{c}_1)}}{\{1+c^*\boldsymbol{\lambda}^{*,\mathrm{T}}_{\mathcal{M}_{\hat{\boldsymbol{\psi}}}(\tilde{c}_1)}\mathbf{g}^{(\mathcal{T})}_{i,\mathcal{M}_{\hat{\boldsymbol{\psi}}}(\tilde{c}_1)}(\hat{\boldsymbol{\psi}})\}^2} \\ &\quad - \sum_{j\in\mathcal{D}_{\hat{\boldsymbol{\psi}}}(\tilde{c}_1)} \nu\rho_2'(0^+)|\lambda_j^*| - \frac{1}{2}\sum_{j\in\mathcal{D}_{\hat{\boldsymbol{\psi}}}(\tilde{c}_1)} \nu\rho_2''(c_j|\lambda_j^*|;\nu)|\lambda_j^*|^2 \\ &\ge \boldsymbol{\lambda}^{*,\mathrm{T}}_{\mathcal{M}_{\hat{\boldsymbol{\psi}}}(\tilde{c}_1)}\boldsymbol{\beta}_{\mathcal{M}_{\hat{\boldsymbol{\psi}}}(\tilde{c}_1)}(\hat{\boldsymbol{\psi}}) - C\delta_n^2\{1+o_{\mathrm{p}}(1)\} - 2\nu\rho_2'(0^+) \sum_{j\in\mathcal{T}:\,\tilde{c}_1\nu\rho_2'(0^+)\le|\bar{g}_j^{(\mathcal{T})}(\hat{\boldsymbol{\psi}})|<\nu\rho_2'(0^+)} |\lambda_j^*| \\ &\ge \delta_n|\boldsymbol{\beta}_{\mathcal{M}_{\hat{\boldsymbol{\psi}}}(\tilde{c}_1)}(\hat{\boldsymbol{\psi}})|_2 - C\delta_n^2\{1+o_{\mathrm{p}}(1)\} \end{aligned}$$

for some $c^*, c_j\in(0,1)$. Thus, $|\boldsymbol{\beta}_{\mathcal{M}_{\hat{\boldsymbol{\psi}}}(\tilde{c}_1)}(\hat{\boldsymbol{\psi}})|_2 = O_{\mathrm{p}}(\delta_n)$. For any $\epsilon_n\to 0$, choose $\boldsymbol{\lambda}^{**}$ satisfying

$\boldsymbol{\lambda}^{**}_{\mathcal{M}_{\hat{\boldsymbol{\psi}}}(\tilde{c}_1)} = \epsilon_n \boldsymbol{\beta}_{\mathcal{M}_{\hat{\boldsymbol{\psi}}}(\tilde{c}_1)}(\hat{\boldsymbol{\psi}})$ and $\boldsymbol{\lambda}^{**}_{\mathcal{M}^{\mathrm{c}}_{\hat{\boldsymbol{\psi}}}(\tilde{c}_1)} = \mathbf{0}$. Then, $|\boldsymbol{\lambda}^{**}|_2 = o_{\mathrm{p}}(\delta_n)$. Using the same arguments given above, we can obtain $\epsilon_n |\boldsymbol{\beta}_{\mathcal{M}_{\hat{\boldsymbol{\psi}}}(\tilde{c}_1)}(\hat{\boldsymbol{\psi}})|_2^2 - C\epsilon_n^2 |\boldsymbol{\beta}_{\mathcal{M}_{\hat{\boldsymbol{\psi}}}(\tilde{c}_1)}(\hat{\boldsymbol{\psi}})|_2^2 \{1 + o_{\mathrm{p}}(1)\} = O_{\mathrm{p}}(r_1 \aleph_n^2)$, which implies that $\epsilon_n |\boldsymbol{\beta}_{\mathcal{M}_{\hat{\boldsymbol{\psi}}}(\tilde{c}_1)}(\hat{\boldsymbol{\psi}})|_2^2 = O_{\mathrm{p}}(r_1 \aleph_n^2)$. Since we can select an arbitrary slow $\epsilon_n \to 0$, we have $|\boldsymbol{\beta}_{\mathcal{M}_{\hat{\boldsymbol{\psi}}}(\tilde{c}_1)}(\hat{\boldsymbol{\psi}})|_2^2 = O_{\mathrm{p}}(r_1 \aleph_n^2)$ following a standard result from probability theory. Then Lemmas A.1 and A.2 imply that $|\hat{\boldsymbol{\lambda}}(\hat{\boldsymbol{\psi}})|_2 = O_{\mathrm{p}}(r_1^{1/2} \aleph_n)$. Recall $\hat{\boldsymbol{\lambda}}(\boldsymbol{\psi}) = \arg\max_{\boldsymbol{\lambda} \in \hat{\Lambda}_n^{(\mathcal{T})}(\boldsymbol{\psi})} f(\boldsymbol{\lambda}; \boldsymbol{\psi})$. Write $\hat{\boldsymbol{\lambda}} = \hat{\boldsymbol{\lambda}}(\hat{\boldsymbol{\psi}})$ and $\hat{\boldsymbol{\lambda}}^* = \hat{\boldsymbol{\lambda}}(\hat{\boldsymbol{\psi}}^*)$. Notice that $S_n(\boldsymbol{\psi}) = \max_{\boldsymbol{\lambda} \in \hat{\Lambda}_n^{(\mathcal{T})}(\boldsymbol{\psi})} f(\boldsymbol{\lambda}; \boldsymbol{\psi}) + \sum_{k \in \mathcal{D}} P_{1,\pi}(|\xi_k|)$ and $\mathcal{S} = \mathcal{P} \cup \mathcal{A}^{\mathrm{c}}$. Then

$$
\begin{aligned}
S_n(\hat{\boldsymbol{\psi}}^*) - S_n(\hat{\boldsymbol{\psi}}) &= f(\hat{\boldsymbol{\lambda}}^*; \hat{\boldsymbol{\psi}}^*) - f(\hat{\boldsymbol{\lambda}}; \hat{\boldsymbol{\psi}}) - \sum_{k \in \mathcal{A}} P_{1,\pi}(|\hat{\xi}_k|) \\
&\leq f(\hat{\boldsymbol{\lambda}}^*; \hat{\boldsymbol{\psi}}^*) - f(\hat{\boldsymbol{\lambda}}^*; \hat{\boldsymbol{\psi}}) - \sum_{k \in \mathcal{A}} P_{1,\pi}(|\hat{\xi}_k|) \qquad \text{(A.15)} \\
&= \frac{1}{n} \sum_{i=1}^n \log\{1 + \hat{\boldsymbol{\lambda}}^{*,\mathrm{T}} \mathbf{g}_i^{(\mathcal{T})}(\hat{\boldsymbol{\psi}}^*)\} - \frac{1}{n} \sum_{i=1}^n \log\{1 + \hat{\boldsymbol{\lambda}}^{*,\mathrm{T}} \mathbf{g}_i^{(\mathcal{T})}(\hat{\boldsymbol{\psi}})\} - \sum_{k \in \mathcal{A}} P_{1,\pi}(|\hat{\xi}_k|)\,.
\end{aligned}
$$

It follows from the Taylor expansion that

$$
S_n(\hat{\boldsymbol{\psi}}^*) \leq S_n(\hat{\boldsymbol{\psi}}) - \underbrace{\frac{1}{n} \sum_{i=1}^n \frac{\hat{\boldsymbol{\lambda}}^{*,\mathrm{T}} \nabla_{\boldsymbol{\psi}_{\mathcal{S}^{\mathrm{c}}}} \mathbf{g}_i^{(\mathcal{T})}(\check{\boldsymbol{\psi}})}{1 + \hat{\boldsymbol{\lambda}}^{*,\mathrm{T}} \mathbf{g}_i^{(\mathcal{T})}(\check{\boldsymbol{\psi}})} \hat{\boldsymbol{\psi}}_{\mathcal{S}^c}}_{\mathrm{I}} - \underbrace{\sum_{k \in \mathcal{A}} P_{1,\pi}(|\hat{\xi}_k|)}_{\mathrm{II}}\,, \qquad \text{(A.16)}
$$

where $\check{\boldsymbol{\psi}}$ is on the jointing line between $\hat{\boldsymbol{\psi}}$ and $\hat{\boldsymbol{\psi}}^*$. We need to show $\mathrm{I} + \mathrm{II} > 0$ w.p.a.1.

To do this, we first use Lemma A.2 to bound $|\hat{\boldsymbol{\lambda}}^*|_2$. Given some $\tilde{c}_2 \in (\tilde{c}_1, 1)$, we define

$$
\boldsymbol{\beta}_{\mathcal{M}_{\hat{\boldsymbol{\psi}}^*(\tilde{c}_2)}}(\hat{\boldsymbol{\psi}}^*) := \begin{pmatrix} \bar{\mathbf{g}}^{(\mathcal{I})}(\hat{\boldsymbol{\psi}}^*) \\ \bar{\mathbf{g}}^{(\mathcal{T})}_{\mathcal{D}_{\hat{\boldsymbol{\psi}}^*(\tilde{c}_2)}}(\hat{\boldsymbol{\psi}}^*) - \nu \rho_2'(0^+) \mathrm{sgn}\{\bar{\mathbf{g}}^{(\mathcal{T})}_{\mathcal{D}_{\hat{\boldsymbol{\psi}}^*(\tilde{c}_2)}}(\hat{\boldsymbol{\psi}}^*)\} \end{pmatrix}.
$$

It holds that

$$
\begin{aligned}
|\boldsymbol{\beta}_{\mathcal{M}_{\hat{\boldsymbol{\psi}}^*(\tilde{c}_2)}}(\hat{\boldsymbol{\psi}}^*)|_2 &\leq |\bar{\mathbf{g}}^{(\mathcal{I})}(\hat{\boldsymbol{\psi}}^*)|_2 + \big|\bar{\mathbf{g}}^{(\mathcal{T})}_{\mathcal{D}_{\hat{\boldsymbol{\psi}}^*(\tilde{c}_2)}}(\hat{\boldsymbol{\psi}}^*) - \nu \rho_2'(0^+) \mathrm{sgn}\big\{\bar{\mathbf{g}}^{(\mathcal{T})}_{\mathcal{D}_{\hat{\boldsymbol{\psi}}^*(\tilde{c}_2)}}(\hat{\boldsymbol{\psi}}^*)\big\}\big|_2 \\
&\leq |\bar{\mathbf{g}}^{(\mathcal{I})}(\hat{\boldsymbol{\psi}})|_2 + \underbrace{\big|\bar{\mathbf{g}}^{(\mathcal{T})}_{\mathcal{D}_{\hat{\boldsymbol{\psi}}^*(\tilde{c}_2) \cap \mathcal{D}_{\hat{\boldsymbol{\psi}}}(\tilde{c}_1)}}(\hat{\boldsymbol{\psi}}^*) - \nu \rho_2'(0^+) \mathrm{sgn}\big\{\bar{\mathbf{g}}^{(\mathcal{T})}_{\mathcal{D}_{\hat{\boldsymbol{\psi}}^*(\tilde{c}_2) \cap \mathcal{D}_{\hat{\boldsymbol{\psi}}}(\tilde{c}_1)}}(\hat{\boldsymbol{\psi}}^*)\big\}\big|_2}_{T_1} \qquad \text{(A.17)} \\
&\quad + \underbrace{|\bar{\mathbf{g}}^{(\mathcal{I})}(\hat{\boldsymbol{\psi}}^*) - \bar{\mathbf{g}}^{(\mathcal{I})}(\hat{\boldsymbol{\psi}})|_2}_{T_2} + \underbrace{\big|\bar{\mathbf{g}}^{(\mathcal{T})}_{\mathcal{D}_{\hat{\boldsymbol{\psi}}^*(\tilde{c}_2) \cap \mathcal{D}^{\mathrm{c}}_{\hat{\boldsymbol{\psi}}}(\tilde{c}_1)}}(\hat{\boldsymbol{\psi}}^*) - \nu \rho_2'(0^+) \mathrm{sgn}\big\{\bar{\mathbf{g}}^{(\mathcal{T})}_{\mathcal{D}_{\hat{\boldsymbol{\psi}}^*(\tilde{c}_2) \cap \mathcal{D}^{\mathrm{c}}_{\hat{\boldsymbol{\psi}}}(\tilde{c}_1)}}(\hat{\boldsymbol{\psi}}^*)\big\}\big|_2}_{T_3}\,.
\end{aligned}
$$

As we have shown that $|\boldsymbol{\beta}_{\mathcal{M}_{\hat{\boldsymbol{\psi}}}(\tilde{c}_1)}(\hat{\boldsymbol{\psi}})|_2^2 = |\bar{\mathbf{g}}^{(\mathcal{I})}(\hat{\boldsymbol{\psi}})|_2^2 + |\bar{\mathbf{g}}^{(\mathcal{T})}_{\mathcal{D}_{\hat{\boldsymbol{\psi}}}(\tilde{c}_1)}(\hat{\boldsymbol{\psi}}) - \nu\rho_2'(0^+)\mathrm{sgn}\{\bar{\mathbf{g}}^{(\mathcal{T})}_{\mathcal{D}_{\hat{\boldsymbol{\psi}}}(\tilde{c}_1)}(\hat{\boldsymbol{\psi}})\}|_2^2 = O_{\mathrm{p}}(r_1\aleph_n^2)$, then $|\bar{\mathbf{g}}^{(\mathcal{I})}(\hat{\boldsymbol{\psi}})|_2 = O_{\mathrm{p}}(r_1^{1/2}\aleph_n) = |\bar{\mathbf{g}}^{(\mathcal{T})}_{\mathcal{D}_{\hat{\boldsymbol{\psi}}}(\tilde{c}_1)}(\hat{\boldsymbol{\psi}}) - \nu\rho_2'(0^+)\mathrm{sgn}\{\bar{\mathbf{g}}^{(\mathcal{T})}_{\mathcal{D}_{\hat{\boldsymbol{\psi}}}(\tilde{c}_1)}(\hat{\boldsymbol{\psi}})\}|_2$. For the term T_1 in (A.17), we have

$$
\begin{aligned}
T_1 &\le \big|\bar{\mathbf{g}}^{(\mathcal{T})}_{\mathcal{D}_{\hat{\boldsymbol{\psi}}}(\tilde{c}_1)}(\hat{\boldsymbol{\psi}}^*) - \nu\rho_2'(0^+)\mathrm{sgn}\big\{\bar{\mathbf{g}}^{(\mathcal{T})}_{\mathcal{D}_{\hat{\boldsymbol{\psi}}}(\tilde{c}_1)}(\hat{\boldsymbol{\psi}}^*)\big\}\big|_2 \\
&\le \big|\bar{\mathbf{g}}^{(\mathcal{T})}_{\mathcal{D}_{\hat{\boldsymbol{\psi}}}(\tilde{c}_1)}(\hat{\boldsymbol{\psi}}) - \nu\rho_2'(0^+)\mathrm{sgn}\big\{\bar{\mathbf{g}}^{(\mathcal{T})}_{\mathcal{D}_{\hat{\boldsymbol{\psi}}}(\tilde{c}_1)}(\hat{\boldsymbol{\psi}})\big\}\big|_2 + \big|\bar{\mathbf{g}}^{(\mathcal{T})}_{\mathcal{D}_{\hat{\boldsymbol{\psi}}}(\tilde{c}_1)}(\hat{\boldsymbol{\psi}}^*) - \bar{\mathbf{g}}^{(\mathcal{T})}_{\mathcal{D}_{\hat{\boldsymbol{\psi}}}(\tilde{c}_1)}(\hat{\boldsymbol{\psi}})\big|_2 \\
&\quad + \nu\rho_2'(0^+)\big|\mathrm{sgn}\big\{\bar{\mathbf{g}}^{(\mathcal{T})}_{\mathcal{D}_{\hat{\boldsymbol{\psi}}}(\tilde{c}_1)}(\hat{\boldsymbol{\psi}})\big\} - \mathrm{sgn}\big\{\bar{\mathbf{g}}^{(\mathcal{T})}_{\mathcal{D}_{\hat{\boldsymbol{\psi}}}(\tilde{c}_1)}(\hat{\boldsymbol{\psi}}^*)\big\}\big|_2 \\
&\le O_{\mathrm{p}}(r_1^{1/2}\aleph_n) + \big|\bar{\mathbf{g}}^{(\mathcal{T})}_{\mathcal{D}_{\hat{\boldsymbol{\psi}}}(\tilde{c}_1)}(\hat{\boldsymbol{\psi}}^*) - \bar{\mathbf{g}}^{(\mathcal{T})}_{\mathcal{D}_{\hat{\boldsymbol{\psi}}}(\tilde{c}_1)}(\hat{\boldsymbol{\psi}})\big|_2 + \nu\rho_2'(0^+)\big|\mathrm{sgn}\big\{\bar{\mathbf{g}}^{(\mathcal{T})}_{\mathcal{D}_{\hat{\boldsymbol{\psi}}}(\tilde{c}_1)}(\hat{\boldsymbol{\psi}})\big\} - \mathrm{sgn}\big\{\bar{\mathbf{g}}^{(\mathcal{T})}_{\mathcal{D}_{\hat{\boldsymbol{\psi}}}(\tilde{c}_1)}(\hat{\boldsymbol{\psi}}^*)\big\}\big|_2 .
\end{aligned}
$$

Notice that $\max_{j\in\mathcal{D}}|\bar{g}_j^{(\mathcal{T})}(\hat{\boldsymbol{\psi}}^*) - \bar{g}_j^{(\mathcal{T})}(\hat{\boldsymbol{\psi}})| \le |\hat{\boldsymbol{\psi}}_{\mathcal{S}^{\mathrm{c}}}|_1 \cdot O_{\mathrm{p}}(1)$. Then $|\bar{\mathbf{g}}^{(\mathcal{T})}_{\mathcal{D}_{\hat{\boldsymbol{\psi}}}(\tilde{c}_1)}(\hat{\boldsymbol{\psi}}^*) - \bar{\mathbf{g}}^{(\mathcal{T})}_{\mathcal{D}_{\hat{\boldsymbol{\psi}}}(\tilde{c}_1)}(\hat{\boldsymbol{\psi}})|_2 \le \ell_n^{1/2}|\hat{\boldsymbol{\psi}}_{\mathcal{S}^{\mathrm{c}}}|_1 \cdot O_{\mathrm{p}}(1) = O_{\mathrm{p}}(\ell_n^{1/2}\aleph_n)$. Recall $|\bar{g}_j^{(\mathcal{T})}(\hat{\boldsymbol{\psi}})| \ge \tilde{c}_1\nu\rho_2'(0^+)$ for any $j \in \mathcal{D}_{\hat{\boldsymbol{\psi}}}(\tilde{c}_1)$. Due to $|\hat{\boldsymbol{\psi}}_{\mathcal{S}^{\mathrm{c}}}|_1 \le \aleph_n = o(\nu)$, it holds that $\mathrm{sgn}\{\bar{g}_j^{(\mathcal{T})}(\hat{\boldsymbol{\psi}}^*)\} = \mathrm{sgn}\{\bar{g}_j^{(\mathcal{T})}(\hat{\boldsymbol{\psi}})\}$ for any $j \in \mathcal{D}_{\hat{\boldsymbol{\psi}}}(\tilde{c}_1)$. Hence, $T_1 = O_{\mathrm{p}}(\ell_n^{1/2}\aleph_n)$. Analogously, we also have $T_2 \le r_1^{1/2}|\hat{\boldsymbol{\psi}}_{\mathcal{S}^{\mathrm{c}}}|_1 \cdot O_{\mathrm{p}}(1) = O_{\mathrm{p}}(\ell_n^{1/2}\aleph_n)$. For the term T_3, notice that for any $j \in \mathcal{D}_{\hat{\boldsymbol{\psi}}^*}(\tilde{c}_2) \cap \mathcal{D}^{\mathrm{c}}_{\hat{\boldsymbol{\psi}}}(\tilde{c}_1)$, we have $|\bar{g}_j^{(\mathcal{T})}(\hat{\boldsymbol{\psi}}^*)| \ge \tilde{c}_2\nu\rho_2'(0^+)$ and $|\bar{g}_j^{(\mathcal{T})}(\hat{\boldsymbol{\psi}})| < \tilde{c}_1\nu\rho_2'(0^+)$ for some $\tilde{c}_2 > \tilde{c}_1$. Since $\max_{j\in\mathcal{T}}|\bar{g}_j^{(\mathcal{T})}(\hat{\boldsymbol{\psi}}) - \bar{g}_j^{(\mathcal{T})}(\hat{\boldsymbol{\psi}}^*)| \le |\hat{\boldsymbol{\psi}}_{\mathcal{S}^{\mathrm{c}}}|_1 \cdot O_{\mathrm{p}}(1) = o_{\mathrm{p}}(\nu)$, it holds that $\mathcal{D}_{\hat{\boldsymbol{\psi}}^*}(\tilde{c}_2) \cap \mathcal{D}^{\mathrm{c}}_{\hat{\boldsymbol{\psi}}}(\tilde{c}_1) = \emptyset$ w.p.a.1, which implies that $T_3 = 0$ w.p.a.1. Therefore, we have $|\boldsymbol{\beta}_{\mathcal{M}_{\hat{\boldsymbol{\psi}}^*(\tilde{c}_2)}}(\hat{\boldsymbol{\psi}}^*)|_2 = O_{\mathrm{p}}(\ell_n^{1/2}\aleph_n)$. Together with Lemma A.2, we have $|\hat{\boldsymbol{\lambda}}^*|_2 = O_{\mathrm{p}}(\ell_n^{1/2}\aleph_n)$, which implies $\max_{i\in[n]}|\hat{\boldsymbol{\lambda}}^{*,\mathrm{T}}\mathbf{g}_i^{(\mathcal{T})}(\check{\boldsymbol{\psi}})| = \max_{i\in[n]}|\hat{\boldsymbol{\lambda}}^{*,\mathrm{T}}_{\mathcal{M}_{\hat{\boldsymbol{\psi}}^*(\tilde{c}_2)}}\mathbf{g}^{(\mathcal{T})}_{i,\mathcal{M}_{\hat{\boldsymbol{\psi}}^*(\tilde{c}_2)}}(\check{\boldsymbol{\psi}})| = o_{\mathrm{p}}(1)$ for $\check{\boldsymbol{\psi}}$ specified in (A.16).

Recall $\mathbf{g}_i^{(\mathcal{T})}(\boldsymbol{\psi}) = \{\mathbf{g}_i^{(\mathcal{I})}(\boldsymbol{\theta})^{\mathrm{T}}, \mathbf{g}_i^{(\mathcal{D})}(\boldsymbol{\theta})^{\mathrm{T}} - \boldsymbol{\xi}^{\mathrm{T}}\}^{\mathrm{T}}$ and $\boldsymbol{\psi}_{\mathcal{S}^{\mathrm{c}}} = \boldsymbol{\xi}_{\mathcal{A}}$. We have $\hat{\boldsymbol{\lambda}}^{*,\mathrm{T}}\nabla_{\boldsymbol{\psi}_{\mathcal{S}^{\mathrm{c}}}}\mathbf{g}_i^{(\mathcal{T})}(\check{\boldsymbol{\psi}}) = -\hat{\boldsymbol{\lambda}}^{*,\mathrm{T}}_{\mathcal{A}}$. For I, since $\max_{i\in[n]}|\hat{\boldsymbol{\lambda}}^{*,\mathrm{T}}\mathbf{g}_i^{(\mathcal{T})}(\check{\boldsymbol{\psi}})| = o_{\mathrm{p}}(1)$, we have

$$
|\mathrm{I}| = \left|\frac{1}{n}\sum_{i=1}^{n}\frac{\hat{\boldsymbol{\lambda}}^{*,\mathrm{T}}_{\mathcal{A}}\hat{\boldsymbol{\psi}}_{\mathcal{S}^{\mathrm{c}}}}{1+\hat{\boldsymbol{\lambda}}^{*,\mathrm{T}}\mathbf{g}_i^{(\mathcal{T})}(\check{\boldsymbol{\psi}})}\right| \le |\hat{\boldsymbol{\lambda}}^*|_\infty|\hat{\boldsymbol{\psi}}_{\mathcal{S}^{\mathrm{c}}}|_1\{1+o_{\mathrm{p}}(1)\} \le |\hat{\boldsymbol{\lambda}}^*|_2|\hat{\boldsymbol{\psi}}_{\mathcal{S}^{\mathrm{c}}}|_1\{1+o_{\mathrm{p}}(1)\}. \tag{A.18}
$$

As we have shown $|\hat{\boldsymbol{\lambda}}^*|_2 = O_{\mathrm{p}}(\ell_n^{1/2}\aleph_n)$, it then holds $|\mathrm{I}| \le |\hat{\boldsymbol{\psi}}_{\mathcal{S}^{\mathrm{c}}}|_1 \cdot O_{\mathrm{p}}(\ell_n^{1/2}\aleph_n)$. On the other hand, II in (A.16) satisfies $\mathrm{II} = \sum_{k\in\mathcal{A}}\pi\rho_1'(c_k|\hat{\xi}_k|;\pi)|\hat{\xi}_k| \ge C\pi|\hat{\boldsymbol{\psi}}_{\mathcal{S}^{\mathrm{c}}}|_1$ for some $c_k \in (0,1)$. Due to $\ell_n^{1/2}\aleph_n = o(\pi)$, we can obtain $\mathrm{I}+\mathrm{II} > 0$ w.p.a.1, which implies that $S_n(\hat{\boldsymbol{\psi}}^*) < S_n(\hat{\boldsymbol{\psi}})$ w.p.a.1. We complete the proof of Part (iii). □

A.4 Proof of Theorem 3.1

Select $\hat{\boldsymbol{\psi}}_{\mathrm{PEL}}$ as the sparse local minimizer given in Proposition 3.1. Recall that the estimate $\hat{\boldsymbol{\lambda}}(\boldsymbol{\psi}) = \arg\max_{\boldsymbol{\lambda}\in\hat{\Lambda}_n^{(\mathcal{T})}(\boldsymbol{\psi})} f(\boldsymbol{\lambda};\boldsymbol{\psi})$ is the Lagrange multiplier associated with $\boldsymbol{\psi}$, where $f(\boldsymbol{\lambda};\boldsymbol{\psi}) = n^{-1}\sum_{i=1}^n \log\{1+\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{g}_i^{(\mathcal{T})}(\boldsymbol{\psi})\} - \sum_{j\in\mathcal{D}}P_{2,\nu}(|\lambda_j|)$ for any $\boldsymbol{\psi} \in \boldsymbol{\Psi}$ and $\boldsymbol{\lambda} = (\lambda_1,\ldots,\lambda_r)^{\mathrm{T}}$. Write

$\hat{\boldsymbol{\lambda}} = \hat{\boldsymbol{\lambda}}(\hat{\boldsymbol{\psi}}_{\mathrm{PEL}}) = (\hat{\lambda}_1, \ldots, \hat{\lambda}_r)^{\mathrm{T}}$. Recall $\mathcal{R}_n = \mathcal{I} \cup \mathrm{supp}(\hat{\boldsymbol{\lambda}}_{\mathcal{D}})$, $\mathcal{A}_* = \{j \in \mathcal{A} : \hat{\lambda}_j \neq 0\}$ and $\mathcal{A}_{*,\mathrm{c}} = \{j \in \mathcal{A}^{\mathrm{c}} : \hat{\lambda}_j \neq 0\}$. Then $\mathcal{R}_n$ can be decomposed into three disjoint sets $\mathcal{R}_n = \mathcal{I} \cup \mathcal{A}_* \cup \mathcal{A}_{*,\mathrm{c}}$. Write $\mathcal{I}^* = \mathcal{I} \cup \mathcal{A}_*$. Notice that $\mathcal{S}_* = \mathcal{P} \cup \mathcal{A}_{*,\mathrm{c}}$ and $\mathcal{S} = \mathcal{P} \cup \mathcal{A}^{\mathrm{c}}$. Then $\mathcal{S}_* \subset \mathcal{S}$ and $s_* := |\mathcal{S}_*| \leq |\mathcal{S}| = s$. For any $\boldsymbol{\psi} \in \boldsymbol{\Psi}$, we have $\boldsymbol{\psi}_{\mathcal{S}_*} = (\boldsymbol{\theta}^{\mathrm{T}}, \boldsymbol{\xi}_{\mathcal{A}_{*,\mathrm{c}}}^{\mathrm{T}})^{\mathrm{T}}$. To prove Theorem 3.1, we also need the following two lemmas. The proof of Lemma A.4 is similar to that of Lemma 3 in Chang et al. (2018) and we omit it here. The proof of Lemma A.5 is given in Section A.7.3.

