Supplementary Materials for "Generalized Linear Models in Distributed Environments: A Framework for Optimal Subset Selection"

Hongmei Lin[1], Xinxin Xiao[1], Chen Guo[2], Yanlin Liu[1], Tiejun Tong[3] and Riquan Zhang[1]

[1] *School of Statistics and Data Science, Shanghai University of International Business and Economics, Shanghai, China*

[2] *KLATASDS-MOE, School of Statistics, East China Normal University, Shanghai, China*

[3] *Department of Mathematics, Hong Kong Baptist University, Hong Kong, China*

The supplementary materials provide the technical conditions required for the theorems in the main text, along with detailed proofs of the lemmas and theorems, as well as additional numerical results. For consistency, we use the same notations as in Sections 2 and 3 of the main text.

S1 Assumptions

To establish the asymptotic theory, we impose the following regularity conditions.

(A1) There exists a constant $c_0 \in (0, 1]$ such that the function $b(\theta)$ is three times continuously differentiable on its domain, with $c_0 \leq b''(\theta) \leq (c_0)^{-1}$ and $|b'''(\theta)| \leq (c_0)^{-1}$.

(A2) The model errors $W_i = y_i - \mathbb{E}(y_i), i = 1, \ldots, N$, are sub-Gaussian; that is, there exist universal positive constants c_1, c_2 such that $\mathbb{P}(|W_i| \geq$

$t) \leq c_1 \exp\{-c_2 t^2\}$ for any $t > 0$. A matrix $\mathbf{M}$ is said to satisfy the spectrum-restricted condition (SRC) of order s with spectrum bounds $\{m_s, M_s\}$ if

$$0 < m_s \leq \frac{\|\mathbf{M}_{\mathcal{A}}\mathbf{u}\|_2^2}{n\|\mathbf{u}\|_2^2} \leq M_s < \infty,$$

for all $\mathbf{u} \neq 0, \mathbf{u} \in \mathbb{R}^{|\mathcal{A}|}$, where $\mathcal{A} \subset \mathcal{F}$ and$|\mathcal{A}| \leq s$.

(A3) The matrix $\boldsymbol{X}$ satisfies the SRC of order $2s$ with spectrum bounds m_{2s}, M_{2s}. This implies that the spectra of the off-diagonal submatrices of $\boldsymbol{X}^\top \boldsymbol{X}$ are bounded by a constant ν_s, defined as the smallest value satisfying

$$\nu_s \geq \frac{\left\|\boldsymbol{X}_{\mathcal{A}}^\top \boldsymbol{X}_{\mathcal{B}}\mathbf{u}\right\|_2}{n\|\mathbf{u}\|_2}, \quad \forall \mathbf{u} \neq 0, \mathbf{u} \in \mathbb{R}^{|\mathcal{B}|}$$

where $\mathcal{A}, \mathcal{B} \subset \mathcal{S}$ with $|\mathcal{A}| \leq s, |\mathcal{B}| \leq s$, and $\mathcal{A} \cap \mathcal{B} = \varnothing$.

(A4) Let the constants C_1 and C_2 be defined as

$$C_1 := \frac{1}{c_0^2}\left(\frac{M_s}{2c_0} + \frac{\nu_s^2}{m_s c_0^3} + \frac{M_s v_s^2}{2m_s^2 c_0^5}\right)\left[\left(2 + \Delta + \frac{2\nu_s}{m_s c_0^2}\right)^2 + (2+\Delta)^2\right]\left(\frac{\nu_s}{m_s c_0^2}\right)^2,$$
$$C_2 := \frac{m_s c_0}{2} - \frac{\nu_s^2}{m_s c_0^3} - \frac{M_s v_s^2}{2m_s^2 c_0^5},$$

where $\Delta > 0$ is a constant such that $\gamma_s := \frac{C_1}{(1-\Delta)C_2} < 1$.

(A5) Assume that ν_s satisfies $\nu_s \leq \kappa_s := \max\{1 - m_{2s}, M_{2s} - 1\}$, and define the function $f(c_0, \kappa_s)$ by

$$\begin{aligned} f(c_0, \kappa_s) := &\frac{(1-\kappa_s)c_0}{2} - \frac{\kappa_s^2}{(1-\kappa_s)c_0^3} - \frac{(1+\kappa_s)\kappa_s^2}{2(1-\kappa_s)^2 c_0^5} \\ &- \frac{1}{c_0^2}\left(\frac{1+\kappa_s}{2c_0} + \frac{\kappa_s^2}{(1-\kappa_s)c_0^3} + \frac{(1+\kappa_s)\kappa_s^2}{2(1-\kappa_s)^2 c_0^5}\right) \\ &\left[\left(2 + \frac{2\kappa_s}{(1-\kappa_s)c_0^2}\right)^2 + 4\right]\frac{\kappa_s^2}{(1-\kappa_s)^2 c_0^4}. \end{aligned}$$

The condition $f(c_0, \kappa_s) > 0$ holds under sufficient assumptions, such as $\{c_0 \geq 0.9, \kappa_s \leq 0.1\}$ or $\{c_0 = 1, \kappa_s \leq 0.183\}$.

(A6) Let $a^* = \min_{j \in \mathcal{A}^*} |\beta_j^*|^2$, then $\frac{1}{a^*} = o\left(\frac{N}{s \log p \log\log N}\right)$.

(A7) The threshold τ_s satisfies $\tau_s = O\left(\frac{s \log p \log\log N}{N}\right)$.

(A8) The maximum sparsity $s_{\max}$ satisfies $s_{\max} = O\left(\left(\frac{N}{\log p}\right)^{\frac{1}{4}}\right)$.

(A9) There exists a constant c_3 such that $\max_{1 \le i \le N} \max_{1 \le j \le p} |x_{ij}| \le c_3$.

(A10) The number of machines m is allowed to grow with the total size N, satisfying $m = o\left(N^{1/2}\right)$.

(A1) and (A2) are the common assumptions in GLMs. They ensure that the variance of response variable remains bounded away from infinity or zero (Rigollet, 2012), thereby guaranteeing the existence of Fisher information (Fan & Tang, 2013) necessary for valid statistical inference. These assumptions are satisfied by commonly used GLMs, including linear, logistic, and Poisson regression. (A3) is a widely adopted identifiability condition in the literature on high-dimensional GLMs (Shi et al., 2019) and has been instrumental in rigorously analyzing the theoretical properties of methods like LASSO and MCP. In distributed settings, we assume that each of the m machines operates within these spectral bounds, ensuring a homogeneous and consistent environment across the distributed framework. (A4) requires ν_s to be sufficiently small to ensure its validity, for which we provide simpler sufficient conditions. Building on (A3) and the Lemma 20 in Huang et al. (2018), (A5) generalizes the corresponding linear model case, with $\kappa_s \le 0.183$ mirroring the Remark 2 in Zhu et al. (2020). (A6) imposes a lower bound on regression coefficients, while (A7) ensures that the threshold τ_s balances error control and computational efficiency. (A8) effectively controls the growth rate of $s_{\max}$, and (A9), a common requirement in high-dimensional GLMs, allows the predictor matrix $\boldsymbol{X}$ to be normalized through preprocessing for practical applications. Finally, (A10) restricts the growth of the number of machines m relative to the total sample size N, ensuring that the distributed framework remains statistically efficient and preventing excessive variance inflation.

S2 Proofs of the Lemmas

S2.1 Proof of Lemma 1

The surrogate likelihood loss function for the generalized linear model is defined as

$$\tilde{\mathcal{L}}(\boldsymbol{\beta}) = \mathcal{L}_1(\boldsymbol{\beta}) + \mathbf{g}^\top\boldsymbol{\beta}, \quad \text{where} \quad \mathbf{g} = \nabla\mathcal{L}(\bar{\boldsymbol{\beta}}) - \nabla\mathcal{L}_1(\bar{\boldsymbol{\beta}}).$$

In addition, $\mathcal{L}_1(\boldsymbol{\beta})$ denotes the local loss function at node 1, while $\nabla\mathcal{L}(\bar{\boldsymbol{\beta}})$ and $\nabla\mathcal{L}_1(\bar{\boldsymbol{\beta}})$ represent the gradients of the global and local loss functions evaluated at $\bar{\boldsymbol{\beta}}$, respectively.

The backward sacrifice quantifies the change in the surrogate likelihood loss when removing the j-th variable from the active set $\mathcal{A}$:

$$\xi_j^* = \tilde{\mathcal{L}}(\hat{\boldsymbol{\beta}}_{-j}) - \tilde{\mathcal{L}}(\hat{\boldsymbol{\beta}}) = \Delta\mathcal{L}_1 + \mathbf{g}^\top\Delta\hat{\boldsymbol{\beta}},$$

where $\hat{\boldsymbol{\beta}}_{-j} = \hat{\boldsymbol{\beta}}|_{\mathcal{A}\setminus\{j\}}$ and $\Delta\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{-j} - \hat{\boldsymbol{\beta}}$. The local loss function at node 1 is

$$\mathcal{L}_1(\boldsymbol{\beta}) = -\frac{1}{n}\sum_{i\in M_1}\left\{y_i\mathbf{x}_i^\top\boldsymbol{\beta} - b(\mathbf{x}_i^\top\boldsymbol{\beta}) + c(y_i,\phi)\right\}.$$

When calculating $\mathcal{L}_1(\hat{\boldsymbol{\beta}}|_{\mathcal{A}\setminus\{j\}}) - \mathcal{L}_1(\hat{\boldsymbol{\beta}})$, since $c(y_i,\phi)$ is independent of $\boldsymbol{\beta}$, it cancels out in the difference calculation.

$$\begin{aligned}\mathcal{L}_1(\hat{\boldsymbol{\beta}}|_{\mathcal{A}\setminus\{j\}} - \mathcal{L}_1(\hat{\boldsymbol{\beta}}) &= -\frac{1}{n}\sum_{i\in M_1}\left[y_i\mathbf{x}_i^\top\hat{\boldsymbol{\beta}}|_{\mathcal{A}\setminus\{j\}} - b(\mathbf{x}_i^\top\hat{\boldsymbol{\beta}}|_{\mathcal{A}\setminus\{j\}}) - (y_i\mathbf{x}_i^\top\hat{\boldsymbol{\beta}} - b(\mathbf{x}_i^\top\hat{\boldsymbol{\beta}}))\right] \\ &= -\frac{1}{n}\sum_{i\in M_1}\left[y_i\mathbf{x}_i^\top(\hat{\boldsymbol{\beta}}|_{\mathcal{A}\setminus\{j\}} - \hat{\boldsymbol{\beta}}) - (b(\mathbf{x}_i^\top\hat{\boldsymbol{\beta}}|_{\mathcal{A}\setminus\{j\}}) - b(\mathbf{x}_i^\top\hat{\boldsymbol{\beta}}))\right].\end{aligned}$$

If $b(\cdot)$ is twice differentiable, to further simplify the above formula, we perform a second-order Taylor expansion of $b(\mathbf{x}_i^\top \hat{\boldsymbol{\beta}}|_{\mathcal{A}\setminus\{j\}})$ at $\mathbf{x}_i^\top \hat{\boldsymbol{\beta}}$ as follows:

$$b(\mathbf{x}_i^\top \hat{\boldsymbol{\beta}}|_{\mathcal{A}\setminus\{j\}}) = b(\mathbf{x}_i^\top \hat{\boldsymbol{\beta}}) + b'(\mathbf{x}_i^\top \hat{\boldsymbol{\beta}})\mathbf{x}_i^\top (\hat{\boldsymbol{\beta}}|_{\mathcal{A}\setminus\{j\}} - \hat{\boldsymbol{\beta}}) + \frac{1}{2} b''(\xi_i) \left[\mathbf{x}_i^\top (\hat{\boldsymbol{\beta}}|_{\mathcal{A}\setminus\{j\}} - \hat{\boldsymbol{\beta}})\right]^2,$$

where ξ_i lies between $\mathbf{x}_i^\top \hat{\boldsymbol{\beta}}$ and $\mathbf{x}_i^\top \hat{\boldsymbol{\beta}}|_{\mathcal{A}\setminus\{j\}}$. This further leads to

$$b(\mathbf{x}_i^\top \hat{\boldsymbol{\beta}}|_{\mathcal{A}\setminus\{j\}}) - b(\mathbf{x}_i^\top \hat{\boldsymbol{\beta}}) = b'(\mathbf{x}_i^\top \hat{\boldsymbol{\beta}})\mathbf{x}_i^\top (\hat{\boldsymbol{\beta}}|_{\mathcal{A}\setminus\{j\}} - \hat{\boldsymbol{\beta}}) + \frac{1}{2} b''(\xi_i) \left[\mathbf{x}_i^\top (\hat{\boldsymbol{\beta}}|_{\mathcal{A}\setminus\{j\}} - \hat{\boldsymbol{\beta}})\right]^2,$$

and

$$\begin{aligned}
&\mathcal{L}_1(\hat{\boldsymbol{\beta}}|_{\mathcal{A}\setminus\{j\}}) - \mathcal{L}_1(\hat{\boldsymbol{\beta}}) \\
=& -\frac{1}{n}\sum_{i\in M_1} \Big[y_i \mathbf{x}_i^\top (\hat{\boldsymbol{\beta}}|_{\mathcal{A}\setminus\{j\}} - \hat{\boldsymbol{\beta}}) - b'(\mathbf{x}_i^\top \hat{\boldsymbol{\beta}})\mathbf{x}_i^\top (\hat{\boldsymbol{\beta}}|_{\mathcal{A}\setminus\{j\}} - \hat{\boldsymbol{\beta}}) \\
&-\frac{1}{2} b''(\xi_i) \left[\mathbf{x}_i^\top (\hat{\boldsymbol{\beta}}|_{\mathcal{A}\setminus\{j\}} - \hat{\boldsymbol{\beta}})\right]^2 \Big] \\
=& -\frac{1}{n}\sum_{i\in M_1} \left[(y_i - b'(\mathbf{x}_i^\top \hat{\boldsymbol{\beta}}))\mathbf{x}_i^\top (\hat{\boldsymbol{\beta}}|_{\mathcal{A}\setminus\{j\}} - \hat{\boldsymbol{\beta}}) - \frac{1}{2} b''(\xi_i) \left[\mathbf{x}_i^\top (\hat{\boldsymbol{\beta}}|_{\mathcal{A}\setminus\{j\}} - \hat{\boldsymbol{\beta}})\right]^2 \right].
\end{aligned}$$

Lastly, we have

$$\Delta\mathcal{L}_1 = -\frac{1}{n}\sum_{i\in M_1} \left[(y_i - b_i')\mathbf{x}_i^\top \Delta\hat{\boldsymbol{\beta}} - \frac{1}{2} b''(\xi_i)(\mathbf{x}_i^\top \Delta\hat{\boldsymbol{\beta}})^2 \right],$$

where $b_i' = b'(\mathbf{x}_i^\top \hat{\boldsymbol{\beta}})$, and ξ_i lies between $\mathbf{x}_i^\top \hat{\boldsymbol{\beta}}$ and $\mathbf{x}_i^\top \hat{\boldsymbol{\beta}}_{-j}$.