Lemma A.4. *Assume the conditions of Proposition* 3.1 *hold. Then* $\sup_{\mathcal{F} \in \mathscr{F}} \big|[\nabla_{\boldsymbol{\psi}_{\mathcal{S}_*}} \bar{\mathbf{g}}_{\mathcal{F}}^{(\mathcal{T})}(\hat{\boldsymbol{\psi}}_{\mathrm{PEL}}) - \mathbb{E}\{\nabla_{\boldsymbol{\psi}_{\mathcal{S}_*}} \mathbf{g}_{i,\mathcal{F}}^{(\mathcal{T})}(\boldsymbol{\psi}_0)\}]\mathbf{z}\big|_2 = |\mathbf{z}|_2 \cdot \{O_{\mathrm{p}}(s^{3/2}\ell_n^{1/2}\phi_n) + O_{\mathrm{p}}(s^{1/2}\ell_n^{1/2}\aleph_n)\}$ *holds uniformly over* $\mathbf{z} \in \mathbb{R}^{s_*}$, *where* $\mathscr{F}$ *is defined in Lemma* A.1.

Lemma A.5. *Assume Condition* 6 *and the conditions of Proposition* 3.1 *hold. It then holds w.p.a.1 that* $\hat{\boldsymbol{\lambda}}(\boldsymbol{\psi})$ *is continuously differentiable at* $\hat{\boldsymbol{\psi}}_{\mathrm{PEL}}$ *and* $\nabla_{\boldsymbol{\psi}} \hat{\boldsymbol{\lambda}}_{\mathcal{R}_n^{\mathrm{c}}}(\hat{\boldsymbol{\psi}}_{\mathrm{PEL}}) = \mathbf{0}$.

Now we begin to prove Theorem 3.1. Define

$$H_n(\boldsymbol{\psi}, \boldsymbol{\lambda}) = \frac{1}{n} \sum_{i=1}^{n} \log\{1 + \boldsymbol{\lambda}^{\mathrm{T}} \mathbf{g}_i^{(\mathcal{T})}(\boldsymbol{\psi})\} + \sum_{k \in \mathcal{D}} P_{1,\pi}(|\xi_k|) - \sum_{j \in \mathcal{D}} P_{2,\nu}(|\lambda_j|)\,. \tag{A.19}$$

By the definition of $\hat{\boldsymbol{\lambda}}$, we have $\nabla_{\boldsymbol{\lambda}} H_n(\hat{\boldsymbol{\psi}}_{\mathrm{PEL}}, \hat{\boldsymbol{\lambda}}) = \mathbf{0}$, that is,

$$\mathbf{0} = \frac{1}{n} \sum_{i=1}^{n} \frac{\mathbf{g}_i^{(\mathcal{T})}(\hat{\boldsymbol{\psi}}_{\mathrm{PEL}})}{1 + \hat{\boldsymbol{\lambda}}^{\mathrm{T}} \mathbf{g}_i^{(\mathcal{T})}(\hat{\boldsymbol{\psi}}_{\mathrm{PEL}})} - \hat{\boldsymbol{\eta}}\,,$$

where $\hat{\boldsymbol{\eta}} = (\hat{\eta}_1, \ldots, \hat{\eta}_r)^{\mathrm{T}}$ with $\hat{\eta}_j = 0$ for $j \in \mathcal{I}$, $\hat{\eta}_j = \nu\rho_2'(|\hat{\lambda}_j|; \nu)\mathrm{sgn}(\hat{\lambda}_j)$ for $j \in \mathcal{D}$ and $\hat{\lambda}_j \neq 0$, and $\hat{\eta}_j \in [-\nu\rho_2'(0^+), \nu\rho_2'(0^+)]$ for $j \in \mathcal{R}_n^{\mathrm{c}}$. It follows from the Taylor expansion that

$$\begin{aligned} \mathbf{0} &= \frac{1}{n} \sum_{i=1}^{n} \mathbf{g}_{i,\mathcal{R}_n}^{(\mathcal{T})}(\hat{\boldsymbol{\psi}}_{\mathrm{PEL}}) - \frac{1}{n} \sum_{i=1}^{n} \frac{\mathbf{g}_{i,\mathcal{R}_n}^{(\mathcal{T})}(\hat{\boldsymbol{\psi}}_{\mathrm{PEL}})^{\otimes 2} \hat{\boldsymbol{\lambda}}_{\mathcal{R}_n}}{\{1 + c\hat{\boldsymbol{\lambda}}_{\mathcal{R}_n}^{\mathrm{T}} \mathbf{g}_{i,\mathcal{R}_n}^{(\mathcal{T})}(\hat{\boldsymbol{\psi}}_{\mathrm{PEL}})\}^2} - \hat{\boldsymbol{\eta}}_{\mathcal{R}_n} \\ &=: \bar{\mathbf{g}}_{\mathcal{R}_n}^{(\mathcal{T})}(\hat{\boldsymbol{\psi}}_{\mathrm{PEL}}) - \mathbf{C}(\hat{\boldsymbol{\psi}}_{\mathrm{PEL}}) \hat{\boldsymbol{\lambda}}_{\mathcal{R}_n} - \hat{\boldsymbol{\eta}}_{\mathcal{R}_n} \end{aligned} \tag{A.20}$$

for some $|c| < 1$. Hence, $\hat{\boldsymbol{\lambda}}_{\mathcal{R}_n} = \{\mathbf{C}(\hat{\boldsymbol{\psi}}_{\mathrm{PEL}})\}^{-1}\{\bar{\mathbf{g}}_{\mathcal{R}_n}^{(\mathcal{T})}(\hat{\boldsymbol{\psi}}_{\mathrm{PEL}}) - \hat{\boldsymbol{\eta}}_{\mathcal{R}_n}\}$. By the definition of $\hat{\boldsymbol{\psi}}_{\mathrm{PEL}}$, we have $\mathbf{0} = \nabla_{\boldsymbol{\psi}} H_n\{\boldsymbol{\psi}, \hat{\boldsymbol{\lambda}}(\boldsymbol{\psi})\}|_{\boldsymbol{\psi} = \hat{\boldsymbol{\psi}}_{\mathrm{PEL}}}$. Notice that

$$\begin{aligned} &\nabla_{\boldsymbol{\psi}} H_n\{\boldsymbol{\psi}, \hat{\boldsymbol{\lambda}}(\boldsymbol{\psi})\}|_{\boldsymbol{\psi} = \hat{\boldsymbol{\psi}}_{\mathrm{PEL}}} \\ &= \frac{\partial H_n(\hat{\boldsymbol{\psi}}_{\mathrm{PEL}}, \hat{\boldsymbol{\lambda}})}{\partial \boldsymbol{\psi}} + \Bigg\{ \underbrace{\frac{\partial H_n(\hat{\boldsymbol{\psi}}_{\mathrm{PEL}}, \hat{\boldsymbol{\lambda}})}{\partial \boldsymbol{\lambda}_{\mathcal{R}_n}^{\mathrm{T}}}}_{\mathrm{I}} \frac{\partial \hat{\boldsymbol{\lambda}}_{\mathcal{R}_n}(\hat{\boldsymbol{\psi}}_{\mathrm{PEL}})}{\partial \boldsymbol{\psi}} + \frac{\partial H_n(\hat{\boldsymbol{\psi}}_{\mathrm{PEL}}, \hat{\boldsymbol{\lambda}})}{\partial \boldsymbol{\lambda}_{\mathcal{R}_n^{\mathrm{c}}}^{\mathrm{T}}} \underbrace{\frac{\partial \hat{\boldsymbol{\lambda}}_{\mathcal{R}_n^{\mathrm{c}}}(\hat{\boldsymbol{\psi}}_{\mathrm{PEL}})}{\partial \boldsymbol{\psi}}}_{\mathrm{II}} \Bigg\}^{\mathrm{T}}. \end{aligned}$$

Due to $\hat{\boldsymbol{\lambda}}(\hat{\boldsymbol{\psi}}_{\text{PEL}}) = \arg\max_{\boldsymbol{\lambda}\in\hat{\Lambda}_n^{(\mathcal{T})}(\hat{\boldsymbol{\psi}}_{\text{PEL}})} f(\boldsymbol{\lambda};\hat{\boldsymbol{\psi}}_{\text{PEL}}) = \arg\max_{\boldsymbol{\lambda}\in\hat{\Lambda}_n^{(\mathcal{T})}(\hat{\boldsymbol{\psi}}_{\text{PEL}})} H_n(\hat{\boldsymbol{\psi}}_{\text{PEL}},\boldsymbol{\lambda})$, then $\mathrm{I}=\mathbf{0}$. On the other hand, Lemma A.5 implies that $\mathrm{II}=\mathbf{0}$. Thus, $\mathbf{0}=\partial H_n(\hat{\boldsymbol{\psi}}_{\text{PEL}},\hat{\boldsymbol{\lambda}})/\partial\boldsymbol{\psi}$. Together with (A.20), we have

$$\begin{aligned}
\mathbf{0} &= \left\{\frac{1}{n}\sum_{i=1}^{n}\frac{\nabla_{\boldsymbol{\psi}_{\mathcal{S}_*}}\mathbf{g}_{i,\mathcal{R}_n}^{(\mathcal{T})}(\hat{\boldsymbol{\psi}}_{\text{PEL}})}{1+\hat{\boldsymbol{\lambda}}_{\mathcal{R}_n}^{\mathrm{T}}\mathbf{g}_{i,\mathcal{R}_n}^{(\mathcal{T})}(\hat{\boldsymbol{\psi}}_{\text{PEL}})}\right\}^{\mathrm{T}}\hat{\boldsymbol{\lambda}}_{\mathcal{R}_n}+\hat{\boldsymbol{\varsigma}}_{\mathcal{S}_*} \\
&= \left\{\frac{1}{n}\sum_{i=1}^{n}\frac{\nabla_{\boldsymbol{\psi}_{\mathcal{S}_*}}\mathbf{g}_{i,\mathcal{R}_n}^{(\mathcal{T})}(\hat{\boldsymbol{\psi}}_{\text{PEL}})}{1+\hat{\boldsymbol{\lambda}}_{\mathcal{R}_n}^{\mathrm{T}}\mathbf{g}_{i,\mathcal{R}_n}^{(\mathcal{T})}(\hat{\boldsymbol{\psi}}_{\text{PEL}})}\right\}^{\mathrm{T}}\{\mathbf{C}(\hat{\boldsymbol{\psi}}_{\text{PEL}})\}^{-1}\{\bar{\mathbf{g}}_{\mathcal{R}_n}^{(\mathcal{T})}(\hat{\boldsymbol{\psi}}_{\text{PEL}})-\hat{\boldsymbol{\eta}}_{\mathcal{R}_n}\}+\hat{\boldsymbol{\varsigma}}_{\mathcal{S}_*} \qquad \text{(A.21)} \\
&=: \{\mathbf{D}(\hat{\boldsymbol{\psi}}_{\text{PEL}})\}^{\mathrm{T}}\{\mathbf{C}(\hat{\boldsymbol{\psi}}_{\text{PEL}})\}^{-1}\{\bar{\mathbf{g}}_{\mathcal{R}_n}^{(\mathcal{T})}(\hat{\boldsymbol{\psi}}_{\text{PEL}})-\hat{\boldsymbol{\eta}}_{\mathcal{R}_n}\}+\hat{\boldsymbol{\varsigma}}_{\mathcal{S}_*}\,,
\end{aligned}$$

where $\hat{\boldsymbol{\varsigma}}_{\mathcal{S}_*} = \{\sum_{k\in\mathcal{D}}\nabla_{\boldsymbol{\psi}_{\mathcal{S}_*}}P_{1,\pi}(|\xi_k|)\}|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}_{\text{PEL}}}$. Recall that $\mathcal{S}_* = \mathcal{P}\cup\mathcal{A}_{*,\mathrm{c}}$. Proposition 3.1 and (3.5) imply that $\hat{\boldsymbol{\varsigma}}_{\mathcal{S}_*}=\mathbf{0}$ w.p.a.1. To construct the asymptotic normality, we need the following lemma. The proof of Lemma A.6 is similar to that of Lemma 2 in Chang et al. (2018) and we omit it here.

Lemma A.6. *Assume the conditions of Proposition* 3.1 *hold. Then* $\|\mathbf{C}(\hat{\boldsymbol{\psi}}_{\text{PEL}})-\widehat{\mathbf{V}}_{\mathcal{R}_n}^{(\mathcal{T})}(\hat{\boldsymbol{\psi}}_{\text{PEL}})\|_2 = O_{\mathrm{p}}(\ell_n n^{1/\gamma}\aleph_n)$*, and* $|\{\mathbf{D}(\hat{\boldsymbol{\psi}}_{\text{PEL}})-\nabla_{\boldsymbol{\psi}_{\mathcal{S}_*}}\bar{\mathbf{g}}_{\mathcal{R}_n}^{(\mathcal{T})}(\hat{\boldsymbol{\psi}}_{\text{PEL}})\}\mathbf{z}|_2 = |\mathbf{z}|_2\cdot O_{\mathrm{p}}(\ell_n s^{1/2}\aleph_n)$ *holds uniformly over* $\mathbf{z}\in\mathbb{R}^{s_*}$.

Recall

$$\mathbf{J}_{\mathcal{R}_n}^{(\mathcal{T})} = ([\mathbb{E}\{\nabla_{\boldsymbol{\psi}_{\mathcal{S}_*}}\mathbf{g}_{i,\mathcal{R}_n}^{(\mathcal{T})}(\boldsymbol{\psi}_0)\}]^{\mathrm{T}}\{\mathbf{V}_{\mathcal{R}_n}^{(\mathcal{T})}(\boldsymbol{\psi}_0)\}^{-1/2})^{\otimes 2}\,. \qquad \text{(A.22)}$$

For any $\boldsymbol{\alpha}\in\mathbb{R}^{s_*}$ with unit L_2-norm, let $\boldsymbol{\delta}=\{\mathbf{J}_{\mathcal{R}_n}^{(\mathcal{T})}\}^{-1/2}\boldsymbol{\alpha}$. Following the same arguments stated in the proof of Proposition A.1, we have $|\boldsymbol{\delta}|_2=O(1)$. Lemma A.2 indicates that $\mathcal{R}_n\subset\mathcal{M}_{\hat{\boldsymbol{\psi}}_{\text{PEL}}}(\tilde{c}) = \mathcal{I}\cup\mathcal{D}_{\hat{\boldsymbol{\psi}}_{\text{PEL}}}(\tilde{c})$ w.p.a.1 for some $\tilde{c}\in(C_*,1)$. As we have shown in the proof of Proposition 3.1, it holds that

$$|\bar{\mathbf{g}}_{\mathcal{R}_n}^{(\mathcal{T})}(\hat{\boldsymbol{\psi}}_{\text{PEL}})-\hat{\boldsymbol{\eta}}_{\mathcal{R}_n}|_2 \le \left|\begin{array}{c}\bar{\mathbf{g}}^{(\mathcal{I})}(\hat{\boldsymbol{\psi}}_{\text{PEL}}) \\ \bar{\mathbf{g}}_{\mathcal{D}_{\hat{\boldsymbol{\psi}}_{\text{PEL}}}(\tilde{c})}^{(\mathcal{T})}(\hat{\boldsymbol{\psi}}_{\text{PEL}})-\nu\rho_2'(0^+)\mathrm{sgn}\{\bar{\mathbf{g}}_{\mathcal{D}_{\hat{\boldsymbol{\psi}}_{\text{PEL}}}(\tilde{c})}^{(\mathcal{T})}(\hat{\boldsymbol{\psi}}_{\text{PEL}})\}\end{array}\right|_2 = O_{\mathrm{p}}(\ell_n^{1/2}\aleph_n)\,. \qquad \text{(A.23)}$$

Together with Lemmas A.1, A.4, and A.6, (A.21) implies that

$$\begin{aligned}
&\boldsymbol{\delta}^{\mathrm{T}}[\mathbb{E}\{\nabla_{\boldsymbol{\psi}_{\mathcal{S}_*}}\mathbf{g}_{i,\mathcal{R}_n}^{(\mathcal{T})}(\boldsymbol{\psi}_0)\}]^{\mathrm{T}}\{\mathbf{V}_{\mathcal{R}_n}^{(\mathcal{T})}(\boldsymbol{\psi}_0)\}^{-1}\{\bar{\mathbf{g}}_{\mathcal{R}_n}^{(\mathcal{T})}(\hat{\boldsymbol{\psi}}_{\text{PEL}})-\hat{\boldsymbol{\eta}}_{\mathcal{R}_n}\} \\
&\qquad = O_{\mathrm{p}}(\ell_n^{3/2}s^{1/2}\aleph_n^2)+O_{\mathrm{p}}(\ell_n s^{3/2}\phi_n\aleph_n)+O_{\mathrm{p}}(\ell_n^{3/2}n^{1/\gamma}\aleph_n^2)\,.
\end{aligned}$$

Notice that $\mathbb{P}(\hat{\boldsymbol{\psi}}_{\text{PEL},\mathcal{S}^{\mathrm{c}}}=\mathbf{0})\to 1$ and $\boldsymbol{\psi}_{0,\mathcal{S}^{\mathrm{c}}}=\mathbf{0}$. By the Taylor expansion, we have $\bar{\mathbf{g}}_{\mathcal{R}_n}^{(\mathcal{T})}(\hat{\boldsymbol{\psi}}_{\text{PEL}}) = \bar{\mathbf{g}}_{\mathcal{R}_n}^{(\mathcal{T})}(\boldsymbol{\psi}_0)+\nabla_{\boldsymbol{\psi}_{\mathcal{S}}}\bar{\mathbf{g}}_{\mathcal{R}_n}^{(\mathcal{T})}(\tilde{\boldsymbol{\psi}})(\hat{\boldsymbol{\psi}}_{\text{PEL},\mathcal{S}}-\boldsymbol{\psi}_{0,\mathcal{S}})$ w.p.a.1, where $\tilde{\boldsymbol{\psi}}$ is on the line joining $\boldsymbol{\psi}_0$ and $\hat{\boldsymbol{\psi}}_{\text{PEL}}$. Recall

that $\mathcal{R}_n = \mathcal{I} \cup \mathcal{A}_* \cup \mathcal{A}_{*,\mathrm{c}}$ with $\mathcal{A}_* = \{j \in \mathcal{A} : \hat{\lambda}_j \neq 0\}$ and $\mathcal{A}_{*,\mathrm{c}} = \{j \in \mathcal{A}^{\mathrm{c}} : \hat{\lambda}_j \neq 0\}$. Notice that $\mathcal{S}_* = \mathcal{P} \cup \mathcal{A}_{*,\mathrm{c}} \subset \mathcal{S}$, $\boldsymbol{\psi}_{\mathcal{S}} = (\boldsymbol{\theta}^{\mathrm{T}}, \boldsymbol{\xi}_{\mathcal{A}^{\mathrm{c}}}^{\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{\psi}_{\mathcal{S}_*} = (\boldsymbol{\theta}^{\mathrm{T}}, \boldsymbol{\xi}_{\mathcal{A}_{*,\mathrm{c}}}^{\mathrm{T}})^{\mathrm{T}}$. For any $j \in \mathcal{R}_n$ and $k \in \mathcal{S} \backslash \mathcal{S}_*$, we know that $g_{i,j}^{(\mathcal{T})}(\boldsymbol{\psi})$ does not involve ψ_k, which implies that $\partial \bar{g}_j^{(\mathcal{T})}(\tilde{\boldsymbol{\psi}})/\partial \psi_k = 0$. Therefore, it holds that $\bar{\mathbf{g}}_{\mathcal{R}_n}^{(\mathcal{T})}(\hat{\boldsymbol{\psi}}_{\mathrm{PEL}}) = \bar{\mathbf{g}}_{\mathcal{R}_n}^{(\mathcal{T})}(\boldsymbol{\psi}_0) + \nabla_{\boldsymbol{\psi}_{\mathcal{S}_*}} \bar{\mathbf{g}}_{\mathcal{R}_n}^{(\mathcal{T})}(\tilde{\boldsymbol{\psi}})(\hat{\boldsymbol{\psi}}_{\mathrm{PEL},\mathcal{S}_*} - \boldsymbol{\psi}_{0,\mathcal{S}_*})$ w.p.a.1, which leads to

$$
\begin{aligned}
&\boldsymbol{\delta}^{\mathrm{T}}[\mathbb{E}\{\nabla_{\boldsymbol{\psi}_{\mathcal{S}_*}} \mathbf{g}_{i,\mathcal{R}_n}^{(\mathcal{T})}(\boldsymbol{\psi}_0)\}]^{\mathrm{T}}\{\mathbf{V}_{\mathcal{R}_n}^{(\mathcal{T})}(\boldsymbol{\psi}_0)\}^{-1}\{\nabla_{\boldsymbol{\psi}_{\mathcal{S}_*}} \bar{\mathbf{g}}_{\mathcal{R}_n}^{(\mathcal{T})}(\tilde{\boldsymbol{\psi}})(\hat{\boldsymbol{\psi}}_{\mathrm{PEL},\mathcal{S}_*} - \boldsymbol{\psi}_{0,\mathcal{S}_*}) - \hat{\boldsymbol{\eta}}_{\mathcal{R}_n}\} \\
=& -\boldsymbol{\delta}^{\mathrm{T}}[\mathbb{E}\{\nabla_{\boldsymbol{\psi}_{\mathcal{S}_*}} \mathbf{g}_{i,\mathcal{R}_n}^{(\mathcal{T})}(\boldsymbol{\psi}_0)\}]^{\mathrm{T}}\{\mathbf{V}_{\mathcal{R}_n}^{(\mathcal{T})}(\boldsymbol{\psi}_0)\}^{-1}\bar{\mathbf{g}}_{\mathcal{R}_n}^{(\mathcal{T})}(\boldsymbol{\psi}_0) + O_{\mathrm{p}}(\ell_n^{3/2} s^{1/2} \aleph_n^2) \\
&+ O_{\mathrm{p}}(\ell_n s^{3/2} \phi_n \aleph_n) + O_{\mathrm{p}}(\ell_n^{3/2} n^{1/\gamma} \aleph_n^2)\,.
\end{aligned} \tag{A.24}
$$

Next, we will specify the convergence rate of $|[\mathbb{E}\{\nabla_{\boldsymbol{\psi}_{\mathcal{S}_*}} \mathbf{g}_{i,\mathcal{R}_n}^{(\mathcal{T})}(\boldsymbol{\psi}_0)\} - \nabla_{\boldsymbol{\psi}_{\mathcal{S}_*}} \bar{\mathbf{g}}_{\mathcal{R}_n}^{(\mathcal{T})}(\tilde{\boldsymbol{\psi}})](\hat{\boldsymbol{\psi}}_{\mathrm{PEL},\mathcal{S}_*} - \boldsymbol{\psi}_{0,\mathcal{S}_*})|_2$. Since $\ell_n \aleph_n = o(\nu)$, it follows from (A.23) that $|\bar{\mathbf{g}}_{\mathcal{R}_n}^{(\mathcal{T})}(\hat{\boldsymbol{\psi}}_{\mathrm{PEL}}) - \bar{\mathbf{g}}_{\mathcal{R}_n}^{(\mathcal{T})}(\boldsymbol{\psi}_0)|_2 \leq |\bar{\mathbf{g}}_{\mathcal{R}_n}^{(\mathcal{T})}(\hat{\boldsymbol{\psi}}_{\mathrm{PEL}})|_2 + |\bar{\mathbf{g}}_{\mathcal{R}_n}^{(\mathcal{T})}(\boldsymbol{\psi}_0)|_2 = O_{\mathrm{p}}(\ell_n^{1/2}\nu) + O_{\mathrm{p}}(\ell_n^{1/2}\aleph_n) = O_{\mathrm{p}}(\ell_n^{1/2}\nu)$. On the other hand, it holds that $|\bar{\mathbf{g}}_{\mathcal{R}_n}^{(\mathcal{T})}(\hat{\boldsymbol{\psi}}_{\mathrm{PEL}}) - \bar{\mathbf{g}}_{\mathcal{R}_n}^{(\mathcal{T})}(\boldsymbol{\psi}_0)|_2 \geq \lambda_{\min}^{1/2}([\{\nabla_{\boldsymbol{\psi}_{\mathcal{S}_*}} \bar{\mathbf{g}}_{\mathcal{R}_n}^{(\mathcal{T})}(\bar{\boldsymbol{\psi}})\}^{\mathrm{T}}]^{\otimes 2})|\hat{\boldsymbol{\psi}}_{\mathrm{PEL},\mathcal{S}_*} - \boldsymbol{\psi}_{0,\mathcal{S}_*}|_2$, where $\bar{\boldsymbol{\psi}}$ is on the line joining $\boldsymbol{\psi}_0$ and $\hat{\boldsymbol{\psi}}_{\mathrm{PEL}}$. Similar to Lemma A.4, Condition 5 implies that $|\hat{\boldsymbol{\psi}}_{\mathrm{PEL},\mathcal{S}_*} - \boldsymbol{\psi}_{0,\mathcal{S}_*}|_2 = O_{\mathrm{p}}(\ell_n^{1/2}\nu)$. Hence, by similar arguments of Lemma A.4, we have $|[\mathbb{E}\{\nabla_{\boldsymbol{\psi}_{\mathcal{S}_*}} \mathbf{g}_{i,\mathcal{R}_n}^{(\mathcal{T})}(\boldsymbol{\psi}_0)\} - \nabla_{\boldsymbol{\psi}_{\mathcal{S}_*}} \bar{\mathbf{g}}_{\mathcal{R}_n}^{(\mathcal{T})}(\tilde{\boldsymbol{\psi}})](\hat{\boldsymbol{\psi}}_{\mathrm{PEL},\mathcal{S}_*} - \boldsymbol{\psi}_{0,\mathcal{S}_*})|_2 = O_{\mathrm{p}}(\ell_n s^{3/2}\phi_n\nu) + O_{\mathrm{p}}(\ell_n s^{1/2}\nu\aleph_n)$. Recall $\hat{\boldsymbol{\zeta}}_{\mathcal{R}_n} = \{\mathbf{J}_{\mathcal{R}_n}^{(\mathcal{T})}\}^{-1}[\mathbb{E}\{\nabla_{\boldsymbol{\psi}_{\mathcal{S}_*}} \mathbf{g}_{i,\mathcal{R}_n}^{(\mathcal{T})}(\boldsymbol{\psi}_0)\}]^{\mathrm{T}}\{\mathbf{V}_{\mathcal{R}_n}^{(\mathcal{T})}(\boldsymbol{\psi}_0)\}^{-1}\hat{\boldsymbol{\eta}}_{\mathcal{R}_n}$ with $\mathbf{J}_{\mathcal{R}_n}^{(\mathcal{T})}$ defined in (A.22). Recall $\boldsymbol{\delta} = \{\mathbf{J}_{\mathcal{R}_n}^{(\mathcal{T})}\}^{-1/2}\boldsymbol{\alpha}$. It follows from (A.24) that

$$
\begin{aligned}
\boldsymbol{\delta}^{\mathrm{T}}\mathbf{J}_{\mathcal{R}_n}^{(\mathcal{T})}(\hat{\boldsymbol{\psi}}_{\mathrm{PEL},\mathcal{S}_*} - \boldsymbol{\psi}_{0,\mathcal{S}_*} - \hat{\boldsymbol{\zeta}}_{\mathcal{R}_n}) =& -\boldsymbol{\alpha}^{\mathrm{T}}\{\mathbf{J}_{\mathcal{R}_n}^{(\mathcal{T})}\}^{-1/2}[\mathbb{E}\{\nabla_{\boldsymbol{\psi}_{\mathcal{S}_*}} \mathbf{g}_{i,\mathcal{R}_n}^{(\mathcal{T})}(\boldsymbol{\psi}_0)\}]^{\mathrm{T}}\{\mathbf{V}_{\mathcal{R}_n}^{(\mathcal{T})}(\boldsymbol{\psi}_0)\}^{-1}\bar{\mathbf{g}}_{\mathcal{R}_n}^{(\mathcal{T})}(\boldsymbol{\psi}_0) \\
&+ O_{\mathrm{p}}(\ell_n^{3/2} n^{1/\gamma}\aleph_n^2) + O_{\mathrm{p}}(\ell_n s^{3/2}\phi_n\nu) + O_{\mathrm{p}}(\ell_n s^{1/2}\nu\aleph_n)\,.
\end{aligned}
$$

Lemma 4 of Chang et al. (2018) yields $n^{1/2}\boldsymbol{\alpha}^{\mathrm{T}}\{\mathbf{J}_{\mathcal{R}_n}^{(\mathcal{T})}\}^{-1/2}[\mathbb{E}\{\nabla_{\boldsymbol{\psi}_{\mathcal{S}_*}} \mathbf{g}_{i,\mathcal{R}_n}^{(\mathcal{T})}(\boldsymbol{\psi}_0)\}]^{\mathrm{T}}\{\mathbf{V}_{\mathcal{R}_n}^{(\mathcal{T})}(\boldsymbol{\psi}_0)\}^{-1}\bar{\mathbf{g}}_{\mathcal{R}_n}^{(\mathcal{T})}(\boldsymbol{\psi}_0) \xrightarrow{d} \mathcal{N}(0,1)$. Then $n^{1/2}\boldsymbol{\alpha}^{\mathrm{T}}\{\mathbf{J}_{\mathcal{R}_n}^{(\mathcal{T})}\}^{1/2}(\hat{\boldsymbol{\psi}}_{\mathrm{PEL},\mathcal{S}_*} - \boldsymbol{\psi}_{0,\mathcal{S}_*} - \hat{\boldsymbol{\zeta}}_{\mathcal{R}_n}) \xrightarrow{d} \mathcal{N}(0,1)$ as $n \to \infty$.

Notice that $\hat{\boldsymbol{\psi}}_{\mathrm{PEL},\mathcal{S}_*} = (\hat{\boldsymbol{\theta}}_{\mathrm{PEL}}^{\mathrm{T}}, \hat{\boldsymbol{\xi}}_{\mathrm{PEL},\mathcal{A}_{*,\mathrm{c}}}^{\mathrm{T}})^{\mathrm{T}}$. In the sequel, we will specify the limiting distribution of $\hat{\boldsymbol{\theta}}_{\mathrm{PEL}}$. Recall $\mathcal{R}_n = \mathcal{I} \cup \mathcal{A}_* \cup \mathcal{A}_{*,\mathrm{c}}$ and $\mathcal{I}^* = \mathcal{I} \cup \mathcal{A}_*$. We write $\{\mathbf{J}_{\mathcal{R}_n}^{(\mathcal{T})}\}^{-1}$, $\mathbb{E}\{\nabla_{\boldsymbol{\psi}_{\mathcal{S}_*}} \mathbf{g}_{i,\mathcal{R}_n}^{(\mathcal{T})}(\boldsymbol{\psi}_0)\}$ and $\{\mathbf{V}_{\mathcal{R}_n}^{(\mathcal{T})}(\boldsymbol{\psi}_0)\}^{-1}$ with following blocks

$$
\{\mathbf{J}_{\mathcal{R}_n}^{(\mathcal{T})}\}^{-1} = \begin{pmatrix} [\{\mathbf{J}_{\mathcal{R}_n}^{(\mathcal{T})}\}^{-1}]_{11} & [\{\mathbf{J}_{\mathcal{R}_n}^{(\mathcal{T})}\}^{-1}]_{12} \\ [\{\mathbf{J}_{\mathcal{R}_n}^{(\mathcal{T})}\}^{-1}]_{21} & [\{\mathbf{J}_{\mathcal{R}_n}^{(\mathcal{T})}\}^{-1}]_{22} \end{pmatrix}, \quad \{\mathbf{V}_{\mathcal{R}_n}^{(\mathcal{T})}(\boldsymbol{\psi}_0)\}^{-1} = \begin{pmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{pmatrix},
$$

$$
\mathbb{E}\{\nabla_{\boldsymbol{\psi}_{\mathcal{S}_*}} \mathbf{g}_{i,\mathcal{R}_n}^{(\mathcal{T})}(\boldsymbol{\psi}_0)\} = \begin{pmatrix} \mathbb{E}\{\nabla_{\boldsymbol{\theta}} \mathbf{g}_{i,\mathcal{I}^*}^{(\mathcal{T})}(\boldsymbol{\theta}_0)\} & \mathbf{0} \\ \mathbb{E}\{\nabla_{\boldsymbol{\theta}} \mathbf{g}_{i,\mathcal{A}_{*,\mathrm{c}}}^{(\mathcal{T})}(\boldsymbol{\theta}_0)\} & -\mathbf{I} \end{pmatrix} =: \begin{pmatrix} \mathbf{G}_{\mathcal{I}^*} & \mathbf{0} \\ \mathbf{G}_{\mathcal{A}_{*,\mathrm{c}}} & -\mathbf{I} \end{pmatrix},
$$

where $[\{\mathbf{J}_{\mathcal{R}_n}^{(\mathcal{T})}\}^{-1}]_{11}$ is a $p \times p$ matrix, and $\mathbf{S}_{11}$ is an $|\mathcal{I}^*| \times |\mathcal{I}^*|$ matrix. Recall $\mathbf{V}_{\mathcal{I}^*}^{(\mathcal{T})}(\boldsymbol{\theta}_0) = \mathbb{E}\{\mathbf{g}_{i,\mathcal{I}^*}^{(\mathcal{T})}(\boldsymbol{\theta}_0)^{\otimes 2}\}$ and $\mathbf{J}_{\mathcal{I}^*}^{(\mathcal{T})} = ([\mathbb{E}\{\nabla_{\boldsymbol{\theta}} \mathbf{g}_{i,\mathcal{I}^*}^{(\mathcal{T})}(\boldsymbol{\theta}_0)\}]^{\mathrm{T}}\{\mathbf{V}_{\mathcal{I}^*}^{(\mathcal{T})}(\boldsymbol{\theta}_0)\}^{-1/2})^{\otimes 2}$. Then $\mathbf{J}_{\mathcal{I}^*}^{(\mathcal{T})} = \mathbf{G}_{\mathcal{I}^*}^{\mathrm{T}}(\mathbf{S}_{11} - \mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21})\mathbf{G}_{\mathcal{I}^*}$. By (A.22), we have $[\{\mathbf{J}_{\mathcal{R}_n}^{(\mathcal{T})}\}^{-1}]_{11} = \{\mathbf{G}_{\mathcal{I}^*}^{\mathrm{T}}(\mathbf{S}_{11} - \mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21})\mathbf{G}_{\mathcal{I}^*}\}^{-1} = \{\mathbf{J}_{\mathcal{I}^*}^{(\mathcal{T})}\}^{-1}$. For any $\tilde{\boldsymbol{\alpha}} \in$

$\mathbb{R}^p$ with unit L_2-norm, let $\boldsymbol{\alpha} = \{\mathbf{J}^{(\mathcal{T})}_{\mathcal{R}_n}\}^{-1/2}[\{\mathbf{J}^{(\mathcal{T})}_{\mathcal{I}^*}\}^{1/2}, \mathbf{0}]^{\mathrm{T}}\tilde{\boldsymbol{\alpha}}$. Then $|\boldsymbol{\alpha}|_2^2 = \tilde{\boldsymbol{\alpha}}^{\mathrm{T}}\tilde{\boldsymbol{\alpha}} = 1$. Hence, $\tilde{\boldsymbol{\alpha}}^{\mathrm{T}}\{\mathbf{J}^{(\mathcal{T})}_{\mathcal{I}^*}\}^{1/2}\{\hat{\boldsymbol{\theta}}_{\mathrm{PEL}} - \boldsymbol{\theta}_0 - \hat{\boldsymbol{\zeta}}_{\mathcal{R}_n,(1)}\} = \boldsymbol{\alpha}^{\mathrm{T}}\{\mathbf{J}^{(\mathcal{T})}_{\mathcal{R}_n}\}^{1/2}(\hat{\boldsymbol{\psi}}_{\mathrm{PEL},\mathcal{S}_*} - \boldsymbol{\psi}_{0,\mathcal{S}_*} - \hat{\boldsymbol{\zeta}}_{\mathcal{R}_n}) \xrightarrow{d} \mathcal{N}(0,1)$, where $\hat{\boldsymbol{\zeta}}_{\mathcal{R}_n,(1)}$ is the first p components of $\hat{\boldsymbol{\zeta}}_{\mathcal{R}_n}$. We complete the proof of Theorem 3.1. □

A.5 Proof of Theorem 4.1

Recall $a_n = \sum_{k\in\mathcal{D}} P_{1,\pi}(|\xi_{0,k}|) + \sum_{l\in\mathcal{P}} P_{1,\pi}(|\theta_{0,l}|)$ and $\mathcal{S} = \mathcal{P}_\sharp \cup \mathcal{A}^{\mathrm{c}}$ with $s = |\mathcal{S}|$ in the current setting. Define $b_{1,n} = \max\{a_n, r_1\aleph_n^2\}$ and $b_{2,n} = \max\{b_{1,n}, \nu^2\}$. Then $\phi_n = \max\{p_\sharp b_{1,n}^{1/2}, b_{2,n}^{1/2}\}$. Notice that $\mathcal{M}^*_{\boldsymbol{\psi}} = \mathcal{I} \cup \mathcal{D}^*_{\boldsymbol{\psi}}$ with $\mathcal{D}^*_{\boldsymbol{\psi}} = \{j \in \mathcal{D} : |\bar{g}_j^{(\mathcal{T})}(\boldsymbol{\psi})| \geq C_*\nu\rho_2'(0^+)\}$ for any $\boldsymbol{\psi} \in \boldsymbol{\Psi}$, where $C_* \in (0,1)$ is a prescribed constant. Recall $\mathcal{D}_{\boldsymbol{\psi}}(c) = \{j \in \mathcal{D} : |\bar{g}_j^{(\mathcal{T})}(\boldsymbol{\psi})| \geq c\nu\rho_2'(0^+)\}$ for any $c \in (C_*, 1)$ and $\mathcal{M}_{\boldsymbol{\psi}}(c) = \mathcal{I} \cup \mathcal{D}_{\boldsymbol{\psi}}(c)$. In this section, we redefine

$$S_n(\boldsymbol{\psi}) = \max_{\boldsymbol{\lambda}\in\hat{\Lambda}_n^{(\mathcal{T})}(\boldsymbol{\psi})} f(\boldsymbol{\lambda};\boldsymbol{\psi}) + \sum_{k\in\mathcal{D}} P_{1,\pi}(|\xi_k|) + \sum_{l\in\mathcal{P}} P_{1,\pi}(|\theta_l|) \tag{A.25}$$

for any $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_r)^{\mathrm{T}}$ and $\boldsymbol{\psi} = (\theta_1, \ldots, \theta_p, \xi_1, \ldots, \xi_{r_2})^{\mathrm{T}} \in \boldsymbol{\Psi}$, where $f(\boldsymbol{\lambda};\boldsymbol{\psi})$ is defined as (A.9) in Section A.3. In comparison to $S_n(\boldsymbol{\psi})$ defined in Section A.3 for a low-dimensional $\boldsymbol{\theta}$, the newly defined $S_n(\boldsymbol{\psi})$ here for a high-dimensional $\boldsymbol{\theta}$ has an extra term $\sum_{l\in\mathcal{P}} P_{1,\pi}(|\theta_l|)$ which is caused by the penalty imposed on $\boldsymbol{\theta}$. Write $\boldsymbol{\xi}_0 = (\xi_{0,1}, \ldots, \xi_{0,r_2})^{\mathrm{T}}$ and $\boldsymbol{\theta}_0 = (\theta_{0,1}, \ldots, \theta_{0,p})^{\mathrm{T}}$. Recall $\mathcal{P}_\sharp = \{k \in \mathcal{P} : \theta_{0,k} \neq 0\}$. Then $\boldsymbol{\psi}_{\mathcal{S}} = (\boldsymbol{\theta}^{\mathrm{T}}_{\mathcal{P}_\sharp}, \boldsymbol{\xi}^{\mathrm{T}}_{\mathcal{A}^{\mathrm{c}}})^{\mathrm{T}}$ and $\boldsymbol{\psi}_{0,\mathcal{S}^{\mathrm{c}}} = \mathbf{0}$. Similar to that in Section A.3, we define $\boldsymbol{\Psi}_* = \{\boldsymbol{\psi} \in \boldsymbol{\Psi} : |\boldsymbol{\psi}_{\mathcal{S}} - \boldsymbol{\psi}_{0,\mathcal{S}}|_\infty \leq \varepsilon, |\boldsymbol{\psi}_{\mathcal{S}^{\mathrm{c}}}|_1 \leq \aleph_n\}$ for some fixed $\varepsilon > 0$. Consider

$$\hat{\boldsymbol{\psi}} = \arg\min_{\boldsymbol{\psi}\in\boldsymbol{\Psi}_*} S_n(\boldsymbol{\psi})\,. \tag{A.26}$$

Analogously to Proposition 3.1, here Proposition A.2 shows that such defined $\hat{\boldsymbol{\psi}}$ is a sparse local minimizer for the nonconvex optimization (4.1).

Proposition A.2. *Let $P_{1,\pi}(\cdot), P_{2,\nu}(\cdot) \in \mathscr{P}$ for $\mathscr{P}$ defined as* (3.2)*, and $P_{2,\nu}(\cdot)$ be convex with bounded second derivative around* 0*. Let $b_{1,n} = \max\{a_n, r_1\aleph_n^2\}$ with $a_n = \sum_{k\in\mathcal{D}} P_{1,\pi}(|\xi_{0,k}|) + \sum_{l\in\mathcal{P}} P_{1,\pi}(|\theta_{0,l}|)$, $b_{2,n} = \max\{b_{1,n}, \nu^2\}$, and $\phi_n = \max\{p_\sharp b_{1,n}^{1/2}, b_{2,n}^{1/2}\}$. For $\hat{\boldsymbol{\psi}}$ defined as* (A.26)*, assume that there exists a constant $\tilde{c} \in (C_*, 1)$ such that $\mathbb{P}[\cup_{j\in\mathcal{T}}\{|\bar{g}_j^{(\mathcal{T})}(\hat{\boldsymbol{\psi}})| \in [\tilde{c}\nu\rho_2'(0^+), \nu\rho_2'(0^+))\}] \to 0$. Under Conditions* 1′*,* 2–4 *and* (4.4)*, if $\log r = o(n^{1/3})$, $s^2\ell_n\phi_n^2 = o(1)$, $b_{2,n} = o(n^{-2/\gamma})$, and $\ell_n\aleph_n = o(\min\{\nu, \pi\})$, then w.p.a.1 such defined $\hat{\boldsymbol{\psi}} = (\hat{\boldsymbol{\theta}}^{\mathrm{T}}, \hat{\boldsymbol{\xi}}^{\mathrm{T}})^{\mathrm{T}}$ provides a local minimizer for the nonconvex optimization* (4.1) *such that* (i) $|\hat{\boldsymbol{\theta}}_{\mathcal{P}_\sharp} - \boldsymbol{\theta}_{0,\mathcal{P}_\sharp}|_\infty = O_{\mathrm{p}}(b_{1,n}^{1/2})$, (ii) $\mathbb{P}(\hat{\boldsymbol{\theta}}_{\mathcal{P}_\sharp^{\mathrm{c}}} = \mathbf{0}) \to 1$ *as* $n \to \infty$, (iii) $|\hat{\boldsymbol{\xi}}_{\mathcal{A}^{\mathrm{c}}} - \boldsymbol{\xi}_{0,\mathcal{A}^{\mathrm{c}}}|_\infty = O_{\mathrm{p}}(\phi_n)$, *and* (iv) $\mathbb{P}(\hat{\boldsymbol{\xi}}_{\mathcal{A}} = \mathbf{0}) \to 1$ *as* $n \to \infty$.

Since $f(\boldsymbol{\lambda};\boldsymbol{\psi})$ involved here for high-dimensional $\boldsymbol{\theta}$ is identical to that used in Section A.3 for low-dimensional $\boldsymbol{\theta}$, Lemmas A.2 and A.3 still hold in the the current setting. With the newly

defined $\mathcal{S}$ for high-dimensional $\boldsymbol{\theta}$, Lemma A.1 also holds in the current setting. The proof of Proposition A.2 is almost identical to that of Proposition 3.1. Using the same arguments as those in the proof of Proposition 3.1, we obtain $\max_{\boldsymbol{\lambda}\in\hat{\Lambda}_n^{(\mathcal{T})}(\boldsymbol{\psi}_0)} f(\boldsymbol{\lambda};\boldsymbol{\psi}_0) = O_{\mathrm{p}}(r_1\aleph_n^2)$. Recall that $b_{1,n} = \max\{a_n, r_1\aleph_n^2\}$ and $a_n = \sum_{k\in\mathcal{D}} P_{1,\pi}(|\xi_{0,k}|) + \sum_{l\in\mathcal{P}} P_{1,\pi}(|\theta_{0,l}|)$. Then $S_n(\boldsymbol{\psi}_0) = O_{\mathrm{p}}(r_1\aleph_n^2) + a_n = O_{\mathrm{p}}(b_{1,n})$. Notice that $\hat{\boldsymbol{\psi}} = \arg\min_{\boldsymbol{\psi}\in\boldsymbol{\Psi}_*} S_n(\boldsymbol{\psi})$ with $\boldsymbol{\Psi}_* = \{\boldsymbol{\psi} = (\boldsymbol{\psi}_{\mathcal{S}}^{\mathrm{T}}, \boldsymbol{\psi}_{\mathcal{S}^{\mathrm{c}}}^{\mathrm{T}})^{\mathrm{T}} : |\boldsymbol{\psi}_{\mathcal{S}} - \boldsymbol{\psi}_{0,\mathcal{S}}|_\infty \le \varepsilon, |\boldsymbol{\psi}_{\mathcal{S}^{\mathrm{c}}}|_1 \le \aleph_n\}$, and $\boldsymbol{\psi}_{0,\mathcal{S}^{\mathrm{c}}} = \mathbf{0}$. We then have $\boldsymbol{\psi}_0 \in \boldsymbol{\Psi}_*$ which implies $S_n(\hat{\boldsymbol{\psi}}) \le S_n(\boldsymbol{\psi}_0) = O_{\mathrm{p}}(b_{1,n})$. We need to show $\hat{\boldsymbol{\psi}} \in \mathrm{int}(\boldsymbol{\Psi}_*)$ w.p.a.1, which indicates that $\hat{\boldsymbol{\psi}}$ is a local minimizer of $S_n(\boldsymbol{\psi})$.