The forward sacrifice measures the change when adding the j-th variable from the inactive set $\mathcal{I}$:

$$\zeta_j^* = \tilde{\mathcal{L}}(\hat{\boldsymbol{\beta}}) - \tilde{\mathcal{L}}(\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{t}}_j) = -\Delta\mathcal{L}_1 - \mathbf{g}^\top \hat{\boldsymbol{t}}_j,$$

where $\hat{\boldsymbol{t}}_j = \arg\min_{\boldsymbol{t}} \tilde{\mathcal{L}}(\hat{\boldsymbol{\beta}} + \boldsymbol{t}|_j)$. Similarly, by performing a Taylor expansion of $b(\mathbf{x}_i^\top(\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{t}}_j))$, we obtain that

$$\Delta\mathcal{L}_1 = -\frac{1}{n}\sum_{i\in M_1} \left[(b_i' - y_i)\mathbf{x}_i^\top \hat{\boldsymbol{t}}_j + \frac{1}{2} b''(\eta_i)(\mathbf{x}_i^\top \hat{\boldsymbol{t}}_j)^2 \right],$$

where $b_i'' = b''(\eta_i)$, and η_i lies between $\mathbf{x}_i^\top \hat{\boldsymbol{\beta}}$ and $\mathbf{x}_i^\top(\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{t}}_j)$.

S2.2 Proof of Lemma 2

Given that $C < \left|\mathcal{A}^k\right|$, we consider the following inequality:

$$\min_{j\in\mathcal{S}_{C,2}^k}\frac{1}{\rho^2}\left|d_j^k\right|^2=\min_{j\in\mathcal{S}_{C,2}^k}\left|\beta_j^k+\frac{1}{\rho}d_j^k\right|^2\geqslant\max_{i\in\mathcal{A}^k}\left|\beta_i^k+\frac{1}{\rho}d_i^k\right|^2=\max_{i\in\mathcal{A}^k}\left|\beta_i^k\right|^2\geqslant\max_{i\in\mathcal{S}_{c,1}^k}\left|\beta_i^k\right|^2.$$

Through algebraic manipulation, we derive the corresponding range for ρ as

$$\rho\leqslant\frac{\min_{j\in\mathcal{S}_{C,2}^k}\left|d_j^k\right|}{\max_{i\in\mathcal{S}_{C,1}^k}\left|\beta_i^k\right|}. \tag{S2.1}$$

Similarly, given $C+1$, the range of ρ is

$$\rho\leqslant\frac{\min_{j\in\mathcal{S}_{C+1,2}}\left|d_j^k\right|}{\max_{i\in\mathcal{S}_{C+1,1}^k}\left|\beta_i^k\right|}. \tag{S2.2}$$

Notably, $\mathcal{S}_{C,1}^k\subseteq\mathcal{S}_{C+1,1}^k$ and $\mathcal{S}_{C,2}^k\subseteq\mathcal{S}_{C+1,2}^k$. If the restricted active and inactive sets exchange C elements, the range of ρ should be the difference between (S2.1) and (S2.2). Consequently, for $C<\left|\mathcal{A}^k\right|$, the range of ρ is specified as

$$\rho\in\left(\frac{\min_{j\in\mathcal{S}_{C+1,2}^k}\left|d_j^k\right|}{\max_{i\in\mathcal{S}_{C+1,1}^k}\left|\beta_i^k\right|},\frac{\min_{j\in\mathcal{S}_{C,2}^k}\left|d_j^k\right|}{\max_{i\in\mathcal{S}_{C,1}^k}\left|\beta_i^k\right|}\right].$$

For the case where $C=\left|\mathcal{A}^k\right|$, it follows that $\mathcal{S}_{C,1}^k=\mathcal{A}^k$. Thus, we have

$$\min_{j\in\mathcal{S}_{C,2}^k}\frac{1}{\rho^2}\left|d_j^k\right|^2=\min_{j\in\mathcal{S}_{C,2}^k}\left|\beta_j^k+\frac{1}{\rho}d_j^k\right|^2\geqslant\max_{i\in\mathcal{A}^k}\left|\beta_i^k+\frac{1}{\rho}d_i^k\right|^2=\max_{i\in\mathcal{A}^k}\left|\beta_i^k\right|^2.$$

Accordingly, after simplifying the algebraic expressions, the range of ρ is determined to be

$$\rho\in\left(0,\frac{\min_{j\in\mathcal{S}_{C,2}^k}\left|d_j^k\right|}{\max_{i\in\mathcal{A}^k}\left|\beta_i^k\right|}\right].$$

S3 Proofs of the Theorems

S3.1 Proof of Theorem 1

Let $(\hat{\boldsymbol{\beta}}, \hat{\mathcal{A}})$ denote the output solution of Algorithm 1 under the sparsity level s. Define

$$\mathcal{A}_1 = \hat{\mathcal{A}} \cap \mathcal{A}^*, \quad \mathcal{A}_2 = \hat{\mathcal{A}} \cap \mathcal{I}^*,$$
$$\mathcal{I}_1 = \hat{\mathcal{I}} \cap \mathcal{A}^*, \quad \mathcal{I}_2 = \hat{\mathcal{I}} \cap \mathcal{I}^*.$$

We employ a proof by contradiction. Suppose $\mathcal{I}_1 \neq \varnothing$ and $k_0 = |\mathcal{I}_1|$. Define the subset used for exchange in the active set as

$$\mathcal{S}_{k_0,1} = \left\{ j \in \hat{\mathcal{A}} : \sum_{i \in \hat{\mathcal{A}}} \mathrm{I}(\xi_j \geq \xi_i) \leq k_0 \right\},$$

We use $\mathcal{S}_1$ to simplify the notation of $\mathcal{S}_{k_0,1}$. Define

$$\mathcal{A}_{11} = \mathcal{A}_1 \cap (\mathcal{S}_1)^c, \quad \mathcal{A}_{12} = \mathcal{A}_1 \cap \mathcal{S}_1,$$
$$\mathcal{A}_{21} = \mathcal{A}_2 \cap (\mathcal{S}_1)^c, \quad \mathcal{A}_{22} = \mathcal{A}_2 \cap \mathcal{S}_1.$$

Let $\mathcal{S}_1 = \mathcal{A}_{12} \cup \mathcal{A}_{22}$. Since $s \geq |\mathcal{A}^*|$, we have

$$|\mathcal{A}_{12}| + |\mathcal{A}_{22}| = k_0 = |\mathcal{S}_1| \leq |\mathcal{A}_2| = |\mathcal{A}_{21}| + |\mathcal{A}_{22}|.$$

Based on $\mathcal{S}_1 = \mathcal{A}_{12} \cup \mathcal{A}_{22}$ and $(\mathcal{S}_1)^c = \mathcal{A}_{11} \cup \mathcal{A}_{21}$, we obtain

$$\max_{j \in \mathcal{A}_{12} \cup \mathcal{A}_{22}} \left. \frac{\partial^2 l_n(\boldsymbol{\beta})}{(\partial \beta_j)^2} \right|_{\hat{\beta}} (\hat{\beta}_j)^2 = \max_{j \in \mathcal{A}_{12} \cup \mathcal{A}_{22}} \xi_j \leq \min_{j \in \mathcal{A}_{11} \cup \mathcal{A}_{21}} \xi_j = \min_{j \in \mathcal{A}_{11} \cup \mathcal{A}_{21}} \left. \frac{\partial^2 l_n(\boldsymbol{\beta})}{(\partial \beta_j)^2} \right|_{\hat{\beta}} (\hat{\beta}_j)^2.$$

According to Assumption (A1), we have

$$N c_0 \leq \left. \frac{\partial^2 l_n(\boldsymbol{\beta})}{(\partial \beta_j)^2} \right|_{\hat{\boldsymbol{\beta}}} = \mathbf{X}_j^\top \Sigma(\boldsymbol{\beta}) \mathbf{X}_j \leq \frac{N}{c_0}.$$

Therefore, when $\mathcal{A}_{12} \neq \varnothing$,

$$\frac{c_0}{\sqrt{|\mathcal{A}_{12}|}} \left\| \hat{\boldsymbol{\beta}}_{\mathcal{A}_{12}} \right\|_2 \leq \frac{1}{\sqrt{|\mathcal{A}_{21}|}} \left\| \hat{\boldsymbol{\beta}}_{\mathcal{A}_{21}} \right\|_2.$$

We further perform a Taylor expansion of $l_n(\hat{\boldsymbol{\beta}})$ as follows:

$$
\begin{aligned}
l_n(\widehat{\boldsymbol{\beta}}) &= l_n(\boldsymbol{\beta}^*) + \left\{ \frac{\partial l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \bigg|_{\boldsymbol{\beta}^*} \right\}^\top (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) + \frac{1}{2} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^\top \frac{\partial^2 l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2} \bigg|_{\overline{\boldsymbol{\beta}}} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \\
&= l_n(\boldsymbol{\beta}^*) - \frac{1}{2} \left\{ \frac{\partial l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\widehat{\mathcal{A}} \cup \mathcal{A}^*}} \bigg|_{\boldsymbol{\beta}^*} \right\}^\top \left(\frac{\partial^2 l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2_{\widehat{\mathcal{A}} \cup \mathcal{A}^*}} \bigg|_{\overline{\boldsymbol{\beta}}} \right)^{-1} \frac{\partial l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\widehat{\mathcal{A}} \cup \mathcal{A}^*}} \bigg|_{\boldsymbol{\beta}^*} \\
&\quad + \frac{1}{2} \left((\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \Big|_{\widehat{\mathcal{A}} \cup \mathcal{A}^*} + \left(\frac{\partial^2 l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2_{\widehat{\mathcal{A}} \cup \mathcal{A}^*}} \bigg|_{\overline{\boldsymbol{\beta}}} \right)^{-1} \frac{\partial l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\widehat{\mathcal{A}} \cup \mathcal{A}^*}} \bigg|_{\boldsymbol{\beta}^*} \right)^\top \frac{\partial^2 l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2_{\widehat{\mathcal{A}} \cup \mathcal{A}^*}} \bigg|_{\overline{\boldsymbol{\beta}}} \\
&\quad \left((\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \Big|_{\widehat{\mathcal{A}} \cup \mathcal{A}^*} + \left(\frac{\partial^2 l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2_{\widehat{\mathcal{A}} \cup \mathcal{A}^*}} \bigg|_{\overline{\boldsymbol{\beta}}} \right)^{-1} \frac{\partial l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\widehat{\mathcal{A}} \cup \mathcal{A}^*}} \bigg|_{\boldsymbol{\beta}^*} \right),
\end{aligned}
$$

where $\overline{\boldsymbol{\beta}} = t\widehat{\boldsymbol{\beta}} + (1-t)\boldsymbol{\beta}^*$ with $0 \leq t \leq 1$. Since $\widehat{\boldsymbol{\beta}}$ minimizes $l_n(\boldsymbol{\beta})$ under the active set $\widehat{\mathcal{A}}$, we have

$$
(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \Big|_{\widehat{\mathcal{A}} \cup \mathcal{A}^*} = - \left(\frac{\partial^2 l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2_{\widehat{\mathcal{A}} \cup \mathcal{A}^*}} \bigg|_{\overline{\boldsymbol{\beta}}} \right)^{-1} \frac{\partial l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\widehat{\mathcal{A}} \cup \mathcal{A}^*}} \Bigg|_{\boldsymbol{\beta}^*}.
$$

Simplifying gives

$$
0 = \frac{\partial l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\widehat{\mathcal{A}}}} \bigg|_{\boldsymbol{\beta}^*} + \left\{ \frac{\partial^2 l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2_{\widehat{\mathcal{A}}}} \bigg|_{\overline{\boldsymbol{\beta}}} \right\}^\top (\hat{\boldsymbol{\beta}}_{\widehat{\mathcal{A}}} - \boldsymbol{\beta}^*_{\widehat{\mathcal{A}}}) + \left\{ \frac{\partial^2 l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\widehat{\mathcal{A}}} \partial \boldsymbol{\beta}_{\mathcal{I}_1}} \bigg|_{\overline{\boldsymbol{\beta}}} \right\} (-\boldsymbol{\beta}^*_{\mathcal{I}_1}).
$$

which immediately implies

$$
\begin{aligned}
\left\| \widehat{\boldsymbol{\beta}}_{\widehat{\mathcal{A}}} - \boldsymbol{\beta}^*_{\widehat{\mathcal{A}}} \right\|_2 &= \left\| \left\{ \frac{\partial^2 l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2_{\widehat{\mathcal{A}}}} \bigg|_{\overline{\boldsymbol{\beta}}} \right\}^{-1} \left\{ \frac{\partial l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\widehat{\mathcal{A}}}} \bigg|_{\boldsymbol{\beta}^*} + \left\{ \frac{\partial^2 l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\widehat{\mathcal{A}}} \partial \boldsymbol{\beta}_{\mathcal{I}_1}} \bigg|_{\overline{\boldsymbol{\beta}}} \right\} (-\boldsymbol{\beta}^*_{\mathcal{I}_1}) \right\} \right\|_2 \\
&\leq \frac{1}{N m_s c_0} \left\| \mathbf{X}^\top_{\widehat{\mathcal{A}}} (\mathbf{y} - b'(\mathbf{X}\boldsymbol{\beta}^*)) \right\|_2 + \frac{\nu_s}{m_s c_0^2} \left\| \boldsymbol{\beta}^*_{\mathcal{I}_1} \right\|_2 \\
&= \frac{1}{N m_s c_0} \left\| \mathbf{X}^\top_{\widehat{\mathcal{A}}} \mathbf{W} \right\|_2 + \frac{\nu_s}{m_s c_0^2} \left\| \boldsymbol{\beta}^*_{\mathcal{I}_1} \right\|_2.
\end{aligned}
$$