We now follow a slightly different line of proof: (i) to show that for any $\epsilon_n \to \infty$ satisfying $b_{1,n}\epsilon_n^2 n^{2/\gamma} = o(1)$ and any $\boldsymbol{\psi} = (\boldsymbol{\theta}^{\mathrm{T}}, \boldsymbol{\xi}^{\mathrm{T}})^{\mathrm{T}} \in \boldsymbol{\Psi}_*$ satisfying $|\boldsymbol{\theta}_{\mathcal{P}_\sharp} - \boldsymbol{\theta}_{0,\mathcal{P}_\sharp}|_\infty > \epsilon_n b_{1,n}^{1/2}$, there exists a universal constant $K > 0$ independent of $\boldsymbol{\psi}$ such that $\mathbb{P}\{S_n(\boldsymbol{\psi}) > Kb_{1,n}\epsilon_n^2\} \to 1$ as $n \to \infty$. Due to $b_{1,n} = o(n^{-2/\gamma})$, we can select an arbitrary slowly diverging ϵ_n satisfying $b_{1,n}\epsilon_n^2 n^{2/\gamma} = o(1)$. Thus, we have $|\hat{\boldsymbol{\theta}}_{\mathcal{P}_\sharp} - \boldsymbol{\theta}_{0,\mathcal{P}_\sharp}|_\infty = O_{\mathrm{p}}(b_{1,n}^{1/2})$; (ii) to show that for any $\varepsilon_n \to \infty$ satisfying $b_{2,n}\varepsilon_n^2 n^{2/\gamma} = o(1)$ and $\boldsymbol{\psi} = (\boldsymbol{\theta}^{\mathrm{T}}, \boldsymbol{\xi}_{\mathcal{A}}^{\mathrm{T}}, \boldsymbol{\xi}_{\mathcal{A}^{\mathrm{c}}}^{\mathrm{T}})^{\mathrm{T}} \in \boldsymbol{\Psi}_*$ satisfying $|\boldsymbol{\theta}_{\mathcal{P}_\sharp} - \boldsymbol{\theta}_{0,\mathcal{P}_\sharp}|_\infty \le O(\varepsilon_n^{1/2} b_{1,n}^{1/2})$ and $|\boldsymbol{\xi}_{\mathcal{A}^{\mathrm{c}}} - \boldsymbol{\xi}_{0,\mathcal{A}^{\mathrm{c}}}|_\infty > \varepsilon_n\phi_n$, there exists a universal constant $M > 0$ independent of $\boldsymbol{\psi}$ such that $\mathbb{P}\{S_n(\boldsymbol{\psi}) > Mb_{2,n}\varepsilon_n^2\} \to 1$ as $n \to \infty$. Since $|\hat{\boldsymbol{\theta}}_{\mathcal{P}_\sharp} - \boldsymbol{\theta}_{0,\mathcal{P}_\sharp}|_\infty \le O(\varepsilon_n^{1/2} b_{1,n}^{1/2})$ w.p.a.1 and we can select an arbitrary slowly diverging ε_n satisfying $b_{2,n}\varepsilon_n^2 n^{2/\gamma} = o(1)$, it holds that $|\hat{\boldsymbol{\xi}}_{\mathcal{A}^{\mathrm{c}}} - \boldsymbol{\xi}_{0,\mathcal{A}^{\mathrm{c}}}|_\infty = O_{\mathrm{p}}(\phi_n)$; (iii) to show that $\hat{\boldsymbol{\psi}}_{\mathcal{S}^{\mathrm{c}}} = \mathbf{0}$ w.p.a.1.

The proof of Part (i) is similar to that of Proposition 2 in Chang et al. (2018). For any $\boldsymbol{\psi} = (\boldsymbol{\theta}^{\mathrm{T}}, \boldsymbol{\xi}^{\mathrm{T}})^{\mathrm{T}} \in \boldsymbol{\Psi}_*$ with $\boldsymbol{\theta} = (\boldsymbol{\theta}_{\mathcal{P}_\sharp}^{\mathrm{T}}, \boldsymbol{\theta}_{\mathcal{P}_\sharp^{\mathrm{c}}}^{\mathrm{T}})^{\mathrm{T}}$ satisfying $|\boldsymbol{\theta}_{\mathcal{P}_\sharp} - \boldsymbol{\theta}_{0,\mathcal{P}_\sharp}|_\infty > \epsilon_n b_{1,n}^{1/2}$, take $\boldsymbol{\theta}^* = (\boldsymbol{\theta}_{\mathcal{P}_\sharp}^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}}$ and $j_0 = \arg\max_{j\in\mathcal{I}} |\mathbb{E}\{g_{i,j}^{(\mathcal{I})}(\boldsymbol{\theta}^*)\}|$. Let $\mu_{j_0} = \mathbb{E}\{g_{i,j_0}^{(\mathcal{I})}(\boldsymbol{\theta})\}$ and $\mu_{j_0}^* = \mathbb{E}\{g_{i,j_0}^{(\mathcal{I})}(\boldsymbol{\theta}^*)\}$. Select $\tilde{\boldsymbol{\lambda}} = \delta b_{1,n}^{1/2}\epsilon_n \mathbf{e}_{j_0}$, where $\delta > 0$ is a constant to be determined later, and $\mathbf{e}_{j_0}$ is an r-dimensional vector with the j_0-th component being 1 and other components being 0. Then $\tilde{\boldsymbol{\lambda}} \in \hat{\Lambda}_n^{(\mathcal{T})}(\boldsymbol{\psi})$ w.p.a.1. For the newly defined $S_n(\boldsymbol{\psi})$ in (A.25), applying the identical arguments for proof of Part (i) in Section A.3, we still have

$$\mathbb{P}\{S_n(\boldsymbol{\psi}) \le Kb_{1,n}\epsilon_n^2\} \le \mathbb{P}\left[\bar{g}_{j_0}^{(\mathcal{I})}(\boldsymbol{\theta}) - \mu_{j_0} \le b_{1,n}^{1/2}\epsilon_n\left\{\frac{K}{\delta} + \frac{\delta}{n}\sum_{i=1}^n |g_{i,j_0}^{(\mathcal{I})}(\boldsymbol{\theta})|^2\right\} - \mu_{j_0}\right] + o(1)\,.$$

Condition 1′ implies that $\mu_{j_0}^* \ge K_1'\epsilon_n b_{1,n}^{1/2}$ with K_1' specified in Condition 1′, and $|\mu_{j_0} - \mu_{j_0}^*| \le K_3'|\boldsymbol{\theta}_{\mathcal{P}_\sharp^c}|_1 \le K_3'\aleph_n \le K_1'\epsilon_n b_{1,n}^{1/2}/2$ for sufficiently large n, which implies $\mu_{j_0} \ge K_1'\epsilon_n b_{1,n}^{1/2}/2$ for sufficiently large n. Using the same arguments stated in the proof of Proposition A.1, we have $\mathbb{P}\{S_n(\boldsymbol{\psi}) > Kb_{1,n}\epsilon_n^2\} \to 1$ as $n \to \infty$. The proof of Part (ii) and Part (iii) are almost identical to that of Proposition 3.1, except some small adjustments. The first difference is for deriving the lower bound of μ_{j_0} appeared in the proof of Part (ii). Notice that $|\mathbb{E}\{g_{i,j_0}^{(\mathcal{D})}(\boldsymbol{\theta})\} - \mathbb{E}\{g_{i,j_0}^{(\mathcal{D})}(\boldsymbol{\theta}_0)\}| \le |\mathbb{E}\{\nabla_{\boldsymbol{\theta}} g_{i,j_0}^{(\mathcal{D})}(\dot{\boldsymbol{\theta}})\}|_\infty |\boldsymbol{\theta} - \boldsymbol{\theta}_0|_1 \le O(\varepsilon_n^{1/2} p_\sharp b_{1,n}^{1/2}) + O(\varepsilon_n^{1/2}\aleph_n) = o(\varepsilon_n\phi_n)$.

Hence, identical to (A.13), we still have $\mu_{j_0} \geq \varepsilon_n b_{2,n}^{1/2}/2$ when n is sufficiently large. The second difference is (A.14). In the current setting, it should be

$$\max_{\boldsymbol{\lambda}\in\hat{\Lambda}_n^{(\mathcal{T})}(\hat{\boldsymbol{\psi}})} f(\boldsymbol{\lambda};\hat{\boldsymbol{\psi}}) \leq \max_{\boldsymbol{\lambda}\in\hat{\Lambda}_n^{(\mathcal{T})}(\boldsymbol{\psi}_0)} f(\boldsymbol{\lambda};\boldsymbol{\psi}_0) + \sum_{k\in\mathcal{S}} P_{1,\pi}(|\psi_{0,k}|) - \sum_{k\in\mathcal{S}} P_{1,\pi}(|\hat{\psi}_k|) = O_{\mathrm{p}}(r_1\aleph_n^2)\,,$$

where $\hat{\boldsymbol{\psi}} = (\hat{\psi}_1,\ldots,\hat{\psi}_{p+r_2})^{\mathrm{T}}$ and $\boldsymbol{\psi}_0 = (\psi_{0,1},\ldots,\psi_{0,p+r_2})^{\mathrm{T}}$. The third difference is that the index set $\mathcal{A}$ in (A.15) should be replaced by $\mathcal{S}^{\mathrm{c}}$ due to the newly defined $S_n(\boldsymbol{\psi})$ in the current setting. Then (A.16) changes to

$$S_n(\hat{\boldsymbol{\psi}}^*) \leq S_n(\hat{\boldsymbol{\psi}}) \underbrace{- \frac{1}{n}\sum_{i=1}^n \frac{\hat{\boldsymbol{\lambda}}^{*,\mathrm{T}}\nabla_{\boldsymbol{\psi}_{\mathcal{S}^{\mathrm{c}}}}\mathbf{g}_i^{(\mathcal{T})}(\check{\boldsymbol{\psi}})}{1+\hat{\boldsymbol{\lambda}}^{*,\mathrm{T}}\mathbf{g}_i^{(\mathcal{T})}(\check{\boldsymbol{\psi}})}\hat{\boldsymbol{\psi}}_{\mathcal{S}^{\mathrm{c}}}}_{\mathrm{I}} \underbrace{- \sum_{k\in\mathcal{S}^{\mathrm{c}}} P_{1,\pi}(|\hat{\psi}_k|)}_{\mathrm{II}}\,.$$

The last difference appears in the upper bound of |I|. Since $\mathbf{g}_i^{(\mathcal{T})}(\boldsymbol{\psi}) = \{\mathbf{g}_i^{(\mathcal{I})}(\boldsymbol{\theta})^{\mathrm{T}}, \mathbf{g}_i^{(\mathcal{D})}(\boldsymbol{\theta})^{\mathrm{T}} - \boldsymbol{\xi}^{\mathrm{T}}\}^{\mathrm{T}}$ and $\boldsymbol{\psi}_{\mathcal{S}^{\mathrm{c}}} = (\boldsymbol{\theta}_{\mathcal{P}_\sharp^{\mathrm{c}}}^{\mathrm{T}}, \boldsymbol{\xi}_{\mathcal{A}}^{\mathrm{T}})^{\mathrm{T}}$, we now have

$$|\mathrm{I}| = \left|\frac{1}{n}\sum_{i=1}^n \frac{\hat{\boldsymbol{\lambda}}^{*,\mathrm{T}}\nabla_{\boldsymbol{\psi}_{\mathcal{S}^{\mathrm{c}}}}\mathbf{g}_i^{(\mathcal{T})}(\check{\boldsymbol{\psi}})}{1+\hat{\boldsymbol{\lambda}}^{*,\mathrm{T}}\mathbf{g}_i^{(\mathcal{T})}(\check{\boldsymbol{\psi}})}\hat{\boldsymbol{\psi}}_{\mathcal{S}^{\mathrm{c}}}\right| \leq \ell_n^{1/2}|\hat{\boldsymbol{\lambda}}^*|_2|\hat{\boldsymbol{\psi}}_{\mathcal{S}^{\mathrm{c}}}|_1\{1+o_{\mathrm{p}}(1)\}\,,$$

where the upper bound has an extra factor $\ell_n^{1/2}$ in comparison to (A.18). Due to $|\hat{\boldsymbol{\lambda}}^*|_2 = O_{\mathrm{p}}(\ell_n^{1/2}\aleph_n)$, it then holds $|\mathrm{I}| \leq |\hat{\boldsymbol{\psi}}_{\mathcal{S}^{\mathrm{c}}}|_1 \cdot O_{\mathrm{p}}(\ell_n\aleph_n)$. Also notice that $\mathrm{II} = \sum_{k\in\mathcal{S}^{\mathrm{c}}} \pi\rho_1'(c_k|\hat{\psi}_k|;\pi)|\hat{\psi}_k| \geq C\pi|\hat{\boldsymbol{\psi}}_{\mathcal{S}^{\mathrm{c}}}|_1$ for some $c_k \in (0,1)$. Then $\ell_n\aleph_n = o(\pi)$ is required for Proposition A.2 rather than $\ell_n^{1/2}\aleph_n = o(\pi)$ required in Proposition 3.1.

Select $\hat{\boldsymbol{\psi}}_{\mathrm{PEL}}$ as the sparse local minimizer given in Proposition A.2. Recall $\mathcal{R}_n = \mathcal{I}\cup\mathrm{supp}(\hat{\boldsymbol{\lambda}}_{\mathcal{D}})$, $\mathcal{A}_* = \{j\in\mathcal{A}: \hat{\lambda}_j \neq 0\}$ and $\mathcal{A}_{*,\mathrm{c}} = \{j\in\mathcal{A}^{\mathrm{c}}: \hat{\lambda}_j \neq 0\}$. Notice that $\mathcal{S}_* = \mathcal{P}_\sharp \cup \mathcal{A}_{*,\mathrm{c}}$ and $\mathcal{S} = \mathcal{P}_\sharp \cup \mathcal{A}^{\mathrm{c}}$. Then $\mathcal{S}_* \subset \mathcal{S}$ and $s_* := |\mathcal{S}_*| \leq |\mathcal{S}| = s$. For any $\boldsymbol{\psi}\in\boldsymbol{\Psi}$, we have $\boldsymbol{\psi}_{\mathcal{S}_*} = (\boldsymbol{\theta}_{\mathcal{P}_\sharp}^{\mathrm{T}}, \boldsymbol{\xi}_{\mathcal{A}_{*,\mathrm{c}}}^{\mathrm{T}})^{\mathrm{T}}$. Under the conditions of Proposition A.2, the results of Lemmas A.4 and A.5 hold with the newly defined $\hat{\boldsymbol{\psi}}_{\mathrm{PEL}}$, $\mathcal{S}_*$, $\mathcal{R}_n$, ℓ_n, s, s_* and ϕ_n. The proof of Theorem 4.1 is almost identical to that of Theorem 3.1 stated in Section A.4. We only point out the difference here. The first difference is the definition of $H_n(\boldsymbol{\psi},\boldsymbol{\lambda})$. In comparison to $H_n(\boldsymbol{\psi},\boldsymbol{\lambda})$ given in (A.19) for the low-dimensional $\boldsymbol{\theta}$, we define

$$H_n(\boldsymbol{\psi},\boldsymbol{\lambda}) = \frac{1}{n}\sum_{i=1}^n \log\{1+\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{g}_i^{(\mathcal{T})}(\boldsymbol{\psi})\} + \sum_{k\in\mathcal{D}} P_{1,\pi}(|\xi_k|) + \sum_{l\in\mathcal{P}} P_{1,\pi}(|\theta_l|) - \sum_{j\in\mathcal{D}} P_{2,\nu}(|\lambda_j|)$$

in current high-dimensional setting. Following the same arguments stated in Section A.4, (A.21) still holds with $\hat{\boldsymbol{\varsigma}}_{\mathcal{S}_*} = \{\sum_{k=1}^{p+r_2}\nabla_{\boldsymbol{\psi}_{\mathcal{S}_*}} P_{1,\pi}(|\psi_k|)\}|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}_{\mathrm{PEL}}}$. It follows from Proposition A.2 that $\hat{\boldsymbol{\varsigma}}_{\mathcal{S}_*} = \mathbf{0}$

w.p.a.1. Notice that Lemma A.6 still holds in the current setting. Identical to the arguments below (A.21), it holds that $n^{1/2}\boldsymbol{\alpha}^{\mathrm{T}}\{\mathbf{J}_{\mathcal{R}_n}^{(\mathcal{T})}\}^{1/2}(\hat{\boldsymbol{\psi}}_{\mathrm{PEL},\mathcal{S}_*}-\boldsymbol{\psi}_{0,\mathcal{S}_*}-\hat{\boldsymbol{\zeta}}_{\mathcal{R}_n})\xrightarrow{d}\mathcal{N}(0,1)$ as $n\to\infty$. Notice that $\hat{\boldsymbol{\psi}}_{\mathrm{PEL},\mathcal{S}_*}=(\hat{\boldsymbol{\theta}}_{\mathrm{PEL},\mathcal{P}_\sharp}^{\mathrm{T}},\hat{\boldsymbol{\xi}}_{\mathcal{A}_{*,\mathrm{c}}}^{\mathrm{T}})^{\mathrm{T}}$. Recall $\mathcal{R}_n=\mathcal{I}\cup\mathcal{A}_*\cup\mathcal{A}_{*,\mathrm{c}}$ and $\mathcal{I}^*=\mathcal{I}\cup\mathcal{A}_*$. We write $\{\mathbf{J}_{\mathcal{R}_n}^{(\mathcal{T})}\}^{-1}$, $\mathbb{E}\{\nabla_{\boldsymbol{\psi}_{\mathcal{S}_*}}\mathbf{g}_{i,\mathcal{R}_n}^{(\mathcal{T})}(\boldsymbol{\psi}_0)\}$ and $\{\mathbf{V}_{\mathcal{R}_n}^{(\mathcal{T})}(\boldsymbol{\psi}_0)\}^{-1}$ with following blocks

$$\{\mathbf{J}_{\mathcal{R}_n}^{(\mathcal{T})}\}^{-1}=\begin{pmatrix}[\{\mathbf{J}_{\mathcal{R}_n}^{(\mathcal{T})}\}^{-1}]_{11} & [\{\mathbf{J}_{\mathcal{R}_n}^{(\mathcal{T})}\}^{-1}]_{12}\\ [\{\mathbf{J}_{\mathcal{R}_n}^{(\mathcal{T})}\}^{-1}]_{21} & [\{\mathbf{J}_{\mathcal{R}_n}^{(\mathcal{T})}\}^{-1}]_{22}\end{pmatrix},\quad \{\mathbf{V}_{\mathcal{R}_n}^{(\mathcal{T})}(\boldsymbol{\psi}_0)\}^{-1}=\begin{pmatrix}\mathbf{S}_{11} & \mathbf{S}_{12}\\ \mathbf{S}_{21} & \mathbf{S}_{22}\end{pmatrix},$$
$$\mathbb{E}\{\nabla_{\boldsymbol{\psi}_{\mathcal{S}_*}}\mathbf{g}_{i,\mathcal{R}_n}^{(\mathcal{T})}(\boldsymbol{\psi}_0)\}=\begin{pmatrix}\mathbb{E}\{\nabla_{\boldsymbol{\theta}_{\mathcal{P}_\sharp}}\mathbf{g}_{i,\mathcal{I}^*}^{(\mathcal{T})}(\boldsymbol{\theta}_0)\} & \mathbf{0}\\ \mathbb{E}\{\nabla_{\boldsymbol{\theta}_{\mathcal{P}_\sharp}}\mathbf{g}_{i,\mathcal{A}_{*,\mathrm{c}}}^{(\mathcal{D})}(\boldsymbol{\theta}_0)\} & -\mathbf{I}\end{pmatrix}=\begin{pmatrix}\mathbf{G}_{\mathcal{I}^*} & \mathbf{0}\\ \mathbf{G}_{\mathcal{A}_{*,\mathrm{c}}} & -\mathbf{I}\end{pmatrix},$$

where $[\{\mathbf{J}_{\mathcal{R}_n}^{(\mathcal{T})}\}^{-1}]_{11}$ is a $p_\sharp\times p_\sharp$ matrix, and $\mathbf{S}_{11}$ is an $|\mathcal{I}^*|\times|\mathcal{I}^*|$ matrix. Recall $\mathbf{V}_{\mathcal{I}^*}^{(\mathcal{T})}(\boldsymbol{\theta}_0)=\mathbb{E}\{\mathbf{g}_{i,\mathcal{I}^*}^{(\mathcal{T})}(\boldsymbol{\theta}_0)^{\otimes 2}\}$ and $\mathbf{W}_{\mathcal{I}^*}^{(\mathcal{T})}=([\mathbb{E}\{\nabla_{\boldsymbol{\theta}_{\mathcal{P}_\sharp}}\mathbf{g}_{i,\mathcal{I}^*}^{(\mathcal{T})}(\boldsymbol{\theta}_0)\}]^{\mathrm{T}}\{\mathbf{V}_{\mathcal{I}^*}^{(\mathcal{T})}(\boldsymbol{\theta}_0)\}^{-1/2})^{\otimes 2}$. Then $[\{\mathbf{J}_{\mathcal{R}_n}^{(\mathcal{T})}\}^{-1}]_{11}=\{\mathbf{G}_{\mathcal{I}^*}^{\mathrm{T}}(\mathbf{S}_{11}-\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21})\mathbf{G}_{\mathcal{I}^*}\}^{-1}=\{\mathbf{W}_{\mathcal{I}^*}^{(\mathcal{T})}\}^{-1}$. By the same arguments of Theorem 3.1, we complete the proof of Theorem 4.1. □

A.6 Proof of Theorem 5.1

To prove Theorem 5.1, we first present the following lemma whose proof is given in Section A.7.4.

Lemma A.7. *Let* $|\boldsymbol{\psi}_{\mathcal{M}}^*-\boldsymbol{\psi}_{0,\mathcal{M}}|_1=O_{\mathrm{p}}(\varpi_{1,n})$ *and* $|\boldsymbol{\psi}_{\mathcal{M}^{\mathrm{c}}}^*-\boldsymbol{\psi}_{0,\mathcal{M}^{\mathrm{c}}}|_1=O_{\mathrm{p}}(\varpi_{2,n})$ *for some* $\varpi_{1,n}\to 0$ *and* $\varpi_{2,n}\to 0$. *Under Conditions* 2–4 *and* 7, *if* $n\varpi_{2,n}^2(\varsigma^2+\varpi_{1,n}^2+\varpi_{2,n}^2)=O(1)$, $m(\omega_n+\varpi_{1,n}+\varpi_{2,n})=o(1)$, $mn^{-1/2+1/\gamma}=o(1)$ *and* $\omega_n^2\log r=O(1)$, *then* $|\tilde{\boldsymbol{\psi}}_{\mathcal{M}}-\boldsymbol{\psi}_{0,\mathcal{M}}|_2=O_{\mathrm{p}}(m^{1/2}n^{-1/2})$.

Now we begin to prove Theorem 5.1. Let $\hat{\boldsymbol{\lambda}}^*=\arg\max_{\boldsymbol{\lambda}\in\tilde{\Lambda}_n(\tilde{\boldsymbol{\psi}}_{\mathcal{M}})}n^{-1}\sum_{i=1}^n\log\{1+\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{f}_i^{\mathbf{A}_n}(\tilde{\boldsymbol{\psi}}_{\mathcal{M}},\boldsymbol{\psi}_{\mathcal{M}^{\mathrm{c}}}^*)\}$. By the same arguments in the proof of Lemma A.7 for bounding $\tilde{\boldsymbol{\lambda}}$ there, we have $|\hat{\boldsymbol{\lambda}}^*|_2=O_{\mathrm{p}}(m^{1/2}n^{-1/2})$. Identical to (A.6), it holds that

$$\mathbf{0}=\{\mathbf{D}^*(\tilde{\boldsymbol{\psi}}_{\mathcal{M}})\}^{\mathrm{T}}\{\mathbf{C}^*(\tilde{\boldsymbol{\psi}}_{\mathcal{M}})\}^{-1}\bar{\mathbf{f}}^{\mathbf{A}_n}(\tilde{\boldsymbol{\psi}}_{\mathcal{M}},\boldsymbol{\psi}_{0,\mathcal{M}^{\mathrm{c}}}^*),$$

where

$$\mathbf{D}^*(\tilde{\boldsymbol{\psi}}_{\mathcal{M}})=\frac{1}{n}\sum_{i=1}^n\frac{\nabla_{\boldsymbol{\psi}_{\mathcal{M}}}\mathbf{f}_i^{\mathbf{A}_n}(\tilde{\boldsymbol{\psi}}_{\mathcal{M}},\boldsymbol{\psi}_{\mathcal{M}^{\mathrm{c}}}^*)}{1+\hat{\boldsymbol{\lambda}}^{*,\mathrm{T}}\mathbf{f}_i^{\mathbf{A}_n}(\tilde{\boldsymbol{\psi}}_{\mathcal{M}},\boldsymbol{\psi}_{\mathcal{M}^{\mathrm{c}}}^*)}\quad\text{and}\quad\mathbf{C}^*(\tilde{\boldsymbol{\psi}}_{\mathcal{M}})=\frac{1}{n}\sum_{i=1}^n\frac{\{\mathbf{f}_i^{\mathbf{A}_n}(\tilde{\boldsymbol{\psi}}_{\mathcal{M}},\boldsymbol{\psi}_{\mathcal{M}^{\mathrm{c}}}^*)\}^{\otimes 2}}{\{1+c\hat{\boldsymbol{\lambda}}^{*,\mathrm{T}}\mathbf{f}_i^{\mathbf{A}_n}(\tilde{\boldsymbol{\psi}}_{\mathcal{M}},\boldsymbol{\psi}_{\mathcal{M}^{\mathrm{c}}}^*)\}^2}$$

for some $|c|<1$. Write $\widehat{\mathbf{V}}_{\mathbf{f}^{\mathbf{A}_n}}(\tilde{\boldsymbol{\psi}}_{\mathcal{M}},\boldsymbol{\psi}_{\mathcal{M}^{\mathrm{c}}}^*)=n^{-1}\sum_{i=1}^n\{\mathbf{f}_i^{\mathbf{A}_n}(\tilde{\boldsymbol{\psi}}_{\mathcal{M}},\boldsymbol{\psi}_{\mathcal{M}^{\mathrm{c}}}^*)\}^{\otimes 2}$. Similar to Lemma A.6, $\|\mathbf{C}^*(\tilde{\boldsymbol{\psi}}_{\mathcal{M}})-\widehat{\mathbf{V}}_{\mathbf{f}^{\mathbf{A}_n}}(\tilde{\boldsymbol{\psi}}_{\mathcal{M}},\boldsymbol{\psi}_{\mathcal{M}^{\mathrm{c}}}^*)\|_2=O_{\mathrm{p}}(mn^{-1/2+1/\gamma})$, and $|\{\mathbf{D}^*(\tilde{\boldsymbol{\psi}}_{\mathcal{M}})-\nabla_{\boldsymbol{\psi}_{\mathcal{M}}}\bar{\mathbf{f}}^{\mathbf{A}_n}(\tilde{\boldsymbol{\psi}}_{\mathcal{M}},\boldsymbol{\psi}_{\mathcal{M}^{\mathrm{c}}}^*)\}\mathbf{z}|_2=|\mathbf{z}|_2\cdot O_{\mathrm{p}}(m^{3/2}n^{-1/2})$ holds uniformly over $\mathbf{z}\in\mathbb{R}^m$. Let $\hat{\mathbf{J}}^*=[\{\nabla_{\boldsymbol{\psi}_{\mathcal{M}}}\bar{\mathbf{f}}^{\mathbf{A}_n}(\tilde{\boldsymbol{\psi}}_{\mathcal{M}},\boldsymbol{\psi}_{\mathcal{M}^{\mathrm{c}}}^*)\}^{\mathrm{T}}\{\widehat{\mathbf{V}}_{\mathbf{f}^{\mathbf{A}_n}}(\tilde{\boldsymbol{\psi}}_{\mathcal{M}},\boldsymbol{\psi}_{\mathcal{M}^{\mathrm{c}}}^*)\}^{-1/2}]^{\otimes 2}$.