Therefore,

$$
\left\| \widehat{\boldsymbol{\beta}}_{\mathcal{A}_{12}} \right\|_2 \geq \left\| \boldsymbol{\beta}^*_{\mathcal{A}_{12}} \right\|_2 - \frac{1}{N m_s c_0} \left\| \mathbf{X}^\top_{\widehat{\mathcal{A}}} \mathbf{W} \right\|_2 - \frac{\nu_s}{m_s c_0} \left\| \boldsymbol{\beta}^*_{\mathcal{I}_1} \right\|_2,
$$

$$
\left\|\widehat{\boldsymbol{\beta}}_{\mathcal{A}_{21}}\right\|_2 \leq \left\|\boldsymbol{\beta}^*_{\mathcal{A}_{21}}\right\|_2 + \frac{1}{Nm_s c_0}\left\|\mathbf{X}^\top_{\widehat{\mathcal{A}}}\mathbf{W}\right\|_2 + \frac{\nu_s}{m_s c_0^2}\left\|\boldsymbol{\beta}^*_{\mathcal{I}_1}\right\|_2.
$$

That is,

$$
\begin{aligned}
\left\|\boldsymbol{\beta}^*_{\mathcal{A}_{12}}\right\|_2 &\leq \left(1+\frac{1}{c_0}\sqrt{\frac{|\mathcal{A}_{12}|}{|\mathcal{A}_{21}|}}\right)\left(\frac{1}{Nm_s c_0}\left\|\mathbf{X}^\top_{\widehat{\mathcal{A}}}\mathbf{W}\right\|_2 + \frac{\nu_s}{m_s c_0}\left\|\boldsymbol{\beta}^*_{\mathcal{I}_1}\right\|_2\right) \\
&\leq \frac{2}{c_0}\left(\frac{1}{Nm_s c_0}\left\|\mathbf{X}^\top_{\widehat{\mathcal{A}}}\mathbf{W}\right\|_2 + \frac{\nu_s}{m_s c_0^2}\left\|\boldsymbol{\beta}^*_{\mathcal{I}_1}\right\|_2\right).
\end{aligned}
$$

By Assumption (A2) on tail probability bounds, for any subset $|\mathcal{B}| \leq s$, we have

$$
\begin{aligned}
\mathbb{P}\left(\left\|\frac{1}{N}\mathbf{X}^\top_{\mathcal{B}}\mathbf{W}\right\|_2 \geq t\right) &\leq \mathbb{P}\left(\max_{|\mathcal{A}|\leq s}\left\|\frac{1}{\sqrt{N}}\mathbf{X}^\top_{\mathcal{A}}\mathbf{W}\right\|_2^2 \geq Nt^2\right) \\
&\leq \sum_{j=1}^{p}\mathbb{P}\left(\left\|\frac{1}{\sqrt{N}}\mathbf{X}^\top_j\mathbf{W}\right\|_2 \geq t\sqrt{\frac{N}{s}}\right) \\
&\leq c_1 p\exp\left\{-\frac{c_2 N t^2}{sM_1^2}\right\}.
\end{aligned}
$$

Consequently,

$$
\begin{aligned}
&\mathbb{P}\left(\frac{1}{Nm_s c_0}\left\|\mathbf{X}^\top_{\widehat{\mathcal{A}}}\mathbf{W}\right\|_2 \geq \frac{\Delta}{2}\frac{\nu_s}{m_s c_0^2}\left\|\boldsymbol{\beta}^*_{\mathcal{I}_1}\right\|_2\right) \\
&\leq c_1 p\exp\left\{-\frac{nc_2\Delta^2\nu_s^2\left\|\boldsymbol{\beta}^*_{\mathcal{I}_1}\right\|_2^2}{4c_0^2 sM_1^2}\right\} \leq \delta_1.
\end{aligned}
$$

That is, with probability at least $1-\delta_1$, $\left\|\boldsymbol{\beta}^*_{\mathcal{A}_{12}}\right\|_2 \leq \frac{1}{c_0}(2+\Delta)\frac{\nu_s}{m_s c_0^2}\left\|\boldsymbol{\beta}^*_{\mathcal{I}_1}\right\|_2$.

For the case of $\mathcal{A}_{12} = \varnothing$, we denote the subset used for exchange in the inactive set by

$$
\mathcal{S}_{k_0,2} = \left\{j \in \widehat{\mathcal{I}} : \sum_{i\in\widehat{\mathcal{I}}} \mathrm{I}(\zeta_j \leq \zeta_i) \leq k_0\right\},
$$

where $k_0 = |\mathcal{I}_1|$. To simplify the notation, we write $\mathcal{S}_2$ instead of $\mathcal{S}_{k_0,2}$. Define

$$
\begin{aligned}
\mathcal{I}_{11} = \mathcal{I}_1 \cap \mathcal{S}_2, \quad \mathcal{I}_{12} = \mathcal{I}_1 \cap (\mathcal{S}_2)^c, \\
\mathcal{I}_{21} = \mathcal{I}_2 \cap \mathcal{S}_2, \quad \mathcal{I}_{22} = \mathcal{I}_2 \cap (\mathcal{S}_2)^c.
\end{aligned}
$$

Since $|\mathcal{S}_2| = |\mathcal{I}_{11}| + |\mathcal{I}_{21}| = |\mathcal{I}_1|$ and $|\mathcal{I}_{12}| = |\mathcal{I}_{21}|$ with $\mathcal{I}_{12} \neq \varnothing$, by the definition of $\mathcal{S}_2$, we have

$$\min_{j \in \mathcal{I}_{21}} \left(\frac{\partial^2 l_n(\boldsymbol{\beta})}{(\partial \beta_j)^2} \Big|_{\widehat{\boldsymbol{\beta}}} \right)^{-1} (d_j)^2 \geq \max_{j \in \mathcal{I}_{12}} \left(\frac{\partial^2 l_n(\boldsymbol{\beta})}{(\partial \beta_j)^2} \Big|_{\widehat{\boldsymbol{\beta}}} \right)^{-1} (d_j)^2.$$

which implies

$$\min_{j \in \mathcal{I}_{21}} |d_j| \geq c_0 \max_{j \in \mathcal{I}_{12}} |d_j| .$$

Recal that

$$\widehat{d}_{\widehat{\mathcal{I}}} = \frac{\partial l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\widehat{\mathcal{I}}}} \Big|_{\boldsymbol{\beta}^*} + \left\{ \frac{\partial^2 l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\widehat{\mathcal{I}}} \partial \boldsymbol{\beta}_{\mathcal{I}_1}} \Big|_{\overline{\boldsymbol{\beta}}} \right\} (-\boldsymbol{\beta}^*_{\mathcal{I}_1}) + \left\{ \frac{\partial^2 l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\widehat{\mathcal{I}}} \partial \boldsymbol{\beta}_{\widehat{\mathcal{A}}}} \Big|_{\overline{\boldsymbol{\beta}}} \right\} \left(\widehat{\boldsymbol{\beta}}_{\widehat{\mathcal{A}}} - \boldsymbol{\beta}^*_{\widehat{\mathcal{A}}} \right).$$

Then,

$$\begin{aligned}
\left\| \widehat{d}_{\mathcal{I}_{12}} \right\|_2 &= \left\| \frac{\partial l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\mathcal{I}_{12}}} \Big|_{\boldsymbol{\beta}^*} + \left\{ \frac{\partial^2 l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\mathcal{I}_{12}} \partial \boldsymbol{\beta}_{\mathcal{I}_1}} \Big|_{\overline{\boldsymbol{\beta}}} \right\} (-\boldsymbol{\beta}^*_{\mathcal{I}_1}) + \left\{ \frac{\partial^2 l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\mathcal{I}_{12}} \partial \boldsymbol{\beta}_{\widehat{\mathcal{A}}}} \Big|_{\overline{\boldsymbol{\beta}}} \right\} \left(\widehat{\boldsymbol{\beta}}_{\widehat{\mathcal{A}}} - \boldsymbol{\beta}^*_{\widehat{\mathcal{A}}} \right) \right\|_2 \\
&\geq N c_0 m_s \left\| \boldsymbol{\beta}^*_{\mathcal{I}_{12}} \right\|_2 - \left\| \frac{\partial l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\mathcal{I}_{12}}} \Big|_{\boldsymbol{\beta}^*} \right\|_2 - \frac{N \nu_s}{c_0} \left\| \widehat{\boldsymbol{\beta}}_{\widehat{\mathcal{A}}} - \boldsymbol{\beta}^*_{\widehat{\mathcal{A}}} \right\|_2 - \frac{N \nu_s}{c_0} \left\| \boldsymbol{\beta}^*_{\mathcal{I}_{11}} \right\|_2 \\
&\geq N c_0 m_s \left\| \boldsymbol{\beta}^*_{\mathcal{I}_{12}} \right\|_2 - \left\| \mathbf{X}^\top_{\mathcal{I}_{12}} \mathbf{W} \right\|_2 - \frac{\nu_s}{m_s c_0^2} \left\| \mathbf{X}^\top_{\widehat{\mathcal{A}}} \mathbf{W} \right\|_2 \\
&- \frac{N \nu_s^2}{m_s c_0^3} \left\| \boldsymbol{\beta}^*_{\mathcal{I}_1} \right\|_2 - \frac{N \nu_s}{c_0} \left\| \boldsymbol{\beta}^*_{\mathcal{I}_{11}} \right\|_2 ,
\end{aligned}$$

and

$$\begin{aligned}
\left\| \widehat{d}_{\mathcal{I}_{21}} \right\|_2 &= \left\| \frac{\partial l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\mathcal{I}_{21}}} \Big|_{\boldsymbol{\beta}^*} + \left\{ \frac{\partial^2 l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\mathcal{I}_{21}} \partial \boldsymbol{\beta}_{\mathcal{I}_1}} \Big|_{\overline{\boldsymbol{\beta}}} \right\} (-\boldsymbol{\beta}^*_{\mathcal{I}_1}) + \left\{ \frac{\partial^2 l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\mathcal{I}_{21}} \partial \boldsymbol{\beta}_{\widehat{\mathcal{A}}}} \Big|_{\overline{\boldsymbol{\beta}}} \right\} \left(\widehat{\boldsymbol{\beta}}_{\widehat{\mathcal{A}}} - \boldsymbol{\beta}^*_{\widehat{\mathcal{A}}} \right) \right\|_2 \\
&\leq \left\| \mathbf{X}^\top_{\mathcal{I}_{21}} \mathbf{W} \right\|_2 + \frac{N \nu_s}{c_0} \left\| \boldsymbol{\beta}^*_{\mathcal{I}_1} \right\|_2 + \frac{N \nu_s}{c_0} \left\| \widehat{\boldsymbol{\beta}}_{\widehat{\mathcal{A}}} - \boldsymbol{\beta}^*_{\widehat{\mathcal{A}}} \right\|_2 \\
&\leq \left\| \mathbf{X}^\top_{\mathcal{I}_{21}} \mathbf{W} \right\|_2 + \frac{N \nu_s}{c_0} \left\| \boldsymbol{\beta}^*_{\mathcal{I}_1} \right\|_2 + \frac{\nu_s}{m_s c_0^2} \left\| \frac{1}{N} \mathbf{X}^\top_{\widehat{\mathcal{A}}} \mathbf{W} \right\|_2 + \frac{N \nu_s^2}{m_s c_0^3} \left\| \boldsymbol{\beta}^*_{\mathcal{I}_1} \right\|_2 .
\end{aligned}$$

Since $|\mathcal{I}_{12}| = |\mathcal{I}_{21}|$, it follows that $\frac{1}{c_0} \left\| \hat{d}_{\mathcal{I}_{21}} \right\|_2 \geq \left\| \hat{d}_{\mathcal{I}_{12}} \right\|_2$, which implies

$$\begin{aligned}
&\frac{2}{c_0} \frac{\nu_s}{c_0} \left\| \boldsymbol{\beta}^*_{\mathcal{I}_1} \right\|_2 + \frac{2}{c_0} \frac{\nu_s^2}{m_s c_0^3} \left\| \boldsymbol{\beta}^*_{\mathcal{I}_1} \right\|_2 \\
\geq & c_0 m_s \left\| \boldsymbol{\beta}^*_{\mathcal{I}_{12}} \right\|_2 - \frac{2}{c_0} \frac{\nu_s}{m_s c_0^2} \left\| \frac{1}{N} \mathbf{X}^\top_{\widehat{\mathcal{A}}} \mathbf{W} \right\|_2 - \left\| \frac{1}{N} \mathbf{X}^\top_{\mathcal{I}_{12}} \mathbf{W} \right\|_2 - \frac{1}{c_0} \left\| \frac{1}{N} \mathbf{X}^\top_{\mathcal{I}_{21}} \mathbf{W} \right\|_2 .
\end{aligned}$$

By Assumption 2, we have

$$\mathbb{P}\left(c_0\left\|\frac{1}{N}\mathbf{X}_{\mathcal{I}_{12}}^\top\mathbf{W}\right\|_2 \geq \frac{\Delta}{3}\frac{\nu_s}{c_0}\left\|\boldsymbol{\beta}_{\mathcal{I}_1}^*\right\|_2\right) \leq c_1 p\exp\left\{-\frac{Nc_2\Delta^2\nu_s^2\left\|\boldsymbol{\beta}_{\mathcal{I}_1}^*\right\|_2^2}{9c_0^4 sM_1^2}\right\} \leq \delta_2/3,$$

$$\mathbb{P}\left(\frac{2\nu_s}{m_s c_0^2}\left\|\frac{1}{N}\mathbf{X}_{\widehat{\mathcal{A}}}^\top\mathbf{W}\right\|_2 \geq \frac{\Delta}{3}\frac{\nu_s}{c_0}\left\|\boldsymbol{\beta}_{\mathcal{I}_1}^*\right\|_2\right) \leq c_1 p\exp\left\{-\frac{N\Delta^2 c_2 c_0^2 m_s^2\left\|\boldsymbol{\beta}_{\mathcal{I}_1}^*\right\|_2^2}{36 sM_1^2}\right\} \leq \delta_2/3,$$

$$\mathbb{P}\left(\left\|\frac{1}{N}\mathbf{X}_{\mathcal{I}_{21}}^\top\mathbf{W}\right\|_2 \geq \frac{\Delta}{3}\frac{\nu_s}{c_0}\left\|\boldsymbol{\beta}_{\mathcal{I}_1}^*\right\|_2\right) \leq c_1 p\exp\left\{-\frac{Nc_2\Delta^2\nu_s^2\left\|\boldsymbol{\beta}_{\mathcal{I}_1}^*\right\|_2^2}{9c_0^4 sM_1^2}\right\} \leq \delta_2/3.$$

Hence,

$$\mathbb{P}\left(\frac{1}{c_0}\left(2+\frac{2\nu_s}{m_s c_0^2}+\Delta\right)\frac{\nu_s}{c_0}\left\|\boldsymbol{\beta}_{\mathcal{I}_1}^*\right\|_2 \geq c_0 m_s\left\|\boldsymbol{\beta}_{\mathcal{I}_{12}}^*\right\|_2\right) \geq 1-\delta_2.$$

We then move $\mathcal{S}_1 = \mathcal{A}_{12}\cup\mathcal{A}_{22}$ to the inactive set and $\mathcal{S}_2 = \mathcal{I}_{11}\cup\mathcal{I}_{21}$ to the active set. Denote the updated active set by $\tilde{\mathcal{A}} = \left(\widehat{\mathcal{A}}\backslash\mathcal{S}_1\right)\cup\mathcal{S}_2$, and the updated inactive set by $\tilde{\mathcal{I}} = (\tilde{\mathcal{A}})^c$. Similarly, let $\tilde{\mathcal{I}}_1 = \tilde{\mathcal{I}}\cap\mathcal{A}^*$. Since $\tilde{\boldsymbol{\beta}} = \arg\min_{\sup\mathbb{P}(\boldsymbol{\beta})=\tilde{\mathcal{A}}} l_n(\boldsymbol{\beta})$, it follows that

$$\left\|\tilde{\boldsymbol{\beta}}_{\tilde{\mathcal{A}}}-\boldsymbol{\beta}_{\tilde{\mathcal{A}}}^*\right\|_2 \leq \frac{1}{Nm_s c_0}\left\|\mathbf{X}_{\tilde{\mathcal{A}}}^\top\mathbf{W}\right\|_2+\frac{\nu_s}{m_s c_0}\left\|\boldsymbol{\beta}_{\tilde{\mathcal{I}}_1}^*\right\|_2.$$

The loss function at $\tilde{\boldsymbol{\beta}}$ can be expanded as

$$\begin{aligned}
l_n(\tilde{\boldsymbol{\beta}}) = l_n\left(\boldsymbol{\beta}^*\right) &+ \left(\left.\frac{\partial l_n(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}_{\tilde{\mathcal{A}}}}\right|_{\boldsymbol{\beta}^*}\right)^\top\left(\tilde{\boldsymbol{\beta}}_{\tilde{\mathcal{A}}}-\boldsymbol{\beta}_{\tilde{\mathcal{A}}}^*\right)+\left(\left.\frac{\partial l_n(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}_{\tilde{\mathcal{I}}_1}}\right|_{\boldsymbol{\beta}^*}\right)^\top\left(-\boldsymbol{\beta}_{\tilde{\mathcal{I}}_1}^*\right)\\
&+\frac{1}{2}\left(\tilde{\boldsymbol{\beta}}_{\tilde{\mathcal{A}}}-\boldsymbol{\beta}_{\tilde{\mathcal{A}}}^*\right)^\top\left.\frac{\partial^2 l_n(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}_{\tilde{\mathcal{A}}}\partial\boldsymbol{\beta}_{\tilde{\mathcal{A}}}}\right|_{\overline{\boldsymbol{\beta}}}\left(\tilde{\boldsymbol{\beta}}_{\tilde{\mathcal{A}}}-\boldsymbol{\beta}_{\tilde{\mathcal{A}}}^*\right)\\
&+\left(-\boldsymbol{\beta}_{\tilde{\mathcal{I}}_1}^*\right)^\top\left.\frac{\partial^2 l_n(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}_{\tilde{\mathcal{I}}_1}\partial\boldsymbol{\beta}_{\tilde{\mathcal{A}}}}\right|_{\overline{\boldsymbol{\beta}}}\left(\tilde{\boldsymbol{\beta}}_{\tilde{\mathcal{A}}}-\boldsymbol{\beta}_{\tilde{\mathcal{A}}}^*\right)\\
&+\frac{1}{2}\left(-\boldsymbol{\beta}_{\tilde{\mathcal{I}}_1}^*\right)^\top\left.\frac{\partial^2 l_n(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}_{\tilde{\mathcal{I}}_1}\partial\boldsymbol{\beta}_{\tilde{\mathcal{I}}_1}}\right|_{\tilde{\boldsymbol{\beta}}}\left(-\boldsymbol{\beta}_{\tilde{\mathcal{I}}_1}^*\right).
\end{aligned}$$

Thus,

$$
\begin{aligned}
l_n(\tilde{\boldsymbol{\beta}}) - l_n\left(\boldsymbol{\beta}^*\right) \leq & \frac{N M_s}{2 c_0}\left\|\boldsymbol{\beta}_{\tilde{\mathcal{I}}_1}^*\right\|_2^2+\frac{N \nu_s}{c_0}\left\|\boldsymbol{\beta}_{\tilde{\mathcal{I}}_1}^*\right\|_2\left\|\tilde{\boldsymbol{\beta}}_{\tilde{\mathcal{A}}}-\boldsymbol{\beta}_{\tilde{\mathcal{A}}}^*\right\|_2+\frac{N M_s}{2 c_0}\left\|\tilde{\boldsymbol{\beta}}_{\tilde{\mathcal{A}}}-\boldsymbol{\beta}_{\tilde{\mathcal{A}}}^*\right\|_2^2 \\
& +\left|\left(\left.\frac{\partial l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\tilde{\mathcal{I}}_1}}\right|_{\boldsymbol{\beta}^*}\right)^{\top}\left(-\boldsymbol{\beta}_{\tilde{\mathcal{I}}_1}^*\right)\right|+\left|\left(\left.\frac{\partial l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\tilde{\mathcal{A}}}}\right|_{\boldsymbol{\beta}^*}\right)^{\top}\left(\tilde{\boldsymbol{\beta}}_{\tilde{\mathcal{A}}}-\boldsymbol{\beta}_{\tilde{\mathcal{A}}}^*\right)\right| \\
\leq & \left(\frac{N M_s}{2 c_0}+\frac{N \nu_s^2}{m_s c_0^3}+\frac{N M_s \nu_s^2}{2 m_s^2 c_0^5}\right)\left\|\boldsymbol{\beta}_{\mathcal{I}_{12} \cup \mathcal{A}_{12}}^*\right\|_2^2+\varepsilon(\tilde{\mathcal{A}}),
\end{aligned}
$$

where

$$
\begin{aligned}
\varepsilon(\mathcal{A})= & \left(\frac{\nu_s M_s}{m_s^2 c_0^4}+\frac{2 v_s}{m_s c_0^2}\right)\left\|\boldsymbol{\beta}_{\mathcal{A}^* \cap \mathcal{A}^c}^*\right\|_2\left\|\mathbf{X}_{\mathcal{A}}^{\top} \mathbf{W}\right\|_2+\left\|\boldsymbol{\beta}_{\mathcal{A}^* \cap \mathcal{A}^c}^*\right\|_2\left\|\mathbf{X}_{\mathcal{A}^* \cap \mathcal{A}^c}^{\top} \mathbf{W}\right\|_2 \\
& +\left(\frac{M_s}{2 N m_s^2 c_0^3}+\frac{1}{N m_s c_0}\right)\left\|\mathbf{X}_{\mathcal{A}}^{\top} \mathbf{W}\right\|_2^2 .
\end{aligned}
$$

From the previous derivations, we thus obtain

$$
\begin{aligned}
& l_n(\tilde{\boldsymbol{\beta}})-l_n\left(\boldsymbol{\beta}^*\right) \\
\leq & \frac{N}{c_0^2}\left(\frac{M_s}{2 c_0}+\frac{\nu_s^2}{m_s c_0^3}+\frac{M_s v_s^2}{2 m_s^2 c_0^5}\right)\left[\left(2+\frac{2 \nu_s}{m_s c_0^2}+\Delta\right)^2+(2+\Delta)^2\right] \\
& \left(\frac{\nu_s}{m_s c_0^2}\right)^2\left\|\boldsymbol{\beta}_{\mathcal{I}_1}^*\right\|_2^2+\varepsilon(\tilde{\mathcal{A}}).
\end{aligned}
$$

Similarly, the loss function at $\hat{\boldsymbol{\beta}}$ can be expanded as

$$
\begin{aligned}
l_n(\widehat{\boldsymbol{\beta}})= & l_n\left(\boldsymbol{\beta}^*\right)+\left(\left.\frac{\partial l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\widehat{\mathcal{A}}}}\right|_{\boldsymbol{\beta}^*}\right)^{\top}\left(\widehat{\boldsymbol{\beta}}_{\widehat{\mathcal{A}}}-\boldsymbol{\beta}_{\widehat{\mathcal{A}}}^*\right)+\left(\left.\frac{\partial l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\mathcal{I}_1}}\right|_{\boldsymbol{\beta}^*}\right)^{\top}\left(-\boldsymbol{\beta}_{\mathcal{I}_1}^*\right) \\
& +\frac{1}{2}\left(\widehat{\boldsymbol{\beta}}_{\widehat{\mathcal{A}}}-\boldsymbol{\beta}_{\widehat{\mathcal{A}}}^*\right)^{\top}\left.\frac{\partial^2 l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\widehat{\mathcal{A}}} \partial \boldsymbol{\beta}_{\widehat{\mathcal{A}}}}\right|_{\overline{\boldsymbol{\beta}}}\left(\widehat{\boldsymbol{\beta}}_{\widehat{\mathcal{A}}}-\boldsymbol{\beta}_{\widehat{\mathcal{A}}}^*\right) \\
& +\left(-\boldsymbol{\beta}_{\mathcal{I}_1}^*\right)^{\top}\left.\frac{\partial^2 l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\mathcal{I}_1} \partial \boldsymbol{\beta}_{\widehat{\mathcal{A}}}}\right|_{\overline{\boldsymbol{\beta}}}\left(\widehat{\boldsymbol{\beta}}_{\widehat{\mathcal{A}}}-\boldsymbol{\beta}_{\widehat{\mathcal{A}}}^*\right) \\
& +\frac{1}{2}\left(-\boldsymbol{\beta}_{\mathcal{I}_1}^*\right)^{\top}\left.\frac{\partial^2 l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\mathcal{I}_1} \partial \boldsymbol{\beta}_{\mathcal{I}_1}}\right|_{\overline{\boldsymbol{\beta}}}\left(-\boldsymbol{\beta}_{\mathcal{I}_1}^*\right) .
\end{aligned}
$$

Consequently, we obtain a lower bound:

$$
\begin{aligned}
& l_n(\widehat{\boldsymbol{\beta}}) - l_n\left(\boldsymbol{\beta}^*\right) \\
\geq & \frac{N m_s c_0}{2}\left\|\boldsymbol{\beta}_{\mathcal{I}_1}^*\right\|_2^2 - \frac{N \nu_s}{c_0}\left\|\boldsymbol{\beta}_{\mathcal{I}_1}^*\right\|_2\left\|\widehat{\boldsymbol{\beta}}_{\widehat{\mathcal{A}}} - \boldsymbol{\beta}_{\widehat{\mathcal{A}}}^*\right\|_2 - \frac{N M_s}{2 c_0}\left\|\widehat{\boldsymbol{\beta}}_{\widehat{\mathcal{A}}} - \boldsymbol{\beta}_{\widehat{\mathcal{A}}}^*\right\|_2^2 \\
& - \left|\left(\left.\frac{\partial l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\mathcal{I}_1}}\right|_{\boldsymbol{\beta}^*}\right)^{\top}\left(-\boldsymbol{\beta}_{\mathcal{I}_1}^*\right)\right| - \left|\left(\left.\frac{\partial l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\widehat{\mathcal{A}}}}\right|_{\boldsymbol{\beta}^*}\right)^{\top}\left(\widehat{\boldsymbol{\beta}}_{\widehat{\mathcal{A}}} - \boldsymbol{\beta}_{\widehat{\mathcal{A}}}^*\right)\right| \\
\geq & \left(\frac{N m_s c_0}{2} - \frac{N \nu_s^2}{m_s c_0^3} - \frac{N M_s v_s^2}{2 m_s^2 c_0^5}\right)\left\|\boldsymbol{\beta}_{\mathcal{I}_1}^*\right\|_2^2 - \varepsilon(\widehat{\mathcal{A}}).
\end{aligned}
$$