For any $\boldsymbol{\alpha} \in \mathbb{R}^m$, let $\boldsymbol{\delta} = (\hat{\mathbf{J}}^*)^{-1/2}\boldsymbol{\alpha}$, and it holds that

$$\boldsymbol{\delta}^{\mathrm{T}}\{\nabla_{\boldsymbol{\psi}_{\mathcal{M}}}\bar{\mathbf{f}}^{\mathbf{A}_n}(\tilde{\boldsymbol{\psi}}_{\mathcal{M}}, \boldsymbol{\psi}^*_{\mathcal{M}^{\mathrm{c}}})\}^{\mathrm{T}}\{\widehat{\mathbf{V}}_{\mathbf{f}^{\mathbf{A}_n}}(\tilde{\boldsymbol{\psi}}_{\mathcal{M}}, \boldsymbol{\psi}^*_{\mathcal{M}^{\mathrm{c}}})\}^{-1}\bar{\mathbf{f}}^{\mathbf{A}_n}(\tilde{\boldsymbol{\psi}}_{\mathcal{M}}, \boldsymbol{\psi}^*_{\mathcal{M}^{\mathrm{c}}}) = O_{\mathrm{p}}(m^{3/2}n^{-1+1/\gamma}) + O_{\mathrm{p}}(m^2n^{-1})\,.$$

Expanding $\bar{\mathbf{f}}^{\mathbf{A}_n}(\tilde{\boldsymbol{\psi}}_{\mathcal{M}}, \boldsymbol{\psi}^*_{\mathcal{M}^{\mathrm{c}}})$ near $\boldsymbol{\psi}_{\mathcal{M}} = \boldsymbol{\psi}_{0,\mathcal{M}}$, we obtain

$$\begin{aligned}
&\boldsymbol{\delta}^{\mathrm{T}}\{\nabla_{\boldsymbol{\psi}_{\mathcal{M}}}\bar{\mathbf{f}}^{\mathbf{A}_n}(\tilde{\boldsymbol{\psi}}_{\mathcal{M}}, \boldsymbol{\psi}^*_{\mathcal{M}^{\mathrm{c}}})\}^{\mathrm{T}}\{\widehat{\mathbf{V}}_{\mathbf{f}^{\mathbf{A}_n}}(\tilde{\boldsymbol{\psi}}_{\mathcal{M}}, \boldsymbol{\psi}^*_{\mathcal{M}^{\mathrm{c}}})\}^{-1}\{\nabla_{\boldsymbol{\psi}_{\mathcal{M}}}\bar{\mathbf{f}}^{\mathbf{A}_n}(\check{\boldsymbol{\psi}}_{\mathcal{M}}, \boldsymbol{\psi}^*_{\mathcal{M}^{\mathrm{c}}})\}(\tilde{\boldsymbol{\psi}}_{\mathcal{M}} - \boldsymbol{\psi}_{0,\mathcal{M}}) \\
=& -\boldsymbol{\delta}^{\mathrm{T}}\{\nabla_{\boldsymbol{\psi}_{\mathcal{M}}}\bar{\mathbf{f}}^{\mathbf{A}_n}(\tilde{\boldsymbol{\psi}}_{\mathcal{M}}, \boldsymbol{\psi}^*_{\mathcal{M}^{\mathrm{c}}})\}^{\mathrm{T}}\{\widehat{\mathbf{V}}_{\mathbf{f}^{\mathbf{A}_n}}(\tilde{\boldsymbol{\psi}}_{\mathcal{M}}, \boldsymbol{\psi}^*_{\mathcal{M}^{\mathrm{c}}})\}^{-1}\bar{\mathbf{f}}^{\mathbf{A}_n}(\boldsymbol{\psi}_{0,\mathcal{M}}, \boldsymbol{\psi}^*_{\mathcal{M}^{\mathrm{c}}}) \\
&+ O_{\mathrm{p}}(m^{3/2}n^{-1+1/\gamma}) + O_{\mathrm{p}}(m^2n^{-1})\,,
\end{aligned} \tag{A.27}$$

where $\check{\boldsymbol{\psi}}_{\mathcal{M}}$ is on the line joining $\tilde{\boldsymbol{\psi}}_{\mathcal{M}}$ and $\boldsymbol{\psi}_{0,\mathcal{M}}$. By Condition 3, we have $|\{\nabla_{\boldsymbol{\psi}_{\mathcal{M}}}\bar{\mathbf{f}}^{\mathbf{A}_n}(\check{\boldsymbol{\psi}}_{\mathcal{M}}, \boldsymbol{\psi}^*_{\mathcal{M}^{\mathrm{c}}}) - \nabla_{\boldsymbol{\psi}_{\mathcal{M}}}\bar{\mathbf{f}}^{\mathbf{A}_n}(\tilde{\boldsymbol{\psi}}_{\mathcal{M}}, \boldsymbol{\psi}^*_{\mathcal{M}^{\mathrm{c}}})\}(\tilde{\boldsymbol{\psi}}_{\mathcal{M}} - \boldsymbol{\psi}_{0,\mathcal{M}})|_2 = O_{\mathrm{p}}(m^{5/2}n^{-1})$. Moreover, by the proof of Theorem 1 in Chang et al. (2021), if $m\omega_n^2 \log r = o(1)$ and $nm\varpi_{2,n}^2(\varsigma^2 + \varpi_{1,n}^2 + \varpi_{2,n}^2) = o(1)$, we have $|\bar{\mathbf{f}}^{\mathbf{A}_n}(\boldsymbol{\psi}_{0,\mathcal{M}}, \boldsymbol{\psi}^*_{\mathcal{M}^{\mathrm{c}}}) - \bar{\mathbf{f}}^{\mathbf{A}}(\boldsymbol{\psi}_0)|_2 = o_{\mathrm{p}}(n^{-1/2})$. Together with (A.27), it holds that

$$\begin{aligned}
n^{1/2}\boldsymbol{\alpha}^{\mathrm{T}}(\hat{\mathbf{J}}^*)^{1/2}(\tilde{\boldsymbol{\psi}}_{\mathcal{M}} - \boldsymbol{\psi}_{0,\mathcal{M}}) =& -n^{1/2}\boldsymbol{\alpha}^{\mathrm{T}}(\hat{\mathbf{J}}^*)^{-1/2}\{\nabla_{\boldsymbol{\psi}_{\mathcal{M}}}\bar{\mathbf{f}}^{\mathbf{A}_n}(\tilde{\boldsymbol{\psi}}_{\mathcal{M}}, \boldsymbol{\psi}^*_{\mathcal{M}^{\mathrm{c}}})\}^{\mathrm{T}}\{\widehat{\mathbf{V}}_{\mathbf{f}^{\mathbf{A}_n}}(\tilde{\boldsymbol{\psi}}_{\mathcal{M}}, \boldsymbol{\psi}^*_{\mathcal{M}^{\mathrm{c}}})\}^{-1}\bar{\mathbf{f}}^{\mathbf{A}}(\boldsymbol{\psi}_0) \\
&+ O_{\mathrm{p}}(m^{3/2}n^{-1/2+1/\gamma}) + O_{\mathrm{p}}(m^{5/2}n^{-1/2}) + o_{\mathrm{p}}(1)\,.
\end{aligned}$$

Let $\mathbf{J} = ([\mathbb{E}\{\nabla_{\boldsymbol{\psi}_{\mathcal{M}}}\mathbf{f}_i^{\mathbf{A}}(\boldsymbol{\psi}_0)\}]^{\mathrm{T}}\{\mathbf{V}_{\mathbf{f}^{\mathbf{A}}}(\boldsymbol{\psi}_0)\}^{-1/2})^{\otimes 2}$ with $\mathbf{V}_{\mathbf{f}^{\mathbf{A}}}(\boldsymbol{\psi}_0) = \mathbb{E}\{\mathbf{f}_i^{\mathbf{A}}(\boldsymbol{\psi}_0)^{\otimes 2}\}$. If $m^{5/2}n^{-1/2} = o(1)$ and $m^{3/2}(\omega_n + \varpi_{2,n}) = o(1)$, then by similar arguments in the proof of Lemma 4 of Chang et al. (2018), we have

$$\begin{aligned}
n^{1/2}\boldsymbol{\alpha}^{\mathrm{T}}(\hat{\mathbf{J}}^*)^{1/2}(\tilde{\boldsymbol{\psi}}_{\mathcal{M}} - \boldsymbol{\psi}_{0,\mathcal{M}}) =& -n^{1/2}\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{J}^{-1/2}[\mathbb{E}\{\nabla_{\boldsymbol{\psi}_{\mathcal{M}}}\mathbf{f}_i^{\mathbf{A}}(\boldsymbol{\psi}_0)\}]^{\mathrm{T}}\{\mathbf{V}_{\mathbf{f}^{\mathbf{A}}}(\boldsymbol{\psi}_0)\}^{-1}\bar{\mathbf{f}}^{\mathbf{A}}(\boldsymbol{\psi}_0) \\
&+ O_{\mathrm{p}}(m^{3/2}n^{-1/2+1/\gamma}) + O_{\mathrm{p}}(m^{5/2}n^{-1/2}) \\
&+ o_{\mathrm{p}}(1) + O_{\mathrm{p}}\{m^{3/2}(\omega_n + \varpi_{2,n})\} \\
\xrightarrow{d}& \mathcal{N}(0,1)\,.
\end{aligned}$$

We complete the proof. □

A.7 Proofs of auxiliary lemmas

A.7.1 Proof of Lemma A.2

To simplify the notation, we write $\mathcal{M}_{\boldsymbol{\psi}_n}(c)$ and $\mathcal{D}_{\boldsymbol{\psi}_n}(c)$ as $\mathcal{M}_{\boldsymbol{\psi}_n}$ and $\mathcal{D}_{\boldsymbol{\psi}_n}$, respectively. Due to the convexity of $P_{2,\nu}(\cdot)$, we know that $f(\boldsymbol{\lambda}; \boldsymbol{\psi}_n)$ is a concave function w.r.t $\boldsymbol{\lambda}$. We only need to show there exists a sparse local maximizer $\hat{\boldsymbol{\lambda}}(\boldsymbol{\psi}_n)$ satisfying the three results. By the definition of $\mathcal{M}_{\boldsymbol{\psi}_n}$ and $\mathcal{M}^*_{\boldsymbol{\psi}_n}$, we have $\mathcal{M}_{\boldsymbol{\psi}_n} \subset \mathcal{M}^*_{\boldsymbol{\psi}_n}$ which implies $|\mathcal{M}_{\boldsymbol{\psi}_n}| \leq m_n$ w.p.a.1. Notice that

$m_n^{1/2} u_n n^{1/\gamma} = o(1)$. Given $\mathcal{M}_{\boldsymbol{\psi}_n}$, we select δ_n satisfying $\delta_n = o(m_n^{-1/2} n^{-1/\gamma})$ and $u_n = o(\delta_n)$. Let $\bar{\boldsymbol{\lambda}}_n = \arg\max_{\boldsymbol{\lambda} \in \Lambda_n} f(\boldsymbol{\lambda}; \boldsymbol{\psi}_n)$ where $\Lambda_n = \{\boldsymbol{\lambda} = (\boldsymbol{\lambda}_{\mathcal{M}_{\boldsymbol{\psi}_n}}^{\mathrm{T}}, \boldsymbol{\lambda}_{\mathcal{M}_{\boldsymbol{\psi}_n}^{\mathrm{c}}}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^r : |\boldsymbol{\lambda}_{\mathcal{M}_{\boldsymbol{\psi}_n}}|_2 \le \delta_n, \boldsymbol{\lambda}_{\mathcal{M}_{\boldsymbol{\psi}_n}^{\mathrm{c}}} = \mathbf{0}\}$. It follows from $\max_{j \in \mathcal{T}} n^{-1} \sum_{i=1}^n |g_{i,j}^{(\mathcal{T})}(\boldsymbol{\psi}_n)|^\gamma = O_{\mathrm{p}}(1)$ that $\max_{i \in [n]} |g_{i,j}^{(\mathcal{T})}(\boldsymbol{\psi}_n)| = O_{\mathrm{p}}(n^{1/\gamma})$ holds uniformly over $j \in \mathcal{T}$, which implies that $\max_{i \in [n]} |\mathbf{g}_{i,\mathcal{M}_{\boldsymbol{\psi}_n}}^{(\mathcal{T})}(\boldsymbol{\psi}_n)|_2 = O_{\mathrm{p}}(m_n^{1/2} n^{1/\gamma})$. Thus, $\max_{i \in [n]} |\bar{\boldsymbol{\lambda}}_n^{\mathrm{T}} \mathbf{g}_i^{(\mathcal{T})}(\boldsymbol{\psi}_n)| = o_{\mathrm{p}}(1)$. By the Taylor expansion, we have

$$\begin{aligned} 0 = f(\mathbf{0}; \boldsymbol{\psi}_n) &\le f(\bar{\boldsymbol{\lambda}}_n; \boldsymbol{\psi}_n) \\ &= \frac{1}{n} \sum_{i=1}^n \bar{\boldsymbol{\lambda}}_n^{\mathrm{T}} \mathbf{g}_i^{(\mathcal{T})}(\boldsymbol{\psi}_n) - \frac{1}{2n} \sum_{i=1}^n \frac{\bar{\boldsymbol{\lambda}}_n^{\mathrm{T}} \mathbf{g}_i^{(\mathcal{T})}(\boldsymbol{\psi}_n)^{\otimes 2} \bar{\boldsymbol{\lambda}}_n}{\{1 + \bar{c} \bar{\boldsymbol{\lambda}}_n^{\mathrm{T}} \mathbf{g}_i^{(\mathcal{T})}(\boldsymbol{\psi}_n)\}^2} - \sum_{j \in \mathcal{D}} P_{2,\nu}(|\bar{\lambda}_{n,j}|)\,, \end{aligned} \tag{A.28}$$

where $\bar{\boldsymbol{\lambda}}_n = (\bar{\lambda}_{n,1}, \ldots, \bar{\lambda}_{n,r})^{\mathrm{T}}$ and $\bar{c} \in (0,1)$. Recall $P_{2,\nu}(t) = \nu \rho_2(t; \nu)$. By the convexity of $P_{2,\nu}(\cdot)$, we have $\rho_2'(t;\nu) \ge \rho_2'(0^+)$ for any $t > 0$. Notice that $\lambda_{\min}\{\widehat{\mathbf{V}}_{\mathcal{M}_{\boldsymbol{\psi}_n}}^{(\mathcal{T})}(\boldsymbol{\psi}_n)\}$ is uniformly bounded away from zero w.p.a.1, and $|\bar{\lambda}_{n,j}| \ge \bar{\lambda}_{n,j} \cdot \mathrm{sgn}\{\bar{g}_j^{(\mathcal{T})}(\boldsymbol{\psi}_n)\}$. Thus, (A.28) leads to

$$0 \le \bar{\boldsymbol{\lambda}}_{n,\mathcal{M}_{\boldsymbol{\psi}_n}}^{\mathrm{T}} \begin{pmatrix} \bar{\mathbf{g}}^{(\mathcal{I})}(\boldsymbol{\psi}_n) \\ \bar{\mathbf{g}}_{\mathcal{D}_{\boldsymbol{\psi}_n}}^{(\mathcal{T})}(\boldsymbol{\psi}_n) - \nu \rho_2'(0^+) \mathrm{sgn}\{\bar{\mathbf{g}}_{\mathcal{D}_{\boldsymbol{\psi}_n}}^{(\mathcal{T})}(\boldsymbol{\psi}_n)\} \end{pmatrix} - C |\bar{\boldsymbol{\lambda}}_{n,\mathcal{M}_{\boldsymbol{\psi}_n}}|_2^2 \{1 + o_{\mathrm{p}}(1)\}\,.$$

Due to $|\bar{\mathbf{g}}^{(\mathcal{I})}(\boldsymbol{\psi}_n)|_2^2 + |\bar{\mathbf{g}}_{\mathcal{D}_{\boldsymbol{\psi}_n}}^{(\mathcal{T})}(\boldsymbol{\psi}_n) - \nu \rho_2'(0^+) \mathrm{sgn}\{\bar{\mathbf{g}}_{\mathcal{D}_{\boldsymbol{\psi}_n}}^{(\mathcal{T})}(\boldsymbol{\psi}_n)\}|_2^2 = O_{\mathrm{p}}(u_n^2)$, it holds that $|\bar{\boldsymbol{\lambda}}_{n,\mathcal{M}_{\boldsymbol{\psi}_n}}|_2 = O_{\mathrm{p}}(u_n) = o_{\mathrm{p}}(\delta_n)$. Recall $\mathcal{M}_{\boldsymbol{\psi}_n} = \mathcal{I} \cup \mathcal{D}_{\boldsymbol{\psi}_n}$. Then $|\bar{\boldsymbol{\lambda}}_n|_2 = |\bar{\boldsymbol{\lambda}}_{n,\mathcal{M}_{\boldsymbol{\psi}_n}}|_2 = O_{\mathrm{p}}(u_n)$ and $\{j \in \mathcal{D} : \bar{\lambda}_{n,j} \ne 0\} \subset \mathcal{D}_{\boldsymbol{\psi}_n}$. Write $\bar{\boldsymbol{\lambda}}_{n,\mathcal{D}_{\boldsymbol{\psi}_n}} = (\bar{\lambda}_1, \ldots, \bar{\lambda}_{|\mathcal{D}_{\boldsymbol{\psi}_n}|})^{\mathrm{T}}$. We have w.p.a.1 that

$$\mathbf{0} = \frac{1}{n} \sum_{i=1}^n \frac{\mathbf{g}_{i,\mathcal{D}_{\boldsymbol{\psi}_n}}^{(\mathcal{T})}(\boldsymbol{\psi}_n)}{1 + \bar{\boldsymbol{\lambda}}_{n,\mathcal{M}_{\boldsymbol{\psi}_n}}^{\mathrm{T}} \mathbf{g}_{i,\mathcal{M}_{\boldsymbol{\psi}_n}}^{(\mathcal{T})}(\boldsymbol{\psi}_n)} - \hat{\boldsymbol{\eta}}\,, \tag{A.29}$$

where $\hat{\boldsymbol{\eta}} = (\hat{\eta}_1, \ldots, \hat{\eta}_{|\mathcal{D}_{\boldsymbol{\psi}_n}|})^{\mathrm{T}}$ with $\hat{\eta}_j = \nu \rho_2'(|\bar{\lambda}_j|; \nu) \mathrm{sgn}(\bar{\lambda}_j)$ for $\bar{\lambda}_j \ne 0$ and $\hat{\eta}_j \in [-\nu\rho_2'(0^+), \nu\rho_2'(0^+)]$ for $\bar{\lambda}_j = 0$. It follows from (A.29) that $\hat{\boldsymbol{\eta}} = \bar{\mathbf{g}}_{\mathcal{D}_{\boldsymbol{\psi}_n}}^{(\mathcal{T})}(\boldsymbol{\psi}_n) + \mathbf{R}$ with

$$\begin{aligned} |\mathbf{R}|_\infty^2 &= \left| \frac{1}{n} \sum_{i=1}^n \frac{\bar{\boldsymbol{\lambda}}_{n,\mathcal{M}_{\boldsymbol{\psi}_n}}^{\mathrm{T}} \mathbf{g}_{i,\mathcal{M}_{\boldsymbol{\psi}_n}}^{(\mathcal{T})}(\boldsymbol{\psi}_n) \mathbf{g}_{i,\mathcal{D}_{\boldsymbol{\psi}_n}}^{(\mathcal{T})}(\boldsymbol{\psi}_n)}{1 + \bar{\boldsymbol{\lambda}}_{n,\mathcal{M}_{\boldsymbol{\psi}_n}}^{\mathrm{T}} \mathbf{g}_{i,\mathcal{M}_{\boldsymbol{\psi}_n}}^{(\mathcal{T})}(\boldsymbol{\psi}_n)} \right|_\infty^2 \\ &\le \max_{j \in \mathcal{D}_{\boldsymbol{\psi}_n}} \left\{ \frac{1}{n} \sum_{i=1}^n |\bar{\boldsymbol{\lambda}}_{n,\mathcal{M}_{\boldsymbol{\psi}_n}}^{\mathrm{T}} \mathbf{g}_{i,\mathcal{M}_{\boldsymbol{\psi}_n}}^{(\mathcal{T})}(\boldsymbol{\psi}_n)| |g_{i,j}^{(\mathcal{T})}(\boldsymbol{\psi}_n)| \right\}^2 \cdot \{1 + o_{\mathrm{p}}(1)\} \\ &\le \max_{j \in \mathcal{D}_{\boldsymbol{\psi}_n}} \{\bar{\boldsymbol{\lambda}}_{n,\mathcal{M}_{\boldsymbol{\psi}_n}}^{\mathrm{T}} \widehat{\mathbf{V}}_{\mathcal{M}_{\boldsymbol{\psi}_n}}^{(\mathcal{T})}(\boldsymbol{\psi}_n) \bar{\boldsymbol{\lambda}}_{n,\mathcal{M}_{\boldsymbol{\psi}_n}}\} \left\{ \frac{1}{n} \sum_{i=1}^n |g_{i,j}^{(\mathcal{T})}(\boldsymbol{\psi}_n)|^2 \right\} \cdot \{1 + o_{\mathrm{p}}(1)\} \\ &= O_{\mathrm{p}}(|\bar{\boldsymbol{\lambda}}_{n,\mathcal{M}_{\boldsymbol{\psi}_n}}|_2^2)\,, \end{aligned}$$

which indicates that $|\mathbf{R}|_\infty = O_{\mathrm{p}}(u_n) = o_{\mathrm{p}}(\nu)$. Hence, w.p.a.1 we have $\mathrm{sgn}(\bar{\lambda}_{n,j}) = \mathrm{sgn}\{\bar{g}_j^{(\mathcal{T})}(\boldsymbol{\psi}_n)\}$ for any $j \in \mathcal{D}_{\boldsymbol{\psi}_n}$ with $\bar{\lambda}_{n,j} \neq 0$. To complete the proof, we need to show that $\bar{\boldsymbol{\lambda}}_n$ is a local maximizer of $f(\boldsymbol{\lambda}; \boldsymbol{\psi}_n)$ w.p.a.1. Our proof includes two steps.

Step 1. Define $\Lambda_n^* = \{\boldsymbol{\lambda} = (\boldsymbol{\lambda}_{\mathcal{M}_{\boldsymbol{\psi}_n}^*}^{\mathrm{T}}, \boldsymbol{\lambda}_{\mathcal{M}_{\boldsymbol{\psi}_n}^{*,\mathrm{c}}}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^r : |\boldsymbol{\lambda}_{\mathcal{M}_{\boldsymbol{\psi}_n}^*}|_2 \leq \varepsilon, \boldsymbol{\lambda}_{\mathcal{M}_{\boldsymbol{\psi}_n}^{*,\mathrm{c}}} = \mathbf{0}\}$ for some sufficiently small $\varepsilon > 0$, where $\mathcal{M}_{\boldsymbol{\psi}_n}^* = \mathcal{I} \cup \mathcal{D}_{\boldsymbol{\psi}_n}^*$ with $\mathcal{D}_{\boldsymbol{\psi}_n}^* = \{j \in \mathcal{D} : |\bar{g}_j^{(\mathcal{T})}(\boldsymbol{\psi}_n)| \geq C_* \nu \rho_2'(0^+)\}$ for some constant $C_* \in (0, 1)$. For $\bar{\boldsymbol{\lambda}}_n$ defined before, we will show in this step that $\bar{\boldsymbol{\lambda}}_n = \arg\max_{\boldsymbol{\lambda} \in \Lambda_n^*} f(\boldsymbol{\lambda}; \boldsymbol{\psi}_n)$ w.p.a.1. Due to $\bar{\boldsymbol{\lambda}}_n \in \Lambda_n$ and $\mathcal{M}_{\boldsymbol{\psi}_n} \subset \mathcal{M}_{\boldsymbol{\psi}_n}^*$, we know $\bar{\boldsymbol{\lambda}}_n \in \Lambda_n^*$ w.p.a.1. Restricted on $\boldsymbol{\lambda} \in \Lambda_n^*$, by the concavity of $f(\boldsymbol{\lambda}; \boldsymbol{\psi}_n)$ w.r.t $\boldsymbol{\lambda}_{\mathcal{M}_{\boldsymbol{\psi}_n}^*}$, it suffices to show that w.p.a.1 for any $j \in \mathcal{M}_{\boldsymbol{\psi}_n}^*$ it holds that

$$\frac{\partial f(\bar{\boldsymbol{\lambda}}_n; \boldsymbol{\psi}_n)}{\partial \lambda_j} = 0\,. \tag{A.30}$$

Due to $\bar{\boldsymbol{\lambda}}_n \in \Lambda_n$ and $|\bar{\boldsymbol{\lambda}}_n|_2 = o_{\mathrm{p}}(\delta_n)$, then $\bar{\boldsymbol{\lambda}}_{n,\mathcal{M}_{\boldsymbol{\psi}_n}}$ is an interior point of the set $\{\boldsymbol{\lambda}_{\mathcal{M}_{\boldsymbol{\psi}_n}} \in \mathbb{R}^{|\mathcal{M}_{\boldsymbol{\psi}_n}|} : |\boldsymbol{\lambda}_{\mathcal{M}_{\boldsymbol{\psi}_n}}|_2 \leq \delta_n\}$. Restricted on $\boldsymbol{\lambda} \in \Lambda_n$, we know $f(\boldsymbol{\lambda}; \boldsymbol{\psi}_n)$ is concave w.r.t $\boldsymbol{\lambda}_{\mathcal{M}_{\boldsymbol{\psi}_n}}$. Notice that $\bar{\boldsymbol{\lambda}}_n = \arg\max_{\boldsymbol{\lambda} \in \Lambda_n} f(\boldsymbol{\lambda}; \boldsymbol{\psi}_n)$. Therefore, (A.30) holds for any $j \in \mathcal{M}_{\boldsymbol{\psi}_n}$. Recall $\bar{\boldsymbol{\lambda}}_n = (\bar{\lambda}_{n,1}, \ldots, \bar{\lambda}_{n,r})^{\mathrm{T}}$. For any $j \in \mathcal{M}_{\boldsymbol{\psi}_n}^* \backslash \mathcal{M}_{\boldsymbol{\psi}_n}$, we have $\bar{\lambda}_{n,j} = 0$ and

$$\frac{1}{n}\sum_{i=1}^n \frac{g_{i,j}^{(\mathcal{T})}(\boldsymbol{\psi}_n)}{1 + \bar{\boldsymbol{\lambda}}_{n,\mathcal{M}_{\boldsymbol{\psi}_n}^*}^{\mathrm{T}} \mathbf{g}_{i,\mathcal{M}_{\boldsymbol{\psi}_n}^*}^{(\mathcal{T})}(\boldsymbol{\psi}_n)} = \bar{g}_j^{(\mathcal{T})}(\boldsymbol{\psi}_n) + O_{\mathrm{p}}(u_n)\,,$$

where the term $O_{\mathrm{p}}(u_n) = o_{\mathrm{p}}(\nu)$ is uniform over $j \in \mathcal{M}_{\boldsymbol{\psi}_n}^* \backslash \mathcal{M}_{\boldsymbol{\psi}_n}$. Such conclusion can be obtained by the same arguments for deriving the convergence rate of $|\mathbf{R}|_\infty$ stated above. By the definition of $\mathcal{M}_{\boldsymbol{\psi}_n}^*$ and $\mathcal{M}_{\boldsymbol{\psi}_n}$, we know $\mathcal{M}_{\boldsymbol{\psi}_n}^* \backslash \mathcal{M}_{\boldsymbol{\psi}_n} = \mathcal{D}_{\boldsymbol{\psi}_n}^* \backslash \mathcal{D}_{\boldsymbol{\psi}_n}$. Then $C_* \nu \rho_2'(0^+) \leq |\bar{g}_j^{(\mathcal{T})}(\boldsymbol{\psi}_n)| < c\nu\rho_2'(0^+)$ for any $j \in \mathcal{M}_{\boldsymbol{\psi}_n}^* \backslash \mathcal{M}_{\boldsymbol{\psi}_n}$. Hence, we have w.p.a.1 that

$$\left| \frac{1}{n}\sum_{i=1}^n \frac{g_{i,j}^{(\mathcal{T})}(\boldsymbol{\psi}_n)}{1 + \bar{\boldsymbol{\lambda}}_{n,\mathcal{M}_{\boldsymbol{\psi}_n}^*}^{\mathrm{T}} \mathbf{g}_{i,\mathcal{M}_{\boldsymbol{\psi}_n}^*}^{(\mathcal{T})}(\boldsymbol{\psi}_n)} \right| \leq \nu \rho_2'(0^+)\,,$$

which implies that there exists some $\hat{\eta}_j^* \in [-\nu\rho_2'(0^+), \nu\rho_2'(0^+)]$ such that

$$0 = \frac{1}{n}\sum_{i=1}^n \frac{g_{i,j}^{(\mathcal{T})}(\boldsymbol{\psi}_n)}{1 + \bar{\boldsymbol{\lambda}}_{n,\mathcal{M}_{\boldsymbol{\psi}_n}^*}^{\mathrm{T}} \mathbf{g}_{i,\mathcal{M}_{\boldsymbol{\psi}_n}^*}^{(\mathcal{T})}(\boldsymbol{\psi}_n)} - \hat{\eta}_j^*\,.$$

Hence, (A.30) holds for any $j \in \mathcal{M}_{\boldsymbol{\psi}_n}^* \backslash \mathcal{M}_{\boldsymbol{\psi}_n}$. Then we have $\bar{\boldsymbol{\lambda}}_n = \arg\max_{\boldsymbol{\lambda} \in \Lambda_n^*} f(\boldsymbol{\lambda}; \boldsymbol{\psi}_n)$ w.p.a.1.

Step 2. Define $\tilde{\Lambda}_n = \{\boldsymbol{\lambda} = (\boldsymbol{\lambda}_{\mathcal{M}_{\boldsymbol{\psi}_n}^*}^{\mathrm{T}}, \boldsymbol{\lambda}_{\mathcal{M}_{\boldsymbol{\psi}_n}^{*,\mathrm{c}}}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^r : |\boldsymbol{\lambda}_{\mathcal{M}_{\boldsymbol{\psi}_n}^*} - \bar{\boldsymbol{\lambda}}_{n,\mathcal{M}_{\boldsymbol{\psi}_n}^*}|_2 \leq o(u_n), |\boldsymbol{\lambda}_{\mathcal{M}_{\boldsymbol{\psi}_n}^{*,\mathrm{c}}}|_1 \leq \min\{O(m_n^{1/2} u_n), o(r_2^{-1/\gamma} n^{-1/\gamma})\}\}$. We will prove in this step that $\bar{\boldsymbol{\lambda}}_n$ is the maximizer of $f(\boldsymbol{\lambda}; \boldsymbol{\psi}_n)$ over $\boldsymbol{\lambda} \in \tilde{\Lambda}_n$. Notice that $\max_{i \in [n], \boldsymbol{\lambda} \in \tilde{\Lambda}_n} |\boldsymbol{\lambda}^{\mathrm{T}} \mathbf{g}_i^{(\mathcal{T})}(\boldsymbol{\psi}_n)| = o_{\mathrm{p}}(1)$. For any $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_r)^{\mathrm{T}} \in \tilde{\Lambda}_n$,

denote by $\tilde{\boldsymbol{\lambda}} = (\boldsymbol{\lambda}_{\mathcal{M}^*_{\boldsymbol{\psi}_n}}^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}}$ the projection of $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_{\mathcal{M}^*_{\boldsymbol{\psi}_n}}^{\mathrm{T}}, \boldsymbol{\lambda}_{\mathcal{M}^{*,\mathrm{c}}_{\boldsymbol{\psi}_n}}^{\mathrm{T}})^{\mathrm{T}}$ onto Λ_n^* for Λ_n^* defined in Step 1. Then it holds that

$$\sup_{\boldsymbol{\lambda}\in\tilde{\Lambda}_n}\{f(\boldsymbol{\lambda};\boldsymbol{\psi}_n) - f(\tilde{\boldsymbol{\lambda}};\boldsymbol{\psi}_n)\} = \sup_{\boldsymbol{\lambda}\in\tilde{\Lambda}_n}\left\{\frac{1}{n}\sum_{i=1}^n \frac{\mathbf{g}_i^{(\mathcal{T})}(\boldsymbol{\psi}_n)^{\mathrm{T}}(\boldsymbol{\lambda}-\tilde{\boldsymbol{\lambda}})}{1+\boldsymbol{\lambda}_*^{\mathrm{T}}\mathbf{g}_i^{(\mathcal{T})}(\boldsymbol{\psi}_n)} - \sum_{j\in\mathcal{M}^{*,\mathrm{c}}_{\boldsymbol{\psi}_n}} P_{2,\nu}(|\lambda_j|)\right\},$$

where $\boldsymbol{\lambda}_*$ is on the jointing line between $\boldsymbol{\lambda}$ and $\tilde{\boldsymbol{\lambda}}$. It follows from the Taylor expansion that

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \frac{\mathbf{g}_i^{(\mathcal{T})}(\boldsymbol{\psi}_n)^{\mathrm{T}}(\boldsymbol{\lambda}-\tilde{\boldsymbol{\lambda}})}{1+\boldsymbol{\lambda}_*^{\mathrm{T}}\mathbf{g}_i^{(\mathcal{T})}(\boldsymbol{\psi}_n)} - \sum_{j\in\mathcal{M}^{*,\mathrm{c}}_{\boldsymbol{\psi}_n}} P_{2,\nu}(|\lambda_j|)\\
&= \boldsymbol{\lambda}_{\mathcal{M}^{*,\mathrm{c}}_{\boldsymbol{\psi}_n}}^{\mathrm{T}}\bar{\mathbf{g}}^{(\mathcal{T})}_{\mathcal{M}^{*,\mathrm{c}}_{\boldsymbol{\psi}_n}}(\boldsymbol{\psi}_n) - \left\{\frac{1}{n}\sum_{i=1}^n \boldsymbol{\lambda}_*^{\mathrm{T}}\mathbf{g}_i^{(\mathcal{T})}(\boldsymbol{\psi}_n)\mathbf{g}^{(\mathcal{T})}_{i,\mathcal{M}^{*,\mathrm{c}}_{\boldsymbol{\psi}_n}}(\boldsymbol{\psi}_n)^{\mathrm{T}}\boldsymbol{\lambda}_{\mathcal{M}^{*,\mathrm{c}}_{\boldsymbol{\psi}_n}}\right\}\{1+o_{\mathrm{p}}(1)\} - \sum_{j\in\mathcal{M}^{*,\mathrm{c}}_{\boldsymbol{\psi}_n}} P_{2,\nu}(|\lambda_j|)\\
&\le \big|\bar{\mathbf{g}}^{(\mathcal{T})}_{\mathcal{M}^{*,\mathrm{c}}_{\boldsymbol{\psi}_n}}(\boldsymbol{\psi}_n)\big|_\infty\big|\boldsymbol{\lambda}_{\mathcal{M}^{*,\mathrm{c}}_{\boldsymbol{\psi}_n}}\big|_1 + \frac{1}{n}\sum_{i=1}^n\sum_{j\in\mathcal{T}}\sum_{k\in\mathcal{M}^{*,\mathrm{c}}_{\boldsymbol{\psi}_n}}|\lambda_{*,j}g_{i,j}^{(\mathcal{T})}(\boldsymbol{\psi}_n)\lambda_k g_{i,k}^{(\mathcal{T})}(\boldsymbol{\psi}_n)|\{1+o_{\mathrm{p}}(1)\}\\
&\qquad - \nu\rho_2'(0^+)\sum_{j\in\mathcal{M}^{*,\mathrm{c}}_{\boldsymbol{\psi}_n}}|\lambda_j|\\
&\le C_*\nu\rho_2'(0^+)\sum_{j\in\mathcal{M}^{*,\mathrm{c}}_{\boldsymbol{\psi}_n}}|\lambda_j| + \max_{j\in\mathcal{T}}\left\{\frac{1}{n}\sum_{i=1}^n|g_{i,j}^{(\mathcal{T})}(\boldsymbol{\psi}_n)|^2\right\}\Big(\sum_{k\in\mathcal{M}^{*,\mathrm{c}}_{\boldsymbol{\psi}_n}}|\lambda_k|\Big)|\boldsymbol{\lambda}_*|_1\{1+o_{\mathrm{p}}(1)\}\\
&\qquad - \nu\rho_2'(0^+)\sum_{j\in\mathcal{M}^{*,\mathrm{c}}_{\boldsymbol{\psi}_n}}|\lambda_j|\\
&\le \big\{-(1-C_*)\nu\rho_2'(0^+) + O_{\mathrm{p}}(m_n^{1/2}u_n)\big\}\sum_{j\in\mathcal{M}^{*,\mathrm{c}}_{\boldsymbol{\psi}_n}}|\lambda_j|\,,
\end{aligned}$$

where the term $O_{\mathrm{p}}(m_n^{1/2}u_n)$ holds uniformly over $\boldsymbol{\lambda}\in\tilde{\Lambda}_n$. Since $m_n^{1/2}u_n = o(\nu)$, then $-(1-C_*)\nu\rho_2'(0^+) + O_{\mathrm{p}}(m_n^{1/2}u_n) < 0$ w.p.a.1. Thus,

$$\mathbb{P}\bigg[\sup_{\boldsymbol{\lambda}\in\tilde{\Lambda}_n}\big\{f(\boldsymbol{\lambda};\boldsymbol{\psi}_n) - f(\tilde{\boldsymbol{\lambda}};\boldsymbol{\psi}_n)\big\} \le 0\bigg] \to 1\,.$$

Hence, $\bar{\boldsymbol{\lambda}}_n$ is a local maximizer of $f(\boldsymbol{\lambda};\boldsymbol{\psi}_n)$ w.p.a.1. We complete the proof of Lemma A.2. □

A.7.2 Proof of Lemma A.3

To simplify the notation, we write $\mathcal{M}_{\boldsymbol{\psi}_0}(c)$ and $\mathcal{D}_{\boldsymbol{\psi}_0}(c)$ as $\mathcal{M}_{\boldsymbol{\psi}_0}$ and $\mathcal{D}_{\boldsymbol{\psi}_0}$, respectively. Recall $\mathcal{M}^*_{\boldsymbol{\psi}_0} = \mathcal{I}\cup\mathcal{D}^*_{\boldsymbol{\psi}_0}$ with $|\mathcal{I}| = r_1$ and $\mathcal{D}^*_{\boldsymbol{\psi}_0} = \{j\in\mathcal{D} : |\bar{g}_j^{(\mathcal{T})}(\boldsymbol{\psi}_0)| \ge C_*\nu\rho_2'(0^+)\}$. Due to $\max_{j\in\mathcal{T}} n^{-1}\sum_{i=1}^n|g_{i,j}^{(\mathcal{T})}(\boldsymbol{\psi}_0)|^2 = O_{\mathrm{p}}(1)$, by the moderate deviation of self-normalized sums (Jing

et al. 2003), it holds that $|\bar{\mathbf{g}}^{(\mathcal{T})}(\boldsymbol{\psi}_0)|_\infty = O_{\mathrm{p}}(\aleph_n)$ provided that $\log r = o(n^{1/3})$. Since $\nu \gg \aleph_n$, we know $\mathbb{P}(\mathcal{D}^*_{\boldsymbol{\psi}_0} = \emptyset) \to 1$ which implies $|\mathcal{M}^*_{\boldsymbol{\psi}_0}| \le 2r_1$ w.p.a.1. Pick δ_n satisfying $\delta_n = o(r_1^{-1/2} n^{-1/\gamma})$ and $r_1^{1/2}\aleph_n = o(\delta_n)$ which can be guaranteed by $r_1\aleph_n = o(n^{-1/\gamma})$. Recall $\mathcal{M}_{\boldsymbol{\psi}_0} = \mathcal{I} \cup \mathcal{D}_{\boldsymbol{\psi}_0}$ with $\mathcal{D}_{\boldsymbol{\psi}_0} = \{j \in \mathcal{D} : |\bar{g}^{(\mathcal{T})}_j(\boldsymbol{\psi}_0)| \ge c\nu\rho_2'(0^+)\}$ for some $c \in (C_*, 1)$. Then $|\mathcal{M}_{\boldsymbol{\psi}_0}| \le |\mathcal{M}^*_{\boldsymbol{\psi}_0}| \le 2r_1$ w.p.a.1. Let $\Lambda_0 = \{\boldsymbol{\lambda} = (\boldsymbol{\lambda}^{\mathrm{T}}_{\mathcal{M}_{\boldsymbol{\psi}_0}}, \boldsymbol{\lambda}^{\mathrm{T}}_{\mathcal{M}^{\mathrm{c}}_{\boldsymbol{\psi}_0}})^{\mathrm{T}} \in \mathbb{R}^r : |\boldsymbol{\lambda}_{\mathcal{M}_{\boldsymbol{\psi}_0}}|_2 \le \delta_n, \boldsymbol{\lambda}_{\mathcal{M}^{\mathrm{c}}_{\boldsymbol{\psi}_0}} = \mathbf{0}\}$ and $\bar{\boldsymbol{\lambda}}_0 = \arg\max_{\boldsymbol{\lambda}\in\Lambda_0} f(\boldsymbol{\lambda}; \boldsymbol{\psi}_0)$. Due to $\max_{i\in[n]} |g^{(\mathcal{T})}_{i,j}(\boldsymbol{\psi}_0)| = O_{\mathrm{p}}(n^{1/\gamma})$ holds uniformly over $j \in \mathcal{T}$, we have $\max_{i\in[n]} |\bar{\boldsymbol{\lambda}}_0^{\mathrm{T}} \mathbf{g}^{(\mathcal{T})}_i(\boldsymbol{\psi}_0)| = o_{\mathrm{p}}(1)$. Write $\bar{\boldsymbol{\lambda}}_0 = (\bar{\lambda}_{0,1}, \ldots, \bar{\lambda}_{0,r})^{\mathrm{T}}$. Similar to (A.28), we have

$$
\begin{aligned}
0 = f(\mathbf{0}; \boldsymbol{\psi}_0) &\le f(\bar{\boldsymbol{\lambda}}_0; \boldsymbol{\psi}_0) \\
&= \frac{1}{n}\sum_{i=1}^n \bar{\boldsymbol{\lambda}}_0^{\mathrm{T}} \mathbf{g}^{(\mathcal{T})}_i(\boldsymbol{\psi}_0) - \frac{1}{2n}\sum_{i=1}^n \frac{\bar{\boldsymbol{\lambda}}_0^{\mathrm{T}} \mathbf{g}^{(\mathcal{T})}_i(\boldsymbol{\psi}_0)^{\otimes 2} \bar{\boldsymbol{\lambda}}_0}{\{1 + \bar{c}\bar{\boldsymbol{\lambda}}_0^{\mathrm{T}} \mathbf{g}^{(\mathcal{T})}_i(\boldsymbol{\psi}_0)\}^2} - \sum_{j\in\mathcal{D}} P_{2,\nu}(|\bar{\lambda}_{0,j}|) \\
&\le \bar{\boldsymbol{\lambda}}^{\mathrm{T}}_{0,\mathcal{M}_{\boldsymbol{\psi}_0}} \bar{\mathbf{g}}^{(\mathcal{T})}_{\mathcal{M}_{\boldsymbol{\psi}_0}}(\boldsymbol{\psi}_0) - \frac{1}{2}\lambda_{\min}\{\widehat{\mathbf{V}}^{(\mathcal{T})}_{\mathcal{M}_{\boldsymbol{\psi}_0}}(\boldsymbol{\psi}_0)\}|\bar{\boldsymbol{\lambda}}_{0,\mathcal{M}_{\boldsymbol{\psi}_0}}|_2^2\{1 + o_{\mathrm{p}}(1)\} \\
&\le \bar{\boldsymbol{\lambda}}^{\mathrm{T}}_{0,\mathcal{M}_{\boldsymbol{\psi}_0}} \bar{\mathbf{g}}^{(\mathcal{T})}_{\mathcal{M}_{\boldsymbol{\psi}_0}}(\boldsymbol{\psi}_0) - C|\bar{\boldsymbol{\lambda}}_{0,\mathcal{M}_{\boldsymbol{\psi}_0}}|_2^2\{1 + o_{\mathrm{p}}(1)\}
\end{aligned}
$$

for some $\bar{c} \in (0, 1)$. Due to $|\bar{\mathbf{g}}^{(\mathcal{T})}(\boldsymbol{\psi}_0)|_\infty = O_{\mathrm{p}}(\aleph_n)$, we have $|\bar{\mathbf{g}}^{(\mathcal{T})}_{\mathcal{M}_{\boldsymbol{\psi}_0}}(\boldsymbol{\psi}_0)|_2 = O_{\mathrm{p}}(r_1^{1/2}\aleph_n)$. Then $|\bar{\boldsymbol{\lambda}}_{0,\mathcal{M}_{\boldsymbol{\psi}_0}}|_2 = O_{\mathrm{p}}(r_1^{1/2}\aleph_n) = o_{\mathrm{p}}(\delta_n)$. Using the same arguments to prove $\bar{\boldsymbol{\lambda}}_n$ is a local maximizer of $f(\boldsymbol{\lambda}; \boldsymbol{\psi}_n)$ w.p.a.1 in the proof of Lemma A.2, we can also show such defined $\bar{\boldsymbol{\lambda}}_0$ is a local maximizer of $f(\boldsymbol{\lambda}; \boldsymbol{\psi}_0)$ w.p.a.1. Notice that $f(\boldsymbol{\lambda}; \boldsymbol{\psi}_0)$ is a concave function w.r.t $\boldsymbol{\lambda}$. We complete the proof. □

A.7.3 Proof of Lemma A.5

Recall $f(\boldsymbol{\lambda}; \boldsymbol{\psi}) = n^{-1}\sum_{i=1}^n \log\{1 + \boldsymbol{\lambda}^{\mathrm{T}} \mathbf{g}^{(\mathcal{T})}_i(\boldsymbol{\psi})\} - \sum_{j\in\mathcal{D}} P_{2,\nu}(|\lambda_j|)$ for any $\boldsymbol{\psi} \in \boldsymbol{\Psi}$ and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_r)^{\mathrm{T}}$, and $\hat{\boldsymbol{\lambda}}(\boldsymbol{\psi}) = \arg\max_{\boldsymbol{\lambda}\in\hat{\Lambda}^{(\mathcal{T})}_n(\boldsymbol{\psi})} f(\boldsymbol{\lambda}; \boldsymbol{\psi})$. Then $\hat{\boldsymbol{\psi}}_{\mathrm{PEL}}$ and its associated Lagrange multiplier $\hat{\boldsymbol{\lambda}}(\hat{\boldsymbol{\psi}}_{\mathrm{PEL}}) = (\hat{\lambda}_1, \ldots, \hat{\lambda}_r)^{\mathrm{T}}$ satisfy the score equation $\nabla_{\boldsymbol{\lambda}} f\{\hat{\boldsymbol{\lambda}}(\hat{\boldsymbol{\psi}}_{\mathrm{PEL}}); \hat{\boldsymbol{\psi}}_{\mathrm{PEL}}\} = \mathbf{0}$, that is,

$$
\mathbf{0} = \frac{1}{n}\sum_{i=1}^n \frac{\mathbf{g}^{(\mathcal{T})}_i(\hat{\boldsymbol{\psi}}_{\mathrm{PEL}})}{1 + \hat{\boldsymbol{\lambda}}(\hat{\boldsymbol{\psi}}_{\mathrm{PEL}})^{\mathrm{T}} \mathbf{g}^{(\mathcal{T})}_i(\hat{\boldsymbol{\psi}}_{\mathrm{PEL}})} - \hat{\boldsymbol{\eta}}\,,
$$

where $\hat{\boldsymbol{\eta}} = (\hat{\eta}_1, \ldots, \hat{\eta}_r)^{\mathrm{T}}$ with $\hat{\eta}_j = 0$ for $j \in \mathcal{I}$, $\hat{\eta}_j = \nu\rho_2'(|\hat{\lambda}_j|; \nu)\mathrm{sgn}(\hat{\lambda}_j)$ for $j \in \mathcal{D}$ and $\hat{\lambda}_j \ne 0$, and $\hat{\eta}_j \in [-\nu\rho_2'(0^+), \nu\rho_2'(0^+)]$ for $j \in \mathcal{D}$ and $\hat{\lambda}_j = 0$. Recall $\mathcal{R}_n = \mathcal{I} \cup \mathrm{supp}\{\hat{\boldsymbol{\lambda}}_{\mathcal{D}}(\hat{\boldsymbol{\psi}}_{\mathrm{PEL}})\}$. Restricted on $\mathcal{R}_n$, for any $\boldsymbol{\psi} \in \mathbb{R}^{p+r_2}$ and $\boldsymbol{\chi} = (\chi_j)_{j\in\mathcal{R}_n} \in \mathbb{R}^{|\mathcal{R}_n|}$ with $\chi_j \ne 0$ for any $j \notin \mathcal{I}$, define $\mathbf{m}(\boldsymbol{\chi}, \boldsymbol{\psi}) = n^{-1}\sum_{i=1}^n \mathbf{g}^{(\mathcal{T})}_{i,\mathcal{R}_n}(\boldsymbol{\psi})\{1 + \boldsymbol{\chi}^{\mathrm{T}} \mathbf{g}^{(\mathcal{T})}_{i,\mathcal{R}_n}(\boldsymbol{\psi})\}^{-1} - \mathbf{w}$, where $\mathbf{w} = (w_j)_{j\in\mathcal{R}_n}$ with $w_j = 0$ for $j \in \mathcal{I}$ and $w_j = \nu\rho_2'(|\chi_j|; \nu)\mathrm{sgn}(\chi_j)$ otherwise. Then $\hat{\boldsymbol{\lambda}}_{\mathcal{R}_n}(\hat{\boldsymbol{\psi}}_{\mathrm{PEL}})$ and $\hat{\boldsymbol{\psi}}_{\mathrm{PEL}}$ satisfy $\mathbf{m}\{\hat{\boldsymbol{\lambda}}_{\mathcal{R}_n}(\hat{\boldsymbol{\psi}}_{\mathrm{PEL}}), \hat{\boldsymbol{\psi}}_{\mathrm{PEL}}\} = \mathbf{0}$. By the implicit function theorem [Theorem 9.28 of Rudin (1976)], for all $\boldsymbol{\psi}$ in a $|\cdot|_2$-neighbourhood of $\hat{\boldsymbol{\psi}}_{\mathrm{PEL}}$, there is a $\boldsymbol{\chi}(\boldsymbol{\psi})$ such that $\mathbf{m}\{\boldsymbol{\chi}(\boldsymbol{\psi}), \boldsymbol{\psi}\} = \mathbf{0}$, $\boldsymbol{\chi}(\hat{\boldsymbol{\psi}}_{\mathrm{PEL}}) = \hat{\boldsymbol{\lambda}}_{\mathcal{R}_n}(\hat{\boldsymbol{\psi}}_{\mathrm{PEL}})$ and $\boldsymbol{\chi}(\boldsymbol{\psi})$ is continuously

differentiable in $\boldsymbol{\psi}$.