By Assumption (A4), it follows that

$$
\begin{aligned}
& (1-\Delta)\left(\frac{m_s c_0}{2} - \frac{\nu_s^2}{m_s c_0^3} - \frac{M_s v_s^2}{2 m_s^2 c_0^5}\right)\left\|\boldsymbol{\beta}_{\mathcal{I}_1}^*\right\|_2^2 \\
\geq & \frac{1}{c_0^2}\left(\frac{M_s}{2 c_0} + \frac{\nu_s^2}{m_s c_0^3} + \frac{M_s v_s^2}{2 m_s^2 c_0^5}\right)\left[\left(2+\Delta+\frac{2 \nu_s}{m_s c_0^2}\right)^2 + (2+\Delta)^2\right]\left(\frac{\nu_s}{m_s c_0^2}\right)^2\left\|\boldsymbol{\beta}_{\mathcal{I}_1}^*\right\|_2^2 .
\end{aligned}
$$

Using probabilistic bounds on the remainder terms, we have

$$
\begin{aligned}
& \mathbb{P}\left(\varepsilon(\tilde{\mathcal{A}}) + \varepsilon(\widehat{\mathcal{A}}) \geq N \Delta\left(\frac{m_s c_0}{2} - \frac{\nu_s^2}{m_s c_0^3} - \frac{M_s v_s^2}{2 m_s^2 c_0^5}\right)\left\|\boldsymbol{\beta}_{\mathcal{I}_1}^*\right\|_2^2\right) \\
\leq & c_1 p \exp\left\{-N C c_2 \Delta^2\left(\frac{m_s c_0}{2} - \frac{\nu_s^2}{m_s c_0^3} - \frac{M_s v_s^2}{2 m_s^2 c_0^5}\right)^2\left\|\boldsymbol{\beta}_{\mathcal{I}_1}^*\right\|_2^2 / \left(s M_1^2\right)\right\} \\
\leq & \delta_3 .
\end{aligned}
$$

where C is a constant. Therefore, with probability at least $1-\delta_1-\delta_2-\delta_3$, we have

$$
\begin{aligned}
l_n(\widehat{\boldsymbol{\beta}}) - l_n(\tilde{\boldsymbol{\beta}}) & > \left(1-\gamma_s\right)(1-\Delta)\left(\frac{N m_s c_0}{2} - \frac{N \nu_s^2}{m_s c_0^3} - \frac{N M_s v_s^2}{2 m_s^2 c_0^5}\right)\left\|\boldsymbol{\beta}_{\mathcal{I}_1}^*\right\|_2^2 \\
& > C_2\left(1-\gamma_s\right)(1-\Delta) N \min_{j \in \mathcal{A}^*}\left|\beta_j^*\right|^2 .
\end{aligned}
$$

Finally, by Assumptions (A5) and (A6), for sufficiently large N, we obtain

$$
l_n(\widehat{\boldsymbol{\beta}}) - l_n(\tilde{\boldsymbol{\beta}}) > \tau_s,
$$

which contradicts the initial assumption that $\mathcal{I}_1 \neq \varnothing$. Hence, we conclude

$$\mathbb{P}\left(\widehat{\mathcal{A}} \supseteq \mathcal{A}^*\right) \geq 1 - \delta_1 - \delta_2 - \delta_3,$$

completing the proof of the theorem.

S3.2 Proof of Theorem 2

Let $(\hat{\boldsymbol{\beta}}^s, \hat{\mathcal{A}}^s)$ denote the output of Algorithm 1 when the support size is s, where $\hat{\boldsymbol{\beta}}^s = \operatorname{argmin}_{\boldsymbol{\beta}_{(\hat{\mathcal{A}}^s)}=0} l_n(\boldsymbol{\beta})$. For notational simplicity, define $\hat{\boldsymbol{\beta}}^* = \operatorname{argmin}_{\boldsymbol{\beta}_{(\mathcal{A}^*)^c=0}} l_n(\boldsymbol{\beta})$ as the optimal estimate, and let $\delta = O\left(p^{-\alpha}\right)$. First, consider the case where $s < s^*$. Define

$$\begin{aligned}
\mathcal{A}_1^s &= \hat{\mathcal{A}}^s \cap \mathcal{A}^*, \quad \mathcal{A}_2^s = \hat{\mathcal{A}}^s \cap \mathcal{I}^*, \\
\mathcal{I}_1^s &= \left(\hat{\mathcal{A}}^s\right)^c \cap \mathcal{A}^*, \quad \mathcal{I}_2^s = \left(\hat{\mathcal{A}}^s\right)^c \cap \mathcal{I}^*.
\end{aligned}$$

Then, the loss function can be written as

$$\begin{aligned}
l_n\left(\hat{\boldsymbol{\beta}}^s\right) =& l_n\left(\boldsymbol{\beta}^*\right) + \left\{\left.\frac{\partial l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\hat{\mathcal{A}}^s}}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}\right\}^\top \left(\hat{\boldsymbol{\beta}}_{\hat{\mathcal{A}}^s} - \boldsymbol{\beta}^*_{\hat{\mathcal{A}}^s}\right) + \left\{\left.\frac{\partial l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\mathcal{I}_1^s}}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}\right\}^\top \left(-\boldsymbol{\beta}^*_{\mathcal{I}_1^s}\right) \\
&+ \frac{1}{2}\left(\hat{\boldsymbol{\beta}}_{\hat{\mathcal{A}}^s} - \boldsymbol{\beta}^*_{\hat{\mathcal{A}}^s}\right)^\top \left.\frac{\partial^2 l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\hat{\mathcal{A}}^s} \partial \boldsymbol{\beta}_{\hat{\mathcal{A}}^s}}\right|_{\boldsymbol{\beta}=\overline{\boldsymbol{\beta}}} \left(\hat{\boldsymbol{\beta}}_{\hat{\mathcal{A}}^s} - \boldsymbol{\beta}^*_{\hat{\mathcal{A}}^s}\right) \\
&+ \left(-\boldsymbol{\beta}^*_{\mathcal{I}_1^s}\right)^\top \left.\frac{\partial^2 l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\mathcal{I}_1^s} \partial \boldsymbol{\beta}_{\hat{\mathcal{A}}^s}}\right|_{\boldsymbol{\beta}=\overline{\boldsymbol{\beta}}} \left(\hat{\boldsymbol{\beta}}_{\hat{\mathcal{A}}^s} - \boldsymbol{\beta}^*_{\hat{\mathcal{A}}^s}\right) + \frac{1}{2}\left(-\boldsymbol{\beta}^*_{\mathcal{I}_1^s}\right)^\top \left.\frac{\partial^2 l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\mathcal{I}_1^s} \partial \boldsymbol{\beta}_{\mathcal{I}_1^s}}\right|_{\boldsymbol{\beta}=\overline{\boldsymbol{\beta}}} \left(-\boldsymbol{\beta}^*_{\mathcal{I}_1^s}\right).
\end{aligned}$$

Clearly, for $0 < \Delta < 1/2$, we have

$$l_n\left(\hat{\boldsymbol{\beta}}^s\right) - l_n\left(\boldsymbol{\beta}^*\right) \geq N(1-\Delta)\left(\frac{c_0 m_s}{2} - \frac{\nu_s^2}{m_s c_0^3} - \frac{\nu_s^2 M_s}{2 m_s^2 c_0^5}\right)\left\|\boldsymbol{\beta}^*_{\mathcal{I}_1^s}\right\|_2^2.$$

Moreover, since $\left\|\widehat{\boldsymbol{\beta}}^*_{\mathcal{A}^*}-\boldsymbol{\beta}^*_{\mathcal{A}^*}\right\|_2 \leq \frac{1}{Nm_sc_0}\left\|\boldsymbol{X}_{\mathcal{A}^*}^\top\boldsymbol{W}\right\|_2$, it follows that

$$
\begin{aligned}
l_n\left(\hat{\boldsymbol{\beta}}^*\right)-l_n\left(\boldsymbol{\beta}^*\right)=&\left\{\left.\frac{\partial l_n(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}_{\mathcal{A}^*}}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}\right\}^\top\left(\hat{\boldsymbol{\beta}}^*_{\mathcal{A}^*}-\boldsymbol{\beta}^*_{\mathcal{A}^*}\right)\\
&+\frac{1}{2}\left(\hat{\boldsymbol{\beta}}^*_{\mathcal{A}^*}-\boldsymbol{\beta}^*_{\mathcal{A}^*}\right)^\top\left.\frac{\partial^2 l_n(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}_{\mathcal{A}^*}\partial\boldsymbol{\beta}_{\mathcal{A}^*}}\right|_{\boldsymbol{\beta}=\overline{\boldsymbol{\beta}}}\left(\hat{\boldsymbol{\beta}}^*_{\mathcal{A}^*}-\boldsymbol{\beta}^*_{\mathcal{A}^*}\right)\\
\leq&\left(\frac{1}{Nm_sc_0}+\frac{M_s}{Nm_s^2c_0^3}\right)\left\|\boldsymbol{X}_{\mathcal{A}^*}^\top\boldsymbol{W}\right\|_2^2.
\end{aligned}
$$

Consequently,

$$
\mathbb{P}\left(\left(\frac{1}{Nm_sc_0}+\frac{M_s}{Nm_s^2c_0^3}\right)\left\|\boldsymbol{X}_{\mathcal{A}^*}^\top\boldsymbol{W}\right\|_2^2\geq N\Delta\left(\frac{c_0m_s}{2}-\frac{\nu_s^2}{m_sc_0^3}-\frac{\nu_s^2M_s}{m_s^2c_0^5}\right)\left\|\boldsymbol{\beta}^*_{\mathcal{I}_1^s}\right\|_2^2\right)\leq\delta.
$$

Therefore, with probability at least $1-\delta$, $F\left(\hat{\boldsymbol{\beta}}^s\right)-F\left(\hat{\boldsymbol{\beta}}^*\right)>0$. For sufficiently large N, by Assumption (A5), we have

$$
\frac{1}{\min_{j\in\mathcal{A}^*}\left|\beta_j^*\right|^2}=o\left(\frac{N}{s_{\max}\log p\log\log N}\right).
$$

Next, consider the case where $s>s^*$. By Theorem 1, with probability at least $1-O\left(\exp\left\{\log p-NC\left(\min_{j\in\mathcal{A}^*}\beta_j^*\right)^2/s\right\}\right)$, we have $\hat{\mathcal{A}}^s\supseteq\mathcal{A}^*$, $s\geq s^*$. Let $\hat{\mathcal{A}}^s=\mathcal{A}^*\cup\mathcal{B}^s$. Since

$$
0=\left.\frac{\partial l_n(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}_{\mathcal{A}^*}}\right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}^*}=\left.\frac{\partial l_n(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}_{\mathcal{A}^*}}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}+\left\{\left.\frac{\partial^2 l_n(\boldsymbol{\beta})}{\left(\partial\boldsymbol{\beta}_{\mathcal{A}^*}\right)^2}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}\right\}^\top\left(\hat{\boldsymbol{\beta}}^*_{\mathcal{A}^*}-\boldsymbol{\beta}^*_{\mathcal{A}^*}\right)+R_3,
$$

where $R_3=\frac{1}{6}\sum_{i=1}^n(b'''\left(\boldsymbol{x}_i\overline{\boldsymbol{\beta}}\right)(\boldsymbol{x}_i^\top(\hat{\boldsymbol{\beta}}^*-\boldsymbol{\beta}^*))^3)$ is a higher-order term, and $\overline{\boldsymbol{\beta}}=t\hat{\boldsymbol{\beta}}^*+(1-t)\boldsymbol{\beta}^*$ for some $t\in[0,1]$. It follows that

$$
\hat{\boldsymbol{\beta}}^*_{\mathcal{A}^*}-\boldsymbol{\beta}^*_{\mathcal{A}^*}=-\left\{\left.\frac{\partial^2 l_n(\boldsymbol{\beta})}{\left(\partial\boldsymbol{\beta}_{\mathcal{A}^*}\right)^2}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}\right\}^{-1}\left(\left.\frac{\partial l_n(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}_{\mathcal{A}^*}}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}+R_3\right).
$$

Applying the Taylor expansion, we have

$$
\begin{aligned}
l_n\left(\hat{\boldsymbol{\beta}}^*\right) - l_n\left(\boldsymbol{\beta}^*\right) = & \left\{ \frac{\partial l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\mathcal{A}^*}}\bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} \right\}^\top \left(\hat{\boldsymbol{\beta}}^*_{\mathcal{A}^*} - \boldsymbol{\beta}^*_{\mathcal{A}^*}\right) + \\
& \frac{1}{2}\left(\hat{\boldsymbol{\beta}}^*_{\mathcal{A}^*} - \boldsymbol{\beta}^*_{\mathcal{A}^*}\right)^\top \frac{\partial^2 l_n(\boldsymbol{\beta})}{\left(\partial \boldsymbol{\beta}_{\mathcal{A}^*}\right)^2}\bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} \left(\hat{\boldsymbol{\beta}}^*_{\mathcal{A}^*} - \boldsymbol{\beta}^*_{\mathcal{A}^*}\right) + R_3^{(1)} \\
= & -\frac{1}{2}\left\{ \frac{\partial l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\mathcal{A}^*}}\bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} \right\}^\top \left(\frac{\partial^2 l_n(\boldsymbol{\beta})}{\left(\partial \boldsymbol{\beta}_{\mathcal{A}^*}\right)^2}\bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} \right)^{-1} \frac{\partial l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\mathcal{A}^*}}\bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} \\
& + f_1\left(R_3, R_3^{(1)}\right) \\
= & -\frac{1}{2}\boldsymbol{W}^\top \boldsymbol{X}_{\mathcal{A}^*}\left(\boldsymbol{X}_{\mathcal{A}^*}^\top \boldsymbol{\Sigma}\left(\boldsymbol{\beta}^*\right)\boldsymbol{X}_{\mathcal{A}^*}\right)^{-1}\boldsymbol{X}_{\mathcal{A}^*}^\top \boldsymbol{W} + f_1\left(R_3, R_3^{(1)}\right),
\end{aligned}
$$

where

$$
f_1\left(R_3, R_3^{(1)}\right) = R_3^\top \left(\frac{\partial^2 l_n(\boldsymbol{\beta})}{\left(\partial \boldsymbol{\beta}_{\mathcal{A}^*}\right)^2}\bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} \right)^{-1} R_3 + 2R_3^\top \left(\frac{\partial^2 l_n(\boldsymbol{\beta})}{\left(\partial \boldsymbol{\beta}_{\mathcal{A}^*}\right)^2}\bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} \right)^{-1} \frac{\partial l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\mathcal{A}^*}}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} + R_3^{(1)},
$$

and $R_3^{(1)}$ is a higher-order term. Define $\boldsymbol{Z}_{\mathcal{A}} = \boldsymbol{\Sigma}\left(\boldsymbol{\beta}^*\right)^{\frac{1}{2}}\boldsymbol{X}_{\mathcal{A}}$. Then

$$
\begin{aligned}
l_n\left(\hat{\boldsymbol{\beta}}^*\right) - l_n\left(\boldsymbol{\beta}^*\right) &= -\frac{1}{2}\boldsymbol{W}^\top \boldsymbol{\Sigma}\left(\boldsymbol{\beta}^*\right)^{-\frac{1}{2}}\boldsymbol{Z}_{\mathcal{A}^*}\left(\boldsymbol{Z}_{\mathcal{A}^*}^\top \boldsymbol{Z}_{\mathcal{A}^*}\right)^{-1}\boldsymbol{Z}_{\mathcal{A}^*}^\top \boldsymbol{\Sigma}\left(\boldsymbol{\beta}^*\right)^{-\frac{1}{2}}\boldsymbol{W} + f_1\left(R_3, R_3^{(1)}\right) \\
&= -\frac{1}{2}\boldsymbol{W}^\top \boldsymbol{\Sigma}\left(\boldsymbol{\beta}^*\right)^{-\frac{1}{2}}\boldsymbol{H}_{\boldsymbol{Z}_{\mathcal{A}^*}}\boldsymbol{\Sigma}\left(\boldsymbol{\beta}^*\right)^{-\frac{1}{2}}\boldsymbol{W} + f_1\left(R_3, R_3^{(1)}\right),
\end{aligned}
$$

where $\boldsymbol{H}_{\boldsymbol{Z}_{\mathcal{A}^*}} = \boldsymbol{Z}_{\mathcal{A}^*}\left(\boldsymbol{Z}_{\mathcal{A}^*}^\top \boldsymbol{Z}_{\mathcal{A}^*}\right)^{-1}\boldsymbol{Z}_{\mathcal{A}^*}^\top$ is the projection matrix onto the column space of $\boldsymbol{Z}_{\mathcal{A}^*}$. Similarly, we obtain

$$
l_n\left(\hat{\boldsymbol{\beta}}^s\right) - l_n\left(\boldsymbol{\beta}^*\right) = -\frac{1}{2}\boldsymbol{W}^\top \boldsymbol{\Sigma}\left(\boldsymbol{\beta}^*\right)^{-\frac{1}{2}}\boldsymbol{H}_{\boldsymbol{Z}_{\mathcal{A}^s}}\boldsymbol{\Sigma}\left(\boldsymbol{\beta}^*\right)^{-\frac{1}{2}}\boldsymbol{W} + f_2\left(R_3^{(2)}, R_3^{(3)}\right).
$$

Therefore,

$$
\begin{aligned}
& l_n\left(\hat{\boldsymbol{\beta}}^*\right) - l_n\left(\hat{\boldsymbol{\beta}}^s\right) \\
= & \frac{1}{2}\boldsymbol{W}^\top \boldsymbol{\Sigma}\left(\boldsymbol{\beta}^*\right)^{-\frac{1}{2}}\left(\boldsymbol{H}_{\boldsymbol{Z}_{\mathcal{A}^s}} - \boldsymbol{H}_{\boldsymbol{Z}_{\mathcal{A}^*}}\right)\boldsymbol{\Sigma}\left(\boldsymbol{\beta}^*\right)^{-\frac{1}{2}}\boldsymbol{W} + f_1\left(R_3, R_3^{(1)}\right) - f_2\left(R_3^{(2)}, R_3^{(3)}\right).
\end{aligned}
$$

Furthermore, for any $x > 0$, there exists a constant $K_1 > 0$ such that

$$P\left\{\sup_{\mathcal{A}\supset\mathcal{A}^*, |\mathcal{A}|\le s_{\max}} \frac{1}{|\mathcal{A}|-|\mathcal{A}^*|} \boldsymbol{W}^\top \boldsymbol{\Sigma}\left(\boldsymbol{\beta}^*\right)^{-\frac{1}{2}}\left(\boldsymbol{H}_{\boldsymbol{Z}_{\mathcal{A}}}-\boldsymbol{H}_{\boldsymbol{Z}_{\mathcal{A}^*}}\right)\boldsymbol{\Sigma}\left(\boldsymbol{\beta}^*\right)^{-\frac{1}{2}}\boldsymbol{W} > x\right\}$$
$$< K_1 \sum_{s=s^*}^{s_{\max}} \left(\frac{pe}{s}\right)^s \exp\left(-K_1 s x\right).$$

Hence, for any $\alpha > 0$, there exists a constant $K_2 > 0$ such that with probability at least $1 - p^{-\alpha}$,

$$\sup_{\mathcal{A}\supset\mathcal{A}^*, |\mathcal{A}|\le s_{\max}} \frac{1}{|\mathcal{A}|-|\mathcal{A}^*|} \boldsymbol{W}^\top \boldsymbol{\Sigma}\left(\boldsymbol{\beta}^*\right)^{-\frac{1}{2}}\left(\boldsymbol{H}_{\boldsymbol{Z}_{\mathcal{A}}}-\boldsymbol{H}_{\boldsymbol{Z}_{\mathcal{A}^*}}\right)\boldsymbol{\Sigma}\left(\boldsymbol{\beta}^*\right)^{-\frac{1}{2}}\boldsymbol{W} \le K_2(1+\alpha)\log\left(K_2 p\right).$$

Now, considering the term R_3, based on its definition, we have With probability at least $1 - p^{-\alpha}$,

$$\begin{aligned}
R_3 &= \frac{1}{6}\sum_{i=1}^{N}\left(\boldsymbol{x}_i^\top\left(\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}^*\right) b'''\left(\boldsymbol{x}_i\overline{\boldsymbol{\beta}}\right)\left(\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}^*\right)^\top \boldsymbol{x}_i\boldsymbol{x}_i^\top\left(\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}^*\right)\right)\\
&\le \frac{1}{6}c_3\sqrt{s}\left\|\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}^*\right\|_2\left(\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}^*\right)^\top \boldsymbol{X}^\top\boldsymbol{X}\left(\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}^*\right)\\
&\le \frac{1}{6}c_3 N\sqrt{s}M_s\left(\left\|\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}^*\right\|_2\right)^3\\
&\le \frac{1}{6}c_3 N\sqrt{s}M_s\left(\frac{1}{N m_s c_0}\left\|\boldsymbol{X}_{\mathcal{A}^*}^\top\boldsymbol{W}\right\|_2\right)^3\\
&\le \frac{1}{6}c_3 N\sqrt{s}M_s\left(\frac{(1+\alpha)sM_s}{CNm_s c_0^2}\log(2p)\right)^{\frac{3}{2}}.
\end{aligned}$$

Under Assumption 8, a similar argument yields that, with probability at least $1 - p^{-\alpha}$,

$$\frac{1}{N}\left|f_1\left(R_3, R_3^{(1)}\right) - f_2\left(R_3^{(2)}, R_3^{(3)}\right)\right| \le O\left(\sqrt{s}\left[\frac{(1+\alpha)s}{N}\log(2p)\right]^{\frac{3}{2}}\right).$$

Consequently,

$$l_n\left(\hat{\boldsymbol{\beta}}^*\right) - l_n\left(\hat{\boldsymbol{\beta}}^s\right) \le K_2(1+\alpha)\left|\mathcal{B}^s\right|\log\left(K_2 p\right) + O\left(\sqrt{s}\left[\frac{(1+\alpha)s}{N}\log(2p)\right]^{\frac{3}{2}}\right).$$

By Assumption 7, the relation $\sqrt{s}N\left(\frac{s}{N}\log p\right)^{\frac{3}{2}} = O(\log p)$ is satisfied. Hence,

$$\begin{aligned} F\left(\hat{\boldsymbol{\beta}}^{*}\right) - F\left(\hat{\boldsymbol{\beta}}^{s}\right) &= \left[l_n\left(\hat{\boldsymbol{\beta}}^{*}\right) - l_n\left(\hat{\boldsymbol{\beta}}^{s}\right)\right] - |\mathcal{B}^s|\log p\log\log N \\ &\leq O\left((1+\alpha)^{\frac{3}{2}}|\mathcal{B}^s|\log p\right) - |\mathcal{B}^s|\log p\log\log N \\ &< 0. \end{aligned}$$

Thus, for sufficiently large N, with probability at least $1-p^{-\alpha}$, $F(\hat{\boldsymbol{\beta}}^{s})$ reaches its minimum value at $\hat{\boldsymbol{\beta}}^{s^*}$.

S3.3 Proof of Theorem 3

To establish a lower bound on the error, we adopt the approach introduced by Raskutti et al. (2011). Define $\mathcal{H}(s) = \{\boldsymbol{z} \in \{-1,0,+1\}^p \mid \|\boldsymbol{z}\|_0 = s\}$. For notational simplicity, we omit the dependence of $\mathcal{H}$ on s. The Hamming distance between any two vectors $\boldsymbol{z}$ and $\boldsymbol{z}'$ is given by $d_H(\boldsymbol{z},\boldsymbol{z}') = \sum_{j=1}^{p} \mathrm{I}\left(z_j \neq z_j'\right)$. When $s < 2p/3$, one can construct a subset $\tilde{\mathcal{H}} \subset \mathcal{H}$ with cardinality

$$|\tilde{\mathcal{H}}| \geq \exp\left(\frac{s}{2}\log\frac{p-s}{s/2}\right),$$

such that for every distinct pair $\boldsymbol{z},\boldsymbol{z}' \in \tilde{\mathcal{H}}$, $d_H(\boldsymbol{z},\boldsymbol{z}') \geq \frac{s}{2}$. Denote the scaled version of $\tilde{\mathcal{H}}$ as $\sqrt{2/s}\Delta_N\tilde{\mathcal{H}}$, where $\Delta_N > 0$. For any elements $\boldsymbol{\beta},\boldsymbol{\beta}' \in \sqrt{\frac{2}{s^*}}\Delta_N\tilde{\mathcal{H}}$, we have

$$\Delta_N^2 \leq \|\boldsymbol{\beta}-\boldsymbol{\beta}'\|_2^2 \leq 8\Delta_N^2.$$

Thus, $\sqrt{\frac{2}{s^*}}\Delta_N\tilde{\mathcal{H}}$ forms a Δ_N-packing set of $\mathbb{B}_0(s^*)$ and contains $|\tilde{\mathcal{H}}|$ elements. Consequently, the minimax lower bound satisfies

$$\mathbb{P}\left(\min_{\hat{\boldsymbol{\beta}}}\max_{\boldsymbol{\beta}^*\in\mathbb{B}_0}\left\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\right\|_2^2 \geq \Delta_N^2/4\right) \geq \min_{\tilde{\boldsymbol{\beta}}}\mathbb{P}(\tilde{\boldsymbol{\beta}} \neq \boldsymbol{B}),$$

where the random vector $\boldsymbol{B}$ is uniformly distributed over the packing set $\{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_M\}$, and the estimator $\tilde{\boldsymbol{\beta}}$ takes value in the packing set. Applying Fano's inequality, we have

$$\mathbb{P}(\tilde{\boldsymbol{\beta}} \neq \boldsymbol{B}) \geq 1 - \frac{I(\boldsymbol{y}; \boldsymbol{B}) + \log 2}{\log |\tilde{\mathcal{H}}|},$$

where $I(\boldsymbol{y}; \boldsymbol{B})$ denotes the mutual information. This mutual information can be bounded via the KL divergence:

$$I(\boldsymbol{y}; \boldsymbol{B}) \leq \frac{2}{M(M-1)} \sum_{i \neq j} D_{KL}\left(\boldsymbol{\beta}_i \| \boldsymbol{\beta}_j\right),$$

where $D_{KL}\left(\boldsymbol{\beta}_i \| \boldsymbol{\beta}_j\right)$ is the KL divergence between the distribution induced by $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_j$. For a generalized linear model, the KL divergence admits the expansion

$$\begin{aligned}
D_{KL}\left(\boldsymbol{\beta}_i \| \boldsymbol{\beta}_j\right) &= E_{\boldsymbol{\beta}_i}\left[l_n\left(\boldsymbol{\beta}_j\right) - l_n\left(\boldsymbol{\beta}_i\right)\right] \\
&= c_2 E_{\boldsymbol{\beta}_i}\left[\sum_{k=1}^{n} \boldsymbol{y}\boldsymbol{x}_k^\top\left(\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\right) - b\left(\boldsymbol{x}_k^\top \boldsymbol{\beta}_i\right) + b\left(\boldsymbol{x}_k^\top \boldsymbol{\beta}_j\right)\right] \\
&= c_2 E_{\boldsymbol{\beta}_i}\left[\sum_{k=1}^{n} \boldsymbol{y}\boldsymbol{x}_k^\top\left(\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\right) + b'\left(\boldsymbol{x}_k^\top \boldsymbol{\beta}_i\right) x_k^\top\left(\boldsymbol{\beta}_j - \boldsymbol{\beta}_i\right)\right. \\
&\quad \left. + \frac{1}{2} b''\left(\boldsymbol{x}_k^\top \tilde{\boldsymbol{\beta}}_i\right)\left(\boldsymbol{\beta}_j - \boldsymbol{\beta}_i\right)^\top \boldsymbol{x}_k \boldsymbol{x}_k^\top\left(\boldsymbol{\beta}_j - \boldsymbol{\beta}_i\right)\right] \\
&= c_2 E_{\boldsymbol{\beta}_i}\left[\sum_{k=1}^{n} \frac{1}{2} b''\left(\boldsymbol{x}_k^\top \tilde{\boldsymbol{\beta}}_i\right)\left(\boldsymbol{\beta}_j - \boldsymbol{\beta}_i\right)^\top \boldsymbol{x}_k \boldsymbol{x}_k^\top\left(\boldsymbol{\beta}_j - \boldsymbol{\beta}_i\right)\right] \\
&\leq \frac{N c_2}{2 c_0} M_{2s^*}\left\|\boldsymbol{\beta}_j - \boldsymbol{\beta}_i\right\|_2^2,
\end{aligned}$$

where $\tilde{\boldsymbol{\beta}}_i$ lying on the line segment between $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_j$. Thus,

$$\mathbb{P}(\tilde{\boldsymbol{\beta}} \neq \boldsymbol{B}) \geq 1 - \frac{\frac{N c_2}{2 c_0} M_{2s^*} 8 \Delta_N^2 + \log 2}{\frac{s^*}{2} \log \frac{p - s^*}{s^*/2}}.$$