By Condition 6, we know the event $\mathcal{E} = \{\max_{j\in\mathcal{R}_n^{\mathrm{c}}} |\hat{\eta}_j| < \nu\rho_2'(0^+)\}$ holds w.p.a.1. Restricted on $\mathcal{E}$, let $\kappa_n = \nu\rho_2'(0^+) - \max_{j\in\mathcal{R}_n^{\mathrm{c}}} |\hat{\eta}_j|$. Define $\boldsymbol{\Psi}_{**} = \{\boldsymbol{\psi} \in \mathbb{R}^{p+r_2} : |\boldsymbol{\psi} - \hat{\boldsymbol{\psi}}_{\mathrm{PEL}}|_1 \leq o[\min\{\zeta_n, \kappa_n\}], |\boldsymbol{\chi}(\boldsymbol{\psi}) - \boldsymbol{\chi}(\hat{\boldsymbol{\psi}}_{\mathrm{PEL}})|_1 \leq o(\kappa_n), |\boldsymbol{\chi}(\boldsymbol{\psi}) - \boldsymbol{\chi}(\hat{\boldsymbol{\psi}}_{\mathrm{PEL}})|_2 \leq o(\ell_n^{-1/2} n^{-1/\gamma})\}$ for some $\zeta_n > 0$. Since $\chi_j(\hat{\boldsymbol{\psi}}_{\mathrm{PEL}}) \neq 0$ for any $j \in \mathcal{R}_n\backslash\mathcal{I}$ and $\boldsymbol{\chi}(\boldsymbol{\psi})$ is continuously differentiable in $\hat{\boldsymbol{\psi}}_{\mathrm{PEL}}$, we can select sufficiently small ζ_n such that $\chi_j(\boldsymbol{\psi}) \neq 0$ for any $\boldsymbol{\psi} \in \boldsymbol{\Psi}_{**}$ and $j \in \mathcal{R}_n\backslash\mathcal{I}$. For any $\boldsymbol{\psi} \in \boldsymbol{\Psi}_{**}$, let $\tilde{\boldsymbol{\lambda}}(\boldsymbol{\psi}) \in \mathbb{R}^r$ satisfy $\tilde{\boldsymbol{\lambda}}_{\mathcal{R}_n}(\boldsymbol{\psi}) = \boldsymbol{\chi}(\boldsymbol{\psi})$ and $\tilde{\boldsymbol{\lambda}}_{\mathcal{R}_n^{\mathrm{c}}}(\boldsymbol{\psi}) = \mathbf{0}$. We will show that $\hat{\boldsymbol{\lambda}}(\boldsymbol{\psi}) = \tilde{\boldsymbol{\lambda}}(\boldsymbol{\psi})$ for any $\boldsymbol{\psi} \in \boldsymbol{\Psi}_{**}$ w.p.a.1. Restricted on $\mathcal{E}$, for any $j \in \mathcal{R}_n^{\mathrm{c}}$, we have

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} \frac{g_{i,j}^{(\mathcal{T})}(\boldsymbol{\psi})}{1 + \tilde{\boldsymbol{\lambda}}(\boldsymbol{\psi})^{\mathrm{T}}\mathbf{g}_i^{(\mathcal{T})}(\boldsymbol{\psi})} \\
&\quad= \frac{1}{n}\sum_{i=1}^{n} \frac{g_{i,j}^{(\mathcal{T})}(\hat{\boldsymbol{\psi}}_{\mathrm{PEL}})}{1 + \tilde{\boldsymbol{\lambda}}(\boldsymbol{\psi})^{\mathrm{T}}\mathbf{g}_i^{(\mathcal{T})}(\hat{\boldsymbol{\psi}}_{\mathrm{PEL}})} \\
&\qquad+ \left[\frac{1}{n}\sum_{i=1}^{n} \frac{\{\nabla_{\boldsymbol{\psi}} g_{i,j}^{(\mathcal{T})}(\check{\boldsymbol{\psi}})\}^{\mathrm{T}}}{1 + \tilde{\boldsymbol{\lambda}}(\boldsymbol{\psi})^{\mathrm{T}}\mathbf{g}_i^{(\mathcal{T})}(\check{\boldsymbol{\psi}})} - \frac{1}{n}\sum_{i=1}^{n} \frac{g_{i,j}^{(\mathcal{T})}(\check{\boldsymbol{\psi}})\tilde{\boldsymbol{\lambda}}(\boldsymbol{\psi})^{\mathrm{T}}\nabla_{\boldsymbol{\psi}}\mathbf{g}_i^{(\mathcal{T})}(\check{\boldsymbol{\psi}})}{\{1 + \tilde{\boldsymbol{\lambda}}(\boldsymbol{\psi})^{\mathrm{T}}\mathbf{g}_i^{(\mathcal{T})}(\check{\boldsymbol{\psi}})\}^2}\right](\boldsymbol{\psi} - \hat{\boldsymbol{\psi}}_{\mathrm{PEL}}) \\
&\quad= \frac{1}{n}\sum_{i=1}^{n} \frac{g_{i,j}^{(\mathcal{T})}(\hat{\boldsymbol{\psi}}_{\mathrm{PEL}})}{1 + \hat{\boldsymbol{\lambda}}(\hat{\boldsymbol{\psi}}_{\mathrm{PEL}})^{\mathrm{T}}\mathbf{g}_i^{(\mathcal{T})}(\hat{\boldsymbol{\psi}}_{\mathrm{PEL}})} + |\boldsymbol{\psi} - \hat{\boldsymbol{\psi}}_{\mathrm{PEL}}|_1 \cdot O_{\mathrm{p}}(1) \\
&\qquad- \left[\frac{1}{n}\sum_{i=1}^{n} \frac{g_{i,j}^{(\mathcal{T})}(\hat{\boldsymbol{\psi}}_{\mathrm{PEL}})\mathbf{g}_i^{(\mathcal{T})}(\hat{\boldsymbol{\psi}}_{\mathrm{PEL}})^{\mathrm{T}}}{\{1 + \check{\boldsymbol{\lambda}}^{\mathrm{T}}\mathbf{g}_i^{(\mathcal{T})}(\hat{\boldsymbol{\psi}}_{\mathrm{PEL}})\}^2}\right]\{\tilde{\boldsymbol{\lambda}}(\boldsymbol{\psi}) - \hat{\boldsymbol{\lambda}}(\hat{\boldsymbol{\psi}}_{\mathrm{PEL}})\} \\
&\quad= \frac{1}{n}\sum_{i=1}^{n} \frac{g_{i,j}^{(\mathcal{T})}(\hat{\boldsymbol{\psi}}_{\mathrm{PEL}})}{1 + \hat{\boldsymbol{\lambda}}(\hat{\boldsymbol{\psi}}_{\mathrm{PEL}})^{\mathrm{T}}\mathbf{g}_i^{(\mathcal{T})}(\hat{\boldsymbol{\psi}}_{\mathrm{PEL}})} + |\boldsymbol{\chi}(\boldsymbol{\psi}) - \boldsymbol{\chi}(\hat{\boldsymbol{\psi}}_{\mathrm{PEL}})|_1 \cdot O_{\mathrm{p}}(1) + |\boldsymbol{\psi} - \hat{\boldsymbol{\psi}}_{\mathrm{PEL}}|_1 \cdot O_{\mathrm{p}}(1) \\
&\quad= \hat{\eta}_j + \kappa_n \cdot o_{\mathrm{p}}(1)\,,
\end{aligned}
$$

where $\check{\boldsymbol{\psi}}$ is on the line joining $\boldsymbol{\psi}$ and $\hat{\boldsymbol{\psi}}_{\mathrm{PEL}}$, $\check{\boldsymbol{\lambda}}$ is on the line joining $\tilde{\boldsymbol{\lambda}}(\boldsymbol{\psi})$ and $\hat{\boldsymbol{\lambda}}(\hat{\boldsymbol{\psi}}_{\mathrm{PEL}})$, and the terms $O_{\mathrm{p}}(1)$ and $o_{\mathrm{p}}(1)$ hold uniformly over $j \in \mathcal{R}_n^{\mathrm{c}}$. Notice that $\mathbb{P}(\mathcal{E}) \to 1$. Therefore, it holds w.p.a.1 that

$$
\max_{j\in\mathcal{R}_n^{\mathrm{c}}} \left|\frac{1}{n}\sum_{i=1}^{n} \frac{g_{i,j}^{(\mathcal{T})}(\boldsymbol{\psi})}{1 + \tilde{\boldsymbol{\lambda}}(\boldsymbol{\psi})^{\mathrm{T}}\mathbf{g}_i^{(\mathcal{T})}(\boldsymbol{\psi})}\right| \leq \nu\rho_2'(0^+)\,.
$$

On the other hand, since $\mathbf{m}\{\boldsymbol{\chi}(\boldsymbol{\psi}), \boldsymbol{\psi}\} = \mathbf{0}$, $\tilde{\boldsymbol{\lambda}}_{\mathcal{R}_n}(\boldsymbol{\psi}) = \boldsymbol{\chi}(\boldsymbol{\psi})$ and $\tilde{\boldsymbol{\lambda}}_{\mathcal{R}_n^{\mathrm{c}}}(\boldsymbol{\psi}) = \mathbf{0}$ for any $\boldsymbol{\psi} \in \boldsymbol{\Psi}_{**}$, then it holds that

$$
0 = \frac{1}{n}\sum_{i=1}^{n} \frac{g_{i,j}^{(\mathcal{T})}(\boldsymbol{\psi})}{1 + \tilde{\boldsymbol{\lambda}}(\boldsymbol{\psi})^{\mathrm{T}}\mathbf{g}_i^{(\mathcal{T})}(\boldsymbol{\psi})} - \nu\rho_2'\{|\tilde{\lambda}_j(\boldsymbol{\psi})|; \nu\}\mathrm{sgn}\{\tilde{\lambda}_j(\boldsymbol{\psi})\}
$$

for any $j \in \mathcal{R}_n \backslash \mathcal{I}$, where $\tilde{\boldsymbol{\lambda}}(\boldsymbol{\psi}) = \{\tilde{\lambda}_1(\boldsymbol{\psi}), \ldots, \tilde{\lambda}_r(\boldsymbol{\psi})\}^{\mathrm{T}}$, and $n^{-1}\sum_{i=1}^n g_{i,j}^{(\mathcal{T})}(\boldsymbol{\psi})\{1+\tilde{\boldsymbol{\lambda}}(\boldsymbol{\psi})^{\mathrm{T}}\mathbf{g}_i^{(\mathcal{T})}(\boldsymbol{\psi})\}^{-1} = 0$ for any $j \in \mathcal{I}$. By the concavity of $f(\boldsymbol{\lambda}; \boldsymbol{\psi})$ with respect to $\boldsymbol{\lambda}$, we know $\hat{\boldsymbol{\lambda}}(\boldsymbol{\psi}) = \tilde{\boldsymbol{\lambda}}(\boldsymbol{\psi})$ for any $\boldsymbol{\psi} \in \boldsymbol{\Psi}_{**}$ w.p.a.1. Therefore, $\hat{\boldsymbol{\lambda}}(\boldsymbol{\psi})$ is continuously differentiable at $\hat{\boldsymbol{\psi}}_{\mathrm{PEL}}$ and $\nabla_{\boldsymbol{\psi}}\hat{\boldsymbol{\lambda}}_{\mathcal{R}_n^{\mathrm{c}}}(\hat{\boldsymbol{\psi}}_{\mathrm{PEL}}) = \mathbf{0}$ w.p.a.1. We complete the proof. □

A.7.4 Proof of Lemma A.7

By the proof of Theorem 1 in Chang et al. (2021), we have $|\bar{\mathbf{f}}^{\mathbf{A}_n}(\boldsymbol{\psi}_{0,\mathcal{M}}, \boldsymbol{\psi}^*_{\mathcal{M}^{\mathrm{c}}})|_2 = O_{\mathrm{p}}(m^{1/2}n^{-1/2})$ provided that $n\varpi_{2,n}^2(\varsigma^2 + \varpi_{1,n}^2 + \varpi_{2,n}^2) = O(1)$ and $\omega_n^2 \log r = O(1)$. Next, we will specify the convergence rate of $|\bar{\mathbf{f}}^{\mathbf{A}_n}(\tilde{\boldsymbol{\psi}}_{\mathcal{M}}, \boldsymbol{\psi}^*_{\mathcal{M}^{\mathrm{c}}})|_2$. Define $B_n(\boldsymbol{\psi}_{\mathcal{M}}, \boldsymbol{\lambda}) = n^{-1}\sum_{i=1}^n \log\{1 + \boldsymbol{\lambda}^{\mathrm{T}}\mathbf{f}_i^{\mathbf{A}_n}(\boldsymbol{\psi}_{\mathcal{M}}, \boldsymbol{\psi}^*_{\mathcal{M}^{\mathrm{c}}})\}$ for any $\boldsymbol{\psi}_{\mathcal{M}} \in \boldsymbol{\Psi}^*_{\mathcal{M}}$ and $\boldsymbol{\lambda} \in \tilde{\Lambda}_n(\boldsymbol{\psi}_{\mathcal{M}})$. Let $\tilde{\boldsymbol{\lambda}} = \arg\max_{\boldsymbol{\lambda} \in \tilde{\Lambda}_n(\boldsymbol{\psi}_{0,\mathcal{M}})} B_n(\boldsymbol{\psi}_{0,\mathcal{M}}, \boldsymbol{\lambda})$. Pick $\delta_n = o(m^{-1/2}n^{-1/\gamma})$ and $m^{1/2}n^{-1/2} = o(\delta_n)$. Let $\bar{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda} \in \Lambda_n} B_n(\boldsymbol{\psi}_{0,\mathcal{M}}, \boldsymbol{\lambda})$ where $\Lambda_n = \{\boldsymbol{\lambda} \in \mathbb{R}^m : |\boldsymbol{\lambda}|_2 \leq \delta_n\}$. Conditions 3 and 7 imply that $\max_{i\in[n]} |\mathbf{f}_i^{\mathbf{A}_n}(\boldsymbol{\psi}_{0,\mathcal{M}}, \boldsymbol{\psi}^*_{\mathcal{M}^{\mathrm{c}}})|_2 = O_{\mathrm{p}}(m^{1/2}n^{1/\gamma})$, which implies $\max_{i\in[n],\boldsymbol{\lambda}\in\Lambda_n} |\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{f}_i^{\mathbf{A}_n}(\boldsymbol{\psi}_{0,\mathcal{M}}, \boldsymbol{\psi}^*_{\mathcal{M}^{\mathrm{c}}})| = o_{\mathrm{p}}(1)$. Under Conditions 4 and 7, if $mn^{-1/2} = o(1)$ and $m(\omega_n^2 + \varpi_{2,n}^2) = o(1)$, Lemma 4 of Chang et al. (2021) implies that the eigenvalues of $n^{-1}\sum_{i=1}^n \mathbf{f}_i^{\mathbf{A}_n}(\boldsymbol{\psi}_{0,\mathcal{M}}, \boldsymbol{\psi}^*_{\mathcal{M}^{\mathrm{c}}})^{\otimes 2}$ are uniformly bounded away from zero and infinity w.p.a.1. By the Taylor expansion, we have

$$
\begin{aligned}
0 = B_n(\boldsymbol{\psi}_{0,\mathcal{M}}, \mathbf{0}) &\leq B_n(\boldsymbol{\psi}_{0,\mathcal{M}}, \bar{\boldsymbol{\lambda}}) \\
&= \bar{\boldsymbol{\lambda}}^{\mathrm{T}}\bar{\mathbf{f}}^{\mathbf{A}_n}(\boldsymbol{\psi}_{0,\mathcal{M}}, \boldsymbol{\psi}^*_{\mathcal{M}^{\mathrm{c}}}) - \frac{1}{2n}\sum_{i=1}^n \frac{\bar{\boldsymbol{\lambda}}^{\mathrm{T}}\mathbf{f}_i^{\mathbf{A}_n}(\boldsymbol{\psi}_{0,\mathcal{M}}, \boldsymbol{\psi}^*_{\mathcal{M}^{\mathrm{c}}})^{\otimes 2}\bar{\boldsymbol{\lambda}}}{\{1 + c\bar{\boldsymbol{\lambda}}^{\mathrm{T}}\mathbf{f}_i^{\mathbf{A}_n}(\boldsymbol{\psi}_{0,\mathcal{M}}, \boldsymbol{\psi}^*_{\mathcal{M}^{\mathrm{c}}})\}^2} \\
&\leq |\bar{\boldsymbol{\lambda}}|_2 |\bar{\mathbf{f}}^{\mathbf{A}_n}(\boldsymbol{\psi}_{0,\mathcal{M}}, \boldsymbol{\psi}^*_{\mathcal{M}^{\mathrm{c}}})|_2 - C|\bar{\boldsymbol{\lambda}}|_2^2\{1 + o_{\mathrm{p}}(1)\}
\end{aligned}
$$

for some $c \in (0, 1)$. Since $|\bar{\mathbf{f}}^{\mathbf{A}_n}(\boldsymbol{\psi}_{0,\mathcal{M}}, \boldsymbol{\psi}^*_{\mathcal{M}^{\mathrm{c}}})|_2 = O_{\mathrm{p}}(m^{1/2}n^{-1/2})$, then $|\bar{\boldsymbol{\lambda}}|_2 = O_{\mathrm{p}}(m^{1/2}n^{-1/2}) = o_{\mathrm{p}}(\delta_n)$ which implies $\bar{\boldsymbol{\lambda}} \in \mathrm{int}(\Lambda_n)$ w.p.a.1. Due to $\Lambda_n \subset \tilde{\Lambda}_n(\boldsymbol{\psi}_{0,\mathcal{M}})$ w.p.a.1 and the concavity of $B_n(\boldsymbol{\psi}_{0,\mathcal{M}}, \boldsymbol{\lambda})$, $\tilde{\boldsymbol{\lambda}} = \bar{\boldsymbol{\lambda}}$ w.p.a.1 and $\max_{\boldsymbol{\lambda}\in\tilde{\Lambda}_n(\boldsymbol{\psi}_{0,\mathcal{M}})} B_n(\boldsymbol{\psi}_{0,\mathcal{M}}, \boldsymbol{\lambda}) = O_{\mathrm{p}}(mn^{-1})$. For δ_n specified above, let $\boldsymbol{\lambda}^* = \delta_n\bar{\mathbf{f}}^{\mathbf{A}_n}(\tilde{\boldsymbol{\psi}}_{\mathcal{M}}, \boldsymbol{\psi}^*_{\mathcal{M}^{\mathrm{c}}})/|\bar{\mathbf{f}}^{\mathbf{A}_n}(\tilde{\boldsymbol{\psi}}_{\mathcal{M}}, \boldsymbol{\psi}^*_{\mathcal{M}^{\mathrm{c}}})|_2$ and then $\boldsymbol{\lambda}^* \in \Lambda_n$. Similar to Lemma 4 of Chang et al. (2021), under Conditions 4 and 7, if $mn^{-1/2} = o(1)$ and $m(\omega_n^2 + \varpi_{1,n}^2 + \varpi_{2,n}^2) = o(1)$, we have the eigenvalues of $n^{-1}\sum_{i=1}^n \mathbf{f}_i^{\mathbf{A}_n}(\tilde{\boldsymbol{\psi}}_{\mathcal{M}}, \boldsymbol{\psi}^*_{\mathcal{M}^{\mathrm{c}}})^{\otimes 2}$ are uniformly bounded away from zero and infinity w.p.a.1. Applying the Taylor expansion again, it holds that

$$
\begin{aligned}
O_{\mathrm{p}}(mn^{-1}) = \max_{\boldsymbol{\lambda}\in\tilde{\Lambda}_n(\boldsymbol{\psi}_{0,\mathcal{M}})} B_n(\boldsymbol{\psi}_{0,\mathcal{M}}, \boldsymbol{\lambda}) &\geq B_n(\tilde{\boldsymbol{\psi}}_{\mathcal{M}}, \boldsymbol{\lambda}^*) \\
&= \boldsymbol{\lambda}^{*,\mathrm{T}}\bar{\mathbf{f}}^{\mathbf{A}_n}(\tilde{\boldsymbol{\psi}}_{\mathcal{M}}, \boldsymbol{\psi}^*_{\mathcal{M}^{\mathrm{c}}}) - \frac{1}{2n}\sum_{i=1}^n \frac{\boldsymbol{\lambda}^{*,\mathrm{T}}\mathbf{f}_i^{\mathbf{A}_n}(\tilde{\boldsymbol{\psi}}_{\mathcal{M}}, \boldsymbol{\psi}^*_{\mathcal{M}^{\mathrm{c}}})^{\otimes 2}\boldsymbol{\lambda}^*}{\{1 + c\boldsymbol{\lambda}^{*,\mathrm{T}}\mathbf{f}_i^{\mathbf{A}_n}(\tilde{\boldsymbol{\psi}}_{\mathcal{M}}, \boldsymbol{\psi}^*_{\mathcal{M}^{\mathrm{c}}})\}^2} \\
&\geq \delta_n|\bar{\mathbf{f}}^{\mathbf{A}_n}(\tilde{\boldsymbol{\psi}}_{\mathcal{M}}, \boldsymbol{\psi}^*_{\mathcal{M}^{\mathrm{c}}})|_2 - C\delta_n^2\{1 + o_{\mathrm{p}}(1)\}
\end{aligned}
$$

for some $c \in (0,1)$, which implies that $|\bar{\mathbf{f}}^{\mathbf{A}_n}(\tilde{\boldsymbol{\psi}}_{\mathcal{M}}, \boldsymbol{\psi}^*_{\mathcal{M}^c})|_2 = O_{\mathrm{p}}(\delta_n)$. Given any $\epsilon_n \to 0$, let $\boldsymbol{\lambda}^{**} = \epsilon_n \bar{\mathbf{f}}^{\mathbf{A}_n}(\tilde{\boldsymbol{\psi}}_{\mathcal{M}}, \boldsymbol{\psi}^*_{\mathcal{M}^c})$. Repeating above arguments again, we have $\epsilon_n |\bar{\mathbf{f}}^{\mathbf{A}_n}(\tilde{\boldsymbol{\psi}}_{\mathcal{M}}, \boldsymbol{\psi}^*_{\mathcal{M}^c})|_2^2 = O_{\mathrm{p}}(mn^{-1})$. Notice that we can select an arbitrary slow $\epsilon_n \to 0$. Then $|\bar{\mathbf{f}}^{\mathbf{A}_n}(\tilde{\boldsymbol{\psi}}_{\mathcal{M}}, \boldsymbol{\psi}^*_{\mathcal{M}^c})|_2 = O_{\mathrm{p}}(m^{1/2}n^{-1/2})$.