By setting $\Delta_N^2 = \frac{1}{16}\frac{c_0}{NM_{2s^*}c_2}\frac{s^*}{2}\log\frac{p-s^*}{s^*/2}$, we obtain

$$\mathbb{P}\left(\min_{\hat{\boldsymbol{\beta}}}\max_{\boldsymbol{\beta}^*\in\mathbb{B}_0}\left\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\right\|_2^2 \geq \frac{1}{32}\frac{c_0}{M_{2s^*}c_2}\frac{s^*}{N}\log\frac{p}{s^*}\right) \geq 1/2.$$

In the regime where $s^* < p/2$ and $s^* \leq N \ll p$, it follows that with probability at least 1/2,

$$\min_{\hat{\boldsymbol{\beta}}}\max_{\boldsymbol{\beta}^*\in\mathbb{B}_0}\left\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\right\|_2^2 \geq \frac{1}{32}\frac{c_0}{M_{2s^*}c_2}\frac{s^*}{N}\log\frac{p}{s^*}.$$

Next, we establish the corresponding upper bound for the estimation error. By Theorem 2, with probability $1-O\left(p^{-\alpha}\right)$, we have $\hat{\mathcal{A}} = \mathcal{A}^*$, which implies

$$\left(\hat{\boldsymbol{\beta}}_{\hat{s}}\right)_{\mathcal{A}^*} - \boldsymbol{\beta}^*_{\mathcal{A}^*} = \left\{\left.\frac{\partial^2 l_n(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^2_{\mathcal{A}^*}}\right|_{\overline{\boldsymbol{\beta}}}\right\}^{-1}\left.\frac{\partial l_n(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}_{\mathcal{A}^*}}\right|_{\boldsymbol{\beta}^*} = \left\{\left.\frac{\partial^2 l_n(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^2_{\mathcal{A}^*}}\right|_{\overline{\boldsymbol{\beta}}}\right\}^{-1}\mathbf{X}^\top_{\mathcal{A}^*}\boldsymbol{W}.$$

Applying a tail bound for quadratic forms of sub-Gaussian variables Hsu et al. (2012) yields

$$\mathbb{P}\left(\|\mathbf{X}'_{\mathcal{A}^*}\boldsymbol{W}\|_2^2 > \frac{1}{c_2}\left(\operatorname{tr}\left(\mathbf{X}_{\mathcal{A}^*}\mathbf{X}'_{\mathcal{A}^*}\right) + 2\left\|\mathbf{X}_{\mathcal{A}^*}\mathbf{X}'_{\mathcal{A}^*}\right\|_F\sqrt{t} + 2\lambda_{\max}\left(\mathbf{X}_{\mathcal{A}^*}\mathbf{X}'_{\mathcal{A}^*}\right)t\right)\right) \leq c_1 e^{-t}.$$

Using the SRC constraint condition, we have the following spectral bounds:

$$\operatorname{tr}\left(\mathbf{X}_{\mathcal{A}^*}\mathbf{X}'_{\mathcal{A}^*}\right) \leq Ns^*M_{s^*}, \quad \left\|\mathbf{X}_{\mathcal{A}^*}\mathbf{X}'_{\mathcal{A}^*}\right\|_F \leq N^2 s^* M_{s^*}^2, \quad \lambda_{\max}\left(\mathbf{X}_{\mathcal{A}^*}\mathbf{X}'_{\mathcal{A}^*}\right) \leq NM_{s^*}.$$

Consequently, $\mathbb{P}\left(\|\mathbf{X}'_{\mathcal{A}^*}\boldsymbol{W}\|_2^2 > N\frac{1}{c_2}M_{s^*}\left(s^* + 2\sqrt{s^*t} + 2t\right)\right) \leq c_1 e^{-t}$.

In the high-dimensional regime $p \gg N \geq s^*$, take $t = 2s^*\log\left(p/s^*\right)$. Provided $p \geq es^*$, we have $s^* \leq t$, which together with the previous tail bound gives

$$\mathbb{P}\left(\|\mathbf{X}'_{\mathcal{A}^*}\boldsymbol{W}\|_2^2 > 10\mathrm{N}\frac{1}{c_2}M_{s^*}s^*\log\left(p/s^*\right)\right) \leq c_1 e^{-2s^*\log(p/s^*)}.$$

Since $\frac{p}{s^*} \leq \left(\frac{ep}{s^*}\right)^{s^*} \leq e^{s^*\log ep/s^*}$, it follows that

$$\begin{aligned}\mathbb{P}(\max_{\mathcal{A}^*}\|\mathbf{X}'_{\mathcal{A}^*}\boldsymbol{W}\|_2^2 > 10\frac{1}{c_2}NM_{s^*}s^*\log\left(p/s^*\right)) &\leq c_1 e^{s^*\log(ep/s^*)}e^{-2s^*\log(p/s^*)}\\ &\leq c_1 e^{-s^*\log(p/(es^*))}.\end{aligned}$$

Equivalently,

$$\begin{aligned}
&\mathbb{P}\left(\left\|\hat{\boldsymbol{\beta}}_{\hat{s}}-\boldsymbol{\beta}^*\right\|_2^2 \geq 10\frac{1}{c_2c_0^2}\frac{M_{s^*}^2}{m_{s^*}^2}\frac{s^*\log\left(p/s^*\right)}{N}\right)\\
\leq&\mathbb{P}\left(\left\|\left\{\left.\frac{\partial^2 l_n(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}_{\mathcal{A}^*}^2}\right|_{\overline{\boldsymbol{\beta}}}\right\}^{-1}\mathbf{X}_{\mathcal{A}^*}^\top\boldsymbol{W}\right\|_2^2 \geq 10\frac{1}{c_2c_0}\frac{M_{s^*}}{m_{s^*}^2}\frac{s^*\log\left(p/s^*\right)}{N}\right)\\
\leq&\mathbb{P}\left(\max_{\mathcal{A}^*}\left\|\frac{1}{Nm_{s^*}c_0}\mathbf{X}'_{\mathcal{A}^*}\boldsymbol{W}\right\|_2^2 \geq 10\frac{1}{c_2c_0^2}\frac{M_{s^*}}{m_{s^*}^2}\frac{s^*\log\left(p/s^*\right)}{N}\right)\\
\leq&\mathbb{P}\left(\max_{\mathcal{A}^*}\|\mathbf{X}'_{\mathcal{A}^*}\boldsymbol{W}\|_2^2 \geq 10\frac{1}{c_2}NM_{s^*}s^*\log\left(p/s^*\right)\right)\\
\leq&c_1e^{-s^*\log(p/es^*)}.
\end{aligned}$$

Thus, the proof is complete.

S3.4 Proof of Theorem 4

Recall that $\beta^*_{A^*}$ denotes the true coefficient vector restricted to the true active set A^* with size s^*, and I_* denotes the corresponding Fisher information matrix. For each observation, define $\mu_i^* := b'(X_{i,A^*}^\top\beta^*_{A^*})$, and $\psi_i := X_{i,A^*}\{y_i-\mu_i^*\}$.

Lemma S1 (Second-order expansion of local restricted MLEs)**.** *For each machine $k=1,\ldots,m$, let $\mathcal{M}_k$ be the index set of its local samples with $|\mathcal{M}_k|=n$, and consider the local negative log-likelihood restricted to the true active set A^*,*

$$\mathcal{L}_k(\theta) := -\frac{1}{n}\sum_{i\in\mathcal{M}_k}\bigl\{y_iX_{i,A^*}^\top\theta - b(X_{i,A^*}^\top\theta)+c(y_i,\phi)\bigr\},\qquad \theta\in\mathbb{R}^{s^*}.$$

Let $\widehat{\theta}_k$ be the local maximum likelihood estimator on machine k,

$$\widehat{\theta}_k := \arg\min_{\theta\in\mathbb{R}^{s^*}}\mathcal{L}_k(\theta),\quad \textit{equivalently}\ \ \widehat{\beta}^{\mathrm{or}}_{k,(A^*)^c}=0,\ \widehat{\beta}^{\mathrm{or}}_{k,A^*}=\widehat{\theta}_k.$$

Then there exist random remainder vectors $r_{k,n} \in \mathbb{R}^{s^*}$ *such that*

$$\widehat{\theta}_k - \beta^*_{A^*} = I_*^{-1} \frac{1}{n} \sum_{i \in \mathcal{M}_k} \psi_i + r_{k,n}, \qquad \|r_{k,n}\|_2 = O_p(n^{-1}). \tag{S3.1}$$

Consequently,

$$\sqrt{n}\,(\widehat{\theta}_k - \beta^*_{A^*}) = I_*^{-1} \frac{1}{\sqrt{n}} \sum_{i \in \mathcal{M}_k} \psi_i + o_p(1), \qquad k = 1, \ldots, m. \tag{S3.2}$$

Let $U_k(\theta) = \nabla \mathcal{L}_k(\theta)$ and $H_k(\theta) = \nabla^2 \mathcal{L}_k(\theta)$ denote the score function and the Hessian matrix, respectively. A direct calculation yields

$$U_k(\theta) = -\frac{1}{n} \sum_{i \in \mathcal{M}_k} \{y_i - b'(X_{i,A^*}^\top \theta)\} X_{i,A^*}, \quad H_k(\theta) = \frac{1}{n} \sum_{i \in \mathcal{M}_k} b''(X_{i,A^*}^\top \theta)\, X_{i,A^*} X_{i,A^*}^\top.$$

The local MLE $\widehat{\theta}_k$ satisfies the first-order condition $U_k(\widehat{\theta}_k) = 0$. Since $\mathcal{L}_k(\theta)$ is three times continuously differentiable in a neighborhood of $\beta^*_{A^*}$, and b''' is uniformly bounded under Assumption (A1), a Taylor expansion of U_k around $\beta^*_{A^*}$ gives

$$0 = U_k(\widehat{\theta}_k) = U_k(\beta^*_{A^*}) + H_k(\beta^*_{A^*})(\widehat{\theta}_k - \beta^*_{A^*}) + R_k, \tag{S3.3}$$

where R_k collects the third order remainder terms, and the expansion holds for some intermediate point $\tilde{\theta}_k$ on the line segment between $\widehat{\theta}_k$ and $\beta^*_{A^*}$. Equivalently, by the mean-value form of the Taylor expansion,

$$0 = U_k(\widehat{\theta}_k) = U_k(\beta^*_{A^*}) + H_k(\tilde{\theta}_k)(\widehat{\theta}_k - \beta^*_{A^*}),$$

which implies

$$\widehat{\theta}_k - \beta^*_{A^*} = -H_k(\tilde{\theta}_k)^{-1} U_k(\beta^*_{A^*}).$$

Under (A1), (A3) and (A9), the eigenvalues of $H_k(\tilde{\theta}_k)$ are uniformly bounded away from zero and infinity with probability tending to one. Consequently, $\|H_k(\tilde{\theta}_k)^{-1}\|_{\mathrm{op}} = O_p(1)$. Moreover, $U_k(\beta^*_{A^*}) = -(1/n) \sum_{i \in \mathcal{M}_k} \psi_i$ has mean zero and finite second moments under (A1), (A2) and (A9). When s^* is fixed, this implies $\|U_k(\beta^*_{A^*})\|_2 = O_p(n^{-1/2})$. Combining the above bounds yields the standard root-n rate, $\|\widehat{\theta}_k - \beta^*_{A^*}\|_2 = O_p(n^{-1/2})$, which follows from classical M-estimation theory; see, for example, Chapter 5 of van der Vaart (1998) for further details.