Notice that $|\bar{\mathbf{f}}^{\mathbf{A}_n}(\boldsymbol{\psi}_{0,\mathcal{M}}, \boldsymbol{\psi}^*_{\mathcal{M}^c}) - \bar{\mathbf{f}}^{\mathbf{A}_n}(\tilde{\boldsymbol{\psi}}_{\mathcal{M}}, \boldsymbol{\psi}^*_{\mathcal{M}^c})|_2 \geq \lambda_{\min}^{1/2}[\{\nabla_{\boldsymbol{\psi}_{\mathcal{M}}} \bar{\mathbf{f}}^{\mathbf{A}_n}(\check{\boldsymbol{\psi}}_{\mathcal{M}}, \boldsymbol{\psi}^*_{\mathcal{M}^c})\}^{\mathrm{T},\otimes 2}] |\tilde{\boldsymbol{\psi}}_{\mathcal{M}} - \boldsymbol{\psi}_{0,\mathcal{M}}|_2$, where $\check{\boldsymbol{\psi}}_{\mathcal{M}}$ is on the jointing line between $\boldsymbol{\psi}_{0,\mathcal{M}}$ and $\tilde{\boldsymbol{\psi}}_{\mathcal{M}}$. Recall $\mathbf{A} = (\mathbf{a}_j)_{j\in\mathcal{M}}^{\mathrm{T}}$ and let $\boldsymbol{\Gamma} = (\boldsymbol{\gamma}_j)_{j\in\mathcal{M}}^{\mathrm{T}}$. Then $\mathbf{A}\mathbb{E}\{\nabla_{\boldsymbol{\psi}} \mathbf{g}_i^{(\mathcal{T})}(\boldsymbol{\psi}_0)\} = \boldsymbol{\Gamma}$. Write $\mathcal{M} = \{j_1, \ldots, j_m\}$ and denote by $\boldsymbol{\Gamma}_{\cdot,\mathcal{M}}$ the columns of $\boldsymbol{\Gamma}$ that are indexed in $\mathcal{M}$. Recall $\boldsymbol{\gamma}_{j_k}$ is a $(p + r_2)$-dimensional vector with its j_k-th component being 1 and all other components being 0. Then $\boldsymbol{\Gamma}_{\cdot,\mathcal{M}} = \mathbf{I}_m$. Hence, the eigenvalues of $[\mathbb{E}\{\nabla_{\boldsymbol{\psi}_{\mathcal{M}}} \mathbf{f}_i^{\mathbf{A}}(\boldsymbol{\psi}_0)\}]^{\mathrm{T},\otimes 2} = \boldsymbol{\Gamma}_{\cdot,\mathcal{M}}^{\mathrm{T},\otimes 2}$ are uniformly bounded away from zero. Analogously to Lemma A.4, we have $|[\nabla_{\boldsymbol{\psi}_{\mathcal{M}}} \bar{\mathbf{f}}^{\mathbf{A}_n}(\check{\boldsymbol{\psi}}_{\mathcal{M}}, \boldsymbol{\psi}^*_{\mathcal{M}^c}) - \mathbb{E}\{\nabla_{\boldsymbol{\psi}_{\mathcal{M}}} \mathbf{f}_i^{\mathbf{A}}(\boldsymbol{\psi}_0)\}]\mathbf{z}|_2 = |\mathbf{z}|_2 \cdot [O_{\mathrm{p}}(mn^{-1/2}) + O_{\mathrm{p}}\{m(\omega_n + \varpi_{1,n} + \varpi_{2,n})\}]$ holds uniformly over $\mathbf{z} \in \mathbb{R}^m$. Then, if $mn^{-1/2} = o(1)$ and $m(\omega_n + \varpi_{1,n} + \varpi_{2,n}) = o(1)$, it holds that $\lambda_{\min}[\{\nabla_{\boldsymbol{\psi}_{\mathcal{M}}} \bar{\mathbf{f}}^{\mathbf{A}_n}(\check{\boldsymbol{\psi}}_{\mathcal{M}}, \boldsymbol{\psi}^*_{\mathcal{M}^c})\}^{\mathrm{T},\otimes 2}] \geq C$ w.p.a.1, which implies $|\tilde{\boldsymbol{\psi}}_{\mathcal{M}} - \boldsymbol{\psi}_{0,\mathcal{M}}|_2 = O_{\mathrm{p}}(m^{1/2}n^{-1/2})$. We complete the proof. □

B Additional numerical results

Due to limitations of space, in the main text we report the simulation results of the linear IV model only. In this section, we display additional results for the linear IV model, and also experiment with a nonlinear model.

B.1 Linear IV model

Although economists are primarily interested in the effect of the endogenous variable, the exogenous variables in $\mathbf{z}_i$ control other sources of heterogeneity. To render a full picture of the performance of the estimation, we report the simulation results of $\beta_{\mathbf{z},1}$ and $\beta_{\mathbf{z},2}$—the coefficients for the exogenous variables z_1 and z_2—from the linear IV model in the main text.

Table S1 displays the RMSE of PEL and 2SLS for $(\beta_{\mathbf{z},1}, \beta_{\mathbf{z},2})$. Here,

$$\mathrm{RMSE} = \sqrt{\frac{1}{S}\sum_{s=1}^{S}\{(\hat{\beta}_{\mathbf{z},1}^{(s)} - \beta_{\mathbf{z},1})^2 + (\hat{\beta}_{\mathbf{z},2}^{(s)} - \beta_{\mathbf{z},2})^2\}},$$

where S denotes the number of repetitions and $(\hat{\beta}_{\mathbf{z},1}^{(s)}, \hat{\beta}_{\mathbf{z},2}^{(s)})$ denotes the estimate of $(\beta_{\mathbf{z},1}, \beta_{\mathbf{z},2})$ in the s-th repetition. The RMSE of 2SLS is significantly larger than that of PEL, showing that the advantages of PEL extend to the estimation for the coefficients of the exogenous variables.

The coverage probabilities for the CIs of $\beta_{\mathbf{z},1}$ and $\beta_{\mathbf{z},2}$ are summarized in Table S2. It reveals the strength of PPEL for the inference of these coefficients. Moreover, Figure S1 characterizes the shape of the confidence regions for the cases $(100, 120, 6)$ and $(200, 240, 6)$ with moderate

Table S1: RMSEs of the point estimations for $(\beta_{\mathbf{z},1}, \beta_{\mathbf{z},2})$

(n, d_w, s)	Correlation:	weak	moderate	strong
Panel A: low-dimensional setting				
(100, 50, 6)	PEL	0.172	0.176	0.167
	2SLS	0.219	0.219	0.219
(200, 100, 6)	PEL	0.114	0.115	0.115
	2SLS	0.152	0.152	0.152
Panel B: high-dimensional setting				
(100, 120, 6)	PEL	0.175	0.183	0.177
	2SLS	0.245	0.245	0.245
(200, 240, 6)	PEL	0.124	0.125	0.122
	2SLS	0.199	0.199	0.199
(100, 120, 8)	PEL	0.207	0.211	0.207
	2SLS	0.300	0.300	0.300
(200, 240, 12)	PEL	0.114	0.113	0.113
	2SLS	0.188	0.188	0.188
(100, 120, 13)	PEL	0.220	0.239	0.213
	2SLS	0.293	0.293	0.293
(200, 240, 17)	PEL	0.130	0.133	0.126
	2SLS	0.199	0.199	0.199

Table S2: Coverage probabilities for the CIs of $\beta_{\mathbf{z},1}$ and $\beta_{\mathbf{z},2}$ by PPEL

	Correlation:	weak			moderate			strong		
(n, d_w, s)	Method	90	95	99	90	95	99	90	95	99
Panel A: low-dimensional setting										
(100, 50, 6)	$\beta_{\mathbf{z},1}$	0.892	0.948	0.992	0.888	0.948	0.992	0.890	0.950	0.992
	$\beta_{\mathbf{z},2}$	0.874	0.930	0.990	0.868	0.926	0.990	0.870	0.928	0.990
(200, 100, 6)	$\beta_{\mathbf{z},1}$	0.898	0.940	0.988	0.896	0.936	0.988	0.894	0.938	0.988
	$\beta_{\mathbf{z},2}$	0.880	0.938	0.984	0.880	0.936	0.986	0.882	0.936	0.986
Panel B: high-dimensional setting										
(100, 120, 6)	$\beta_{\mathbf{z},1}$	0.878	0.940	0.984	0.870	0.942	0.980	0.874	0.932	0.980
	$\beta_{\mathbf{z},2}$	0.892	0.930	0.978	0.884	0.926	0.972	0.886	0.926	0.970
(200, 240, 6)	$\beta_{\mathbf{z},1}$	0.896	0.936	0.976	0.894	0.936	0.976	0.892	0.938	0.976
	$\beta_{\mathbf{z},2}$	0.878	0.934	0.986	0.872	0.934	0.986	0.868	0.936	0.986
(100, 120, 8)	$\beta_{\mathbf{z},1}$	0.870	0.924	0.976	0.866	0.926	0.974	0.864	0.928	0.974
	$\beta_{\mathbf{z},2}$	0.854	0.928	0.976	0.850	0.924	0.960	0.834	0.916	0.964
(200, 240, 12)	$\beta_{\mathbf{z},1}$	0.886	0.954	0.994	0.884	0.952	0.994	0.874	0.942	0.994
	$\beta_{\mathbf{z},2}$	0.892	0.940	0.980	0.892	0.938	0.980	0.886	0.938	0.980
(100, 120, 13)	$\beta_{\mathbf{z},1}$	0.872	0.934	0.990	0.862	0.918	0.988	0.860	0.918	0.984
	$\beta_{\mathbf{z},2}$	0.864	0.930	0.982	0.842	0.916	0.974	0.848	0.916	0.976
(200, 240, 17)	$\beta_{\mathbf{z},1}$	0.898	0.944	0.988	0.896	0.944	0.990	0.892	0.942	0.988
	$\beta_{\mathbf{z},2}$	0.908	0.948	0.976	0.902	0.948	0.974	0.900	0.940	0.974

correlation between ϵ_i and $\mathbf{w}_{3i}$.

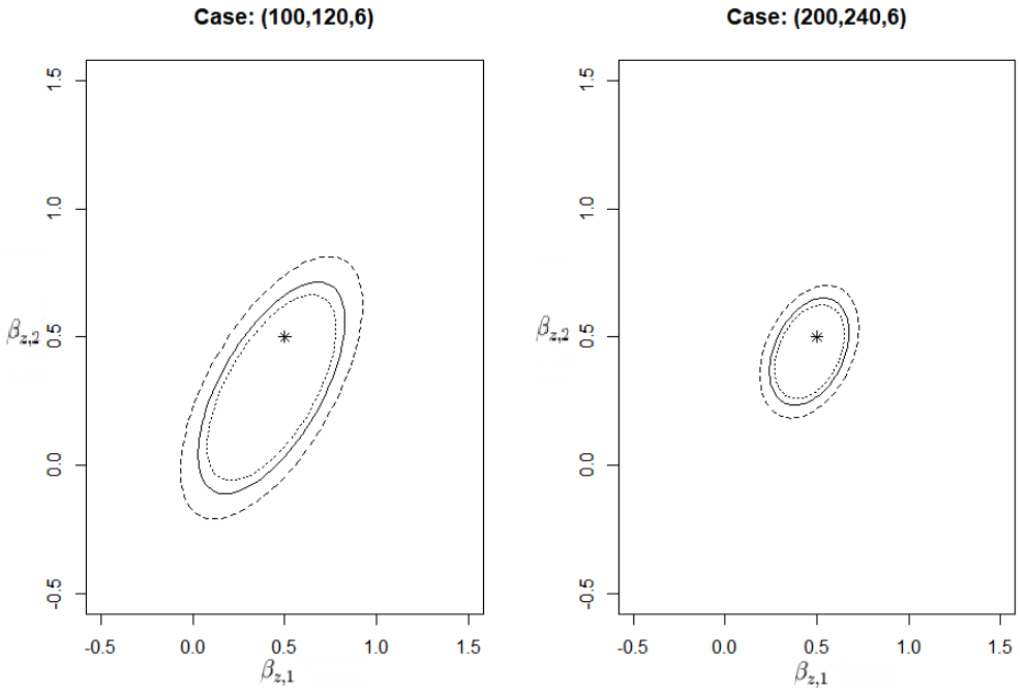

Figure S1: Confidence regions for $(\beta_{\mathbf{z},1}, \beta_{\mathbf{z},2})$ at levels 90%, 95% and 99%

B.2 Dynamic panel data model

Consider a simplified panel data model with time-varying individual heterogeneity (Han et al. 2005):

$$y_{i,j} = \lambda_j(\beta_1)\alpha_i + \beta_2 + e_{i,j}\,,$$

where the zero mean error term $e_{i,j}$ may potentially correlate with the individual-specific fixed effect α_i, and the nonlinear specification $\lambda_j(\beta_1) = 2/\{1+\exp(\beta_1 j)\}$, $j = 0, 1\ldots, h+s$, is originated from Kumbhakar (1990). Here we use j, instead of t, to represent the panel's time dimension in order to be consistent with the notations throughout the paper.

Notice that the fixed effect α_i can be canceled out by

$$\begin{aligned} e_{i,j} &= y_{i,j} - \lambda_j(\beta_1)\alpha_i - \beta_2 = y_{i,j} - \lambda_j(\beta_1)(y_{i,0} - \beta_2 - e_{i,0}) - \beta_2 \\ &= y_{i,j} + \{\lambda_j(\beta_1) - 1\}\beta_2 - \lambda_j(\beta_1)y_{i,0} + \lambda_j(\beta_1)e_{i,0} = g_{i,j}(\boldsymbol{\theta}) + \lambda_j(\beta_1)e_{i,0}\,, \end{aligned}$$

where $\boldsymbol{\theta} = (\beta_1, \beta_2)^{\mathrm{T}}$ and $g_{i,j}(\boldsymbol{\theta}) = y_{i,j} + \{\lambda_j(\beta_1) - 1\}\beta_2 - \lambda_j(\beta_1)y_{i,0}$. The above equation implies many moment conditions $\mathbb{E}\{g_{i,j}(\boldsymbol{\theta})\} = \mathbb{E}\{e_{i,j} - \lambda_j(\beta_1)e_{i,0}\} = 0$ when j varies.

The parameter of interest lies in β_1, which determines the speed of decay of the individual-specific shock α_i, whereas β_2 is an intercept. In the simulation exercises, the true parameter is set as $(\beta_1^0, \beta_2^0) = (0.5, -2)$. Let $r = h + s$ and we specify $(n, r) = (100, 120)$ and $(200, 240)$, $\alpha_i \sim \mathcal{N}(1, 1)$, and $e_{i,j} \sim 2^{-1/2}(\alpha_i - 1) + \mathcal{N}(0, 1/2)$ i.i.d across time. All data across individuals are independent. The number of known valid moments is fixed to be $r_1 = 5$, and then the number of validity-unknown moments is $r - r_1$. After h periods, there occurs a structural break in the mean where the intercept shifts from β_2 to $\beta_2 + \sigma$ and these corresponding moments become invalid at $j = h+1, \ldots, h+s$. The number of invalid moments is $s = (6, 8, 13)$ and $(6, 12, 17)$ when $n = 100$ and 200, respectively, and $\sigma = (0.2, 0.3, 0.4)$ for small, moderate, and large shifts, respectively. The choice of the tuning parameters stays the same as that in the linear model.

Tables S3–S5 are the counterparts of Tables 1–3 in the high-dimensional setting. In Table S3, correct moment selection improves with the sample size, and FN quickly goes to zero as σ increases. As shown in Table S4, PEL offers reasonable estimation of the parameter while bias correction is difficult in this nonlinear model as well. In Table S5, PPEL again exhibits more accurate coverage probability than DB-PEL. Our proposed estimation and inference procedures are effective in terms of moment selection, parameter estimation and coverage probability in this nonlinear panel data model with many moments.

Table S3: PEL's performance in moment selection

	σ	0.2		0.3		0.4	
(n, r, s)	Method	FP	FN	FP	FN	FP	FN
(100, 120, 6)	PEL	0.0874	0.1600	0.0980	0.0060	0.0984	0.0000
	DB-PEL	0.0872	0.1600	0.0980	0.0060	0.0983	0.0000
(200, 240, 6)	PEL	0.0168	0.0607	0.0237	0.0000	0.0166	0.0000
	DB-PEL	0.0168	0.0607	0.0237	0.0000	0.0166	0.0000
(100, 120, 8)	PEL	0.0986	0.1218	0.0937	0.0080	0.0914	0.0000
	DB-PEL	0.0985	0.1220	0.0936	0.0080	0.0913	0.0000
(200, 240, 12)	PEL	0.0305	0.0253	0.0303	0.0000	0.0167	0.0000
	DB-PEL	0.0305	0.0253	0.0303	0.0000	0.0167	0.0000
(100, 120, 13)	PEL	0.0988	0.1292	0.1009	0.0055	0.0926	0.0000
	DB-PEL	0.0987	0.1292	0.1007	0.0055	0.0924	0.0000
(200, 240, 17)	PEL	0.0394	0.0178	0.0305	0.0000	0.0395	0.0000
	DB-PEL	0.0394	0.0178	0.0305	0.0000	0.0395	0.0000

Table S4: Point estimations for β_1

	σ	0.2			0.3			0.4		
(n, r, s)	Method	RMSE	BIAS	STD	RMSE	BIAS	STD	RMSE	BIAS	STD
(100, 120, 6)	PEL	0.053	0.000	0.053	0.053	0.003	0.053	0.053	0.003	0.053
	DB-PEL	0.062	0.003	0.062	0.058	0.004	0.058	0.058	0.004	0.058
(200, 240, 6)	PEL	0.038	0.005	0.038	0.039	0.005	0.038	0.039	0.005	0.038
	DB-PEL	0.042	0.004	0.041	0.043	0.004	0.042	0.042	0.005	0.041
(100, 120, 8)	PEL	0.053	0.002	0.053	0.053	0.003	0.053	0.053	0.003	0.053
	DB-PEL	0.058	0.004	0.058	0.058	0.003	0.058	0.058	0.004	0.058
(200, 240, 12)	PEL	0.039	0.005	0.038	0.039	0.005	0.038	0.039	0.005	0.038
	DB-PEL	0.042	0.004	0.041	0.041	0.005	0.041	0.042	0.004	0.042
(100, 120, 13)	PEL	0.053	0.001	0.053	0.053	0.003	0.053	0.053	0.003	0.053
	DB-PEL	0.056	0.002	0.056	0.056	0.004	0.056	0.056	0.004	0.056
(200, 240, 17)	PEL	0.038	0.005	0.038	0.039	0.005	0.038	0.039	0.005	0.038
	DB-PEL	0.042	0.004	0.042	0.041	0.004	0.041	0.042	0.004	0.041

Table S5: Coverage probabilities for the CIs of β_1

	σ	0.2			0.3			0.4		
(n, r, s)	Method	90	95	99	90	95	99	90	95	99
(100, 120, 6)	PPEL	0.894	0.954	0.994	0.882	0.934	0.986	0.914	0.952	0.990
	DB-PEL	0.836	0.900	0.974	0.862	0.932	0.988	0.844	0.922	0.986
(200, 240, 6)	PPEL	0.888	0.930	0.980	0.896	0.952	0.990	0.882	0.926	0.974
	DB-PEL	0.846	0.904	0.980	0.834	0.906	0.978	0.828	0.904	0.984
(100, 120, 8)	PPEL	0.886	0.926	0.986	0.902	0.954	0.980	0.908	0.960	0.984
	DB-PEL	0.868	0.924	0.978	0.848	0.938	0.974	0.834	0.920	0.982
(200, 240, 12)	PPEL	0.894	0.948	0.986	0.896	0.948	0.990	0.898	0.952	0.978
	DB-PEL	0.846	0.906	0.980	0.848	0.914	0.978	0.834	0.910	0.974
(100, 120, 13)	PPEL	0.906	0.960	0.984	0.896	0.946	0.978	0.876	0.938	0.984
	DB-PEL	0.878	0.932	0.982	0.870	0.920	0.984	0.864	0.928	0.980
(200, 240, 17)	PPEL	0.926	0.952	0.990	0.894	0.950	0.992	0.912	0.960	0.990
	DB-PEL	0.844	0.896	0.982	0.844	0.918	0.980	0.838	0.908	0.978

C Robustness of empirical application

To check the stability of PEL estimator in the real data analysis, we draw bootstrap samples from the original data and apply the proposed method to these sub-samples. Specifically, to witness the effect of sample sizes, we draw from the original sample with no replacement bootstrap observations $n^* = 55, 51, 45$ and 40, which correspond to the sub-samples of size $(n-1)$, $\lceil 0.9n \rceil$, $\lceil 0.8n \rceil$ and $\lceil 0.7n \rceil$, respectively. Given a sub-sample, the proposed PEL method is used to estimate the coefficient β_x. We repeat the bootstrap exercise for 500 times. The results are summarized in Figure S2. The point estimate from the true sample is 0.937, and the bootstrap point estimates are well centered around it. Regarding the histograms, despite a thin right tail when $n^* = 40$, the dispersion quickly concentrates as n^* increases. The normal distribution appears to be an effective approximation when n^* is above 50.

References

Chang, J., Chen, S. X., Tang, C. Y. and Wu, T. T. (2021), 'High-dimensional empirical likelihood inference', *Biometrika* **108**(1), 127–147.

Chang, J., Tang, C. Y. and Wu, T. T. (2018), 'A new scope of penalized empirical likelihood with high-dimensional estimating equations', *The Annals of Statistics* **46**(6B), 3185–3216.

Chang, J., Tang, C. Y. and Wu, Y. (2013), 'Marginal empirical likelihood and sure independence feature screening', *The Annals of Statistics* **41**(4), 2123–2148.

Han, C., Orea, L. and Schmidt, P. (2005), 'Estimation of a panel data model with parametric temporal variation in individual effects', *Journal of Econometrics* **126**(2), 241–267.

Jing, B.-Y., Shao, Q.-M. and Wang, Q. (2003), 'Self-normalized cramér-type large deviations for independent random variables', *The Annals of Probability* **31**(4), 2167–2215.

Koenker, R. and Machado, J. (1999), 'Gmm inference when the number of moment conditions is large', *Journal of Econometrics* **93**(2), 327–344.

Kumbhakar, S. C. (1990), 'Production frontiers, panel data, and time-varying technical inefficiency', *Journal of Econometrics* **46**(1-2), 201–211.

Petrov, V. V. (1995), *Limit Theorems of Probability Theory: Sequences of Independent Random Variables*, Clarendon Press, Oxford.

Rudin, W. (1976), *Principles of Mathematical Analysis*, McGraw-Hill, New York.

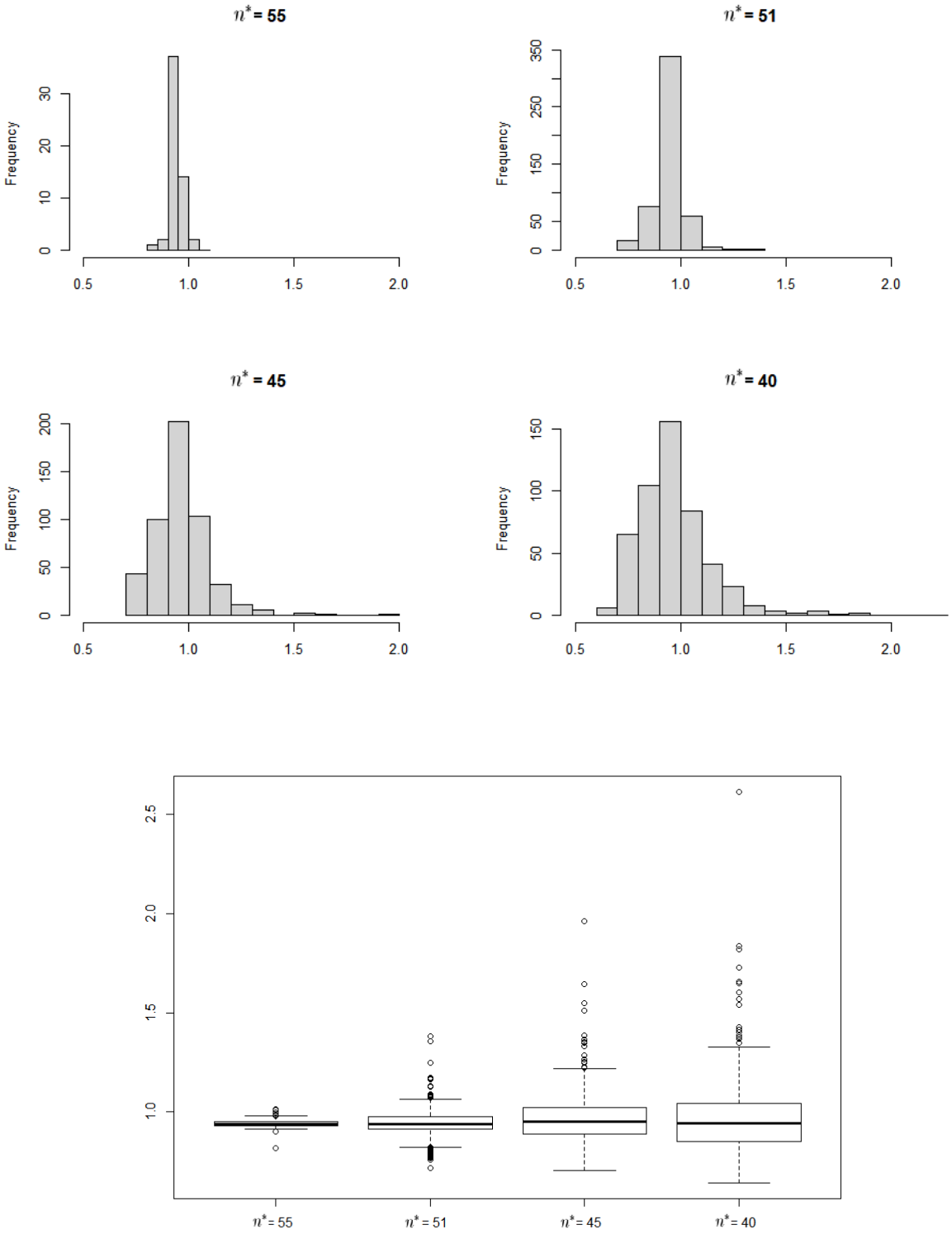

Figure S2: PEL estimation in bootstrap samples