By the boundedness of $b'''(\cdot)$ under Assumption (A1) and the bounded design condition in Assumption (A9), the third derivative of $\mathcal{L}_k$ with respect to θ is uniformly bounded in a neighborhood of $\beta^*_{A^*}$. Consequently, there exists a constant $C > 0$ such that

$$\|R_k\|_2 \leq C\|\widehat{\theta}_k - \beta^*_{A^*}\|_2^2. \tag{S3.4}$$

Combining (S3.4) with the root-n rate established in Step 1 yields $\|R_k\|_2 = O_p(n^{-1})$. Rearranging the Taylor expansion (S3.3) gives

$$\widehat{\theta}_k - \beta^*_{A^*} = -H_k(\beta^*_{A^*})^{-1}U_k(\beta^*_{A^*}) - H_k(\beta^*_{A^*})^{-1}R_k.$$

Since $U_k(\beta^*_{A^*}) = -(1/n)\sum_{i\in\mathcal{M}_k}\psi_i$, the leading term can be written as

$$-H_k(\beta^*_{A^*})^{-1}U_k(\beta^*_{A^*}) = I_*^{-1}\frac{1}{n}\sum_{i\in\mathcal{M}_k}\psi_i + \left\{H_k(\beta^*_{A^*})^{-1} - I_*^{-1}\right\}\frac{1}{n}\sum_{i\in\mathcal{M}_k}\psi_i.$$

By the law of large numbers, together with the boundedness of X_{i,A^*} and $b''(\cdot)$, we have $\|H_k(\beta^*_{A^*}) - I_*\|_{\mathrm{op}} = O_p(n^{-1/2})$. Since $\lambda_{\min}(I_*) > 0$, a standard matrix perturbation argument implies $\|H_k(\beta^*_{A^*})^{-1} - I_*^{-1}\|_{\mathrm{op}} = O_p(n^{-1/2})$. Using again that $\|(1/n)\sum_{i\in\mathcal{M}_k}\psi_i\|_2 = O_p(n^{-1/2})$, we obtain

$$\left\|\left\{H_k(\beta^*_{A^*})^{-1} - I_*^{-1}\right\}\frac{1}{n}\sum_{i\in\mathcal{M}_k}\psi_i\right\|_2 = O_p(n^{-1}).$$

Together with $\|H_k(\beta^*_{A^*})^{-1}R_k\|_2 = O_p(n^{-1})$ from (S3.4), this establishes the expansion in (S3.1). Multiplying both sides of that expansion by $\sqrt{n}$ yields (S3.2).

Let $(\widehat{\beta}, \widehat{A})$ be the output of Algorithm 1 or Algorithm 3, and define $B_N = \{\widehat{A} = A^*\}$ as the event of exact support recovery. By Theorem 2 in the main text, $\mathbb{P}(B_N) \to 1$. On the event B_N, the local estimator $\widehat{\beta}_k$ obtained in Stage 2 coincides with the oracle local MLE in Lemma S1; that is, $\widehat{\beta}_{k,A^*} = \widehat{\theta}_k$ and $\widehat{\beta}_{k,(A^*)^c} = 0$. Averaging over machines yields

$$\widehat{\beta}_{A^*} = \frac{1}{m} \sum_{k=1}^{m} \widehat{\theta}_k = \beta^*_{A^*} + I_*^{-1} \frac{1}{N} \sum_{i=1}^{N} \psi_i + \frac{1}{m} \sum_{k=1}^{m} r_{k,n},$$

where $N = mn$. By Lemma S1, $\|r_{k,n}\|_2 = O_p(n^{-1})$, and therefore

$$\left\| \frac{1}{m} \sum_{k=1}^{m} r_{k,n} \right\|_2 \le \frac{1}{m} \sum_{k=1}^{m} \|r_{k,n}\|_2 = O_p(n^{-1}).$$

Consequently,

$$\sqrt{N} \left\| \frac{1}{m} \sum_{k=1}^{m} r_{k,n} \right\|_2 = O_p\left(\frac{\sqrt{N}}{n} \right) = O_p\left(\frac{m}{\sqrt{N}} \right) \longrightarrow 0$$

under Assumption (A10). This establishes the linear representation

$$\sqrt{N}\,(\widehat{\beta}_{A^*} - \beta^*_{A^*}) = I_*^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \psi_i + o_p(1)$$

on the event B_N.

Next, we verify the central limit theorem for $N^{-1/2} \sum_{i=1}^{N} \psi_i$. Under the GLM specification, $\mathbb{E}(y_i \mid X_i) = \mu_i^*$ and $\mathrm{Var}(y_i \mid X_i) = b''(X_{i,A^*}^\top \beta^*_{A^*})$, which implies $\mathbb{E}(\psi_i) = 0$ and

$$\mathrm{Var}(\psi_i) = \mathbb{E}\left[b''(X_{A^*}^\top \beta^*_{A^*})\, X_{A^*} X_{A^*}^\top \right] = I_*.$$

Assumptions (A1), (A2) and (A9) ensure that ψ_i has sub-Gaussian tails and finite $(2+\delta)$-th moments for some $\delta > 0$. Hence, the Lindeberg condition holds, and by the classical central limit theorem,

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \psi_i \xrightarrow{d} \mathcal{N}(0, I_*).$$

Applying Slutsky's theorem yields, conditionally on B_N,

$$\sqrt{N}\,(\widehat{\beta}_{A^*} - \beta^*_{A^*}) \ \xrightarrow{d}\ \mathcal{N}(0, I_*^{-1}).$$

Finally, we remove the conditioning. For any bounded continuous function f, we have

$$\big|\mathbb{E}f(Z_N) - \mathbb{E}f(Z)\big| \le \big|\mathbb{E}[f(Z_N) \mid B_N] - \mathbb{E}f(Z)\big| + 2\|f\|_\infty\, \mathbb{P}(B_N^c),$$

where $Z_N = \sqrt{N}(\widehat{\beta}_{A^*} - \beta^*_{A^*})$ and $Z \sim \mathcal{N}(0, I_*^{-1})$. The first term converges to zero by the conditional convergence established above, and the second term converges to zero because $\mathbb{P}(B_N^c) \to 0$. Therefore,

$$\sqrt{N}\,(\widehat{\beta}_{A^*} - \beta^*_{A^*}) \ \xrightarrow{d}\ \mathcal{N}(0, I_*^{-1}).$$

Moreover, on B_N we have $\widehat{\beta}_{(A^*)^c} \equiv 0$, so $\widehat{\beta}_{(A^*)^c} = 0$ with probability tending to one. This completes the proof of Theorem 4.

S4 Simulation Results for the Poisson Regression

This section presents the simulation results for the Poisson regression model, complementing the logistic regression experiments presented in Section 4 of the main text. The experimental setup follows the same distributed framework: the number of covariates is $p = 100$, the total sample size is $N = 50000$, and the data were evenly distributed across $m \in \{20, 40, 50\}$ machines. The sparsity level is set to $s^* = 5$, with non-zero coefficients given by $\beta^*_{\mathcal{A}^*} = (1, -1, -1, -1, 1)^\top$. Figures S1 and S2 display the trajectories of the estimation error under zero-vector initialization and one-shot averaging initialization, respectively. All results are averaged over 100 independent replications. Across all settings, DBESS-GLM consistently achieves smaller estimation errors and faster convergence compared with CSL-GLM and

CEASE-GLM. The one-Shot initialization further enhances numerical stability for all algorithms, particularly as the number of machines increases. Sparse subset recovery performance is measured by TPR,TNR, MCC and ReEE. As reported in Table S1, DBESS-GLM achieves higher TNR and MCC together with lower ReEE, indicating more accurate exclusion of irrelevant variables and more precise coefficient estimation. Overall, the Poisson regression exhibits patterns similar to those observed for logistic regression in the main text. These findings further confirm the stability, consistency, and effectiveness of DBESS-GLM across different members of the GLM family, highlighting its broad applicability in distributed settings. Table S2 presents descriptive statistics for the real dataset analyzed in Section 5 of the main text.

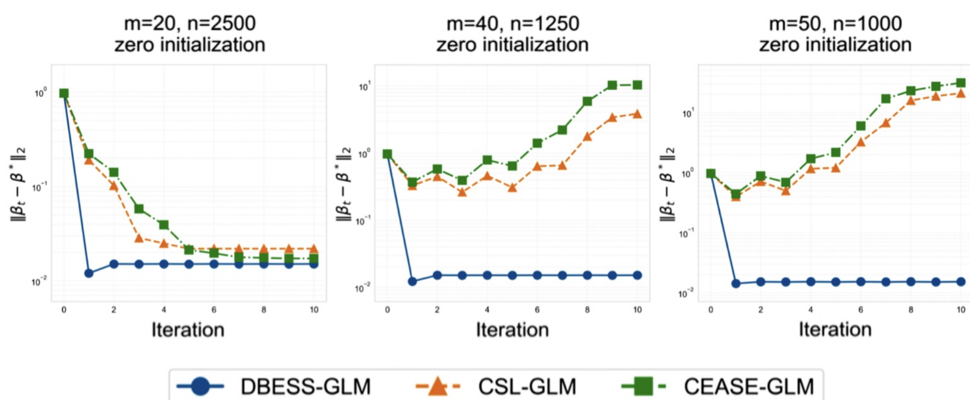

Figure S1: Comparison of convergence rates for zero-vector initialization when m takes values of {20, 40, 50}.

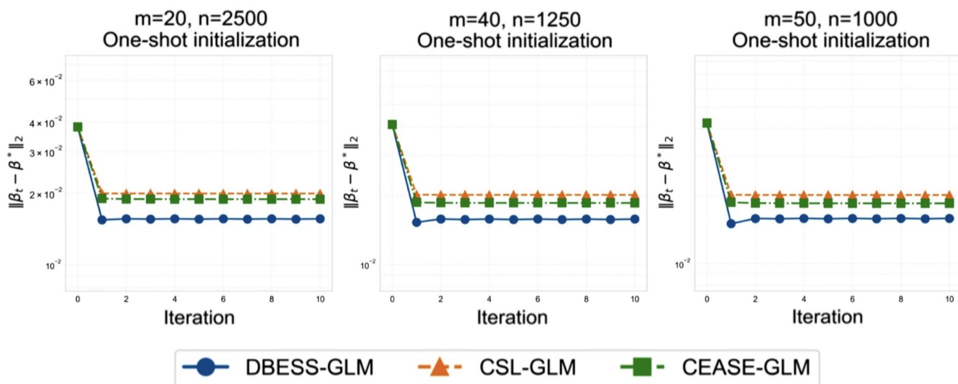

Figure S2: Comparison of convergence rates for one-shot initialization when m takes values of $\{20, 40, 50\}$.

Table S1: Comparison of TPR, TNR, MCC and ReEE in Poisson Regression

Method	TPR	TNR	MCC	ReEE
		m=20		
CSL-GLM	1.000(0.000)	0.963(0.025)	0.882(0.081)	0.049(0.053)
CEASE-GLM	1.000(0.000)	0.987(0.010)	0.942(0.043)	0.045(0.046)
DBESS-GLM	1.000(0.000)	1.000(0.000)	0.949(0.027)	0.016(0.023)
		m=40		
CSL-GLM	0.999(0.010)	0.958(0.072)	0.829(0.092)	0.051(0.064)
CEASE-GLM	1.000(0.000)	0.968(0.034)	0.936(0.044)	0.047(0.065)
DBESS-GLM	1.000(0.000)	0.989(0.002)	0.948(0.042)	0.016(0.032)
		m=50		
CSL-GLM	0.995(0.026)	0.949(0.142)	0.839(0.094)	0.062(0.074)
CEASE-GLM	0.996(0.019)	0.935(0.136)	0.931(0.048)	0.052(0.057)
DBESS-GLM	1.000(0.000)	0.988(0.004)	0.946(0.043)	0.016(0.039)

Table S2: Descriptive Statistical Analysis of Selected Features.

Feature	Mean	SD	Median	Meaning
age	61.99	16.82	64.00	Patient age
bmi	29.11	8.26	27.56	Patient BMI index
d1_heartrate_min	70.41	17.14	70.00	Minimum heart rate within the first 24 hours
d1_heartrate_max	103.30	21.98	101.00	Maximum heart rate within the first 24 hours
h1_heartrate_min	83.79	20.22	82.00	Minimum heart rate within the first 1 hour
h1_heartrate_max	92.45	21.78	90.00	Maximum heart rate within the first 1 hour
d1_glucose_min	114.44	38.22	108.00	Minimum glucose concentration in blood or plasma within the first 24 hours
d1_glucose_max	174.15	86.55	150.00	Maximum glucose concentration in blood or plasma within the first 24 hours
d1_creatinine_min	1.36	1.34	0.94	Minimum creatinine concentration in blood or plasma within the first 24 hours
d1_creatinine_max	1.49	1.53	1.00	Maximum creatinine concentration in blood or plasma within the first 24 hours
h1_glucose_min	158.91	88.30	134.00	Minimum glucose concentration in blood or plasma within the first 1 hour
h1_glucose_max	167.70	94.03	140.00	Maximum glucose concentration in blood or plasma within the first 1 hour

References

Fan, Y. & Tang, C. Y. (2012), 'Tuning parameter selection in high dimensional penalized likelihood', *Journal of the Royal Statistical Society Series B: Statistical Methodology* **75**(3), 531–552.

Hsu, D., Kakade, S. & Zhang, T. (2012), 'A tail inequality for quadratic forms of subgaussian random vectors', *Electronic Communications in Probability* **17**, 1–6.

Huang, J., Jiao, Y., Liu, Y. & Lu, X. (2018), 'A constructive approach to l_0 penalized regression', *Journal of Machine Learning Research* **19**(10), 1–37.

Raskutti, G., Wainwright, M. J. & Yu, B. (2011), 'Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls', *IEEE Transactions on Information Theory* **57**(10), 6976–6994.

Rigollet, P. (2012), 'Kullback-Leibler aggregation and misspecified generalized linear models', *The Annals of Statistics* **40**(2), 639–665.

Shi, C., Song, R., Chen, Z. & Li, R. (2019), 'Linear hypothesis testing for high dimensional generalized linear models', *The Annals of Statistics* **47**(5), 2671–2703.

van der Vaart, A. W. (1998), *Asymptotic Statistics*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press.

Zhu, J., Wen, C., Zhu, J., Zhang, H. & Wang, X. (2020), 'A polynomial algorithm for best-subset selection problem', *Proceedings of the National Academy of Sciences* **117**(52), 33117–33123.