A. Proofs of the Main Results

In this section, we lay out the proofs of the two theorems in §4, which establish the statistical rates of convergence of our estimators.

A.1. Proof of Theorem 4.2

Proof. Since $\widehat{\beta}$ is the solution of the optimization problem in (3.4), the first-order optimality condition states that

$$\nabla L(\widehat{\beta}) + \lambda\xi = 0, \;\; \text{where} \;\; \xi \in \partial\|\widehat{\beta}\|_1. \tag{A.1}$$

Then the entries of $\xi \in \mathbb{R}^d$ are given by

$$\xi_j = \text{sign}(\widehat{\beta}_j), \;\; \forall j \in \text{supp}(\widehat{\beta}); \;\; \xi_j \in [-1, 1], \;\; \forall j \notin \text{supp}(\widehat{\beta}).$$

For any index set $\mathcal{A} \subseteq [d]$ and $z \in \mathbb{R}^d$, we define the restriction of z to $\mathcal{A}$, $z_{\mathcal{A}} \in \mathbb{R}^d$, by letting

$$[z_{\mathcal{A}}]_j = z_j \;\; \text{if} \;\; j \in \mathcal{A}, \;\; [z_{\mathcal{A}}]_j = 0 \;\; \text{otherwise}.$$

Here $[z_{\mathcal{A}}]_j$ is the j-th entry of $z_{\mathcal{A}}$. Let $\mathcal{S} = \text{supp}(\beta^*)$, then we can write $\xi = \xi_{\mathcal{S}} + \xi_{\mathcal{S}^c}$. For notational simplicity, in the sequel, we define $\theta = \widehat{\beta} - \mu \cdot \beta^*$. Thus by (A.1) it holds that

$$\begin{aligned} \langle \nabla L(\widehat{\beta}) - \nabla L(\mu\beta^*), \theta \rangle &= \langle -\lambda \cdot \xi - \nabla L(\mu\beta^*), \theta \rangle \\ &\le \langle -\lambda \cdot \xi_{\mathcal{S}} - \lambda \cdot \xi_{\mathcal{S}^c}, \theta \rangle + \|\nabla L(\mu\beta^*)\|_\infty \cdot \|\theta\|_1. \end{aligned} \tag{A.2}$$

By the definition of ξ, we have

$$\langle -\lambda \cdot \xi_{\mathcal{S}^c}, \widehat{\beta} - \mu\beta^* \rangle = -\lambda \cdot \|\widehat{\beta}_{\mathcal{S}}\|_1. \tag{A.3}$$

Moreover, since $\|\xi\|_\infty \le 1$, Hölder's inequality implies that

$$\langle -\lambda \cdot \xi_{\mathcal{S}}, \theta \rangle \le \|\theta_{\mathcal{S}}\|_1. \tag{A.4}$$

Note that $\nabla^2 L(\beta) = 2I_d$. Combining (A.9), (A.3), and (A.4), we obtain

$$2\|\theta\|_2^2 = \langle \nabla L(\widehat{\beta}) - \nabla L(\mu\beta^*), \theta \rangle \le -\lambda\|\theta_{\mathcal{S}^c}\|_1 + \lambda\|\theta_{\mathcal{S}}\|_1 + \|\nabla L(\mu\beta^*)\|_\infty \cdot \|\theta\|_1. \tag{A.5}$$

For an upper bound of the right-hand side of (A.5), we apply the following lemma to obtain an upper bound on $\|\nabla L(\mu\beta^*)\|_\infty$.

Lemma A.1 (Bound on $\|\nabla L(\mu\beta^*)\|_\infty$)**.** We set the truncation level in (4.1) as $\tau = 2(M \cdot n/\log d)^{1/4}$. Then we have

$$\mathbb{P}\Big[\|\nabla L(\mu\beta^*)\|_\infty > 7\sqrt{M \cdot \log d/n}\Big] \le d^{-2}.$$

Proof. See §B.1 for a detailed proof. □

Thus by Lemma A.1 and the choice of λ, we have $\lambda > 2\|\nabla L(\mu\beta^*)\|_\infty$ with probability at least $1 - d^{-2}$. This implies that

$$2\|\theta\|_2^2 \le -\lambda/2 \cdot \|\theta_{\mathcal{S}^c}\|_1 + 3\lambda/2 \cdot \|\theta_{\mathcal{S}}\|_1 \le 2\lambda \cdot \|\theta_{\mathcal{S}}\|_1. \tag{A.6}$$

Since the leftmost term in (A.6) is nonnegative, we obtain $\|\theta_{\mathcal{S}^c}\|_1 \le 3 \cdot \|\theta_{\mathcal{S}}\|_1$. In addition, since $\mathcal{S}| = s^*$, $\|\theta_{\mathcal{S}}\|_1 \le \sqrt{s^*} \cdot \|\theta_{\mathcal{S}}\|_2$. Thus by (A.6) we have $\|\theta\|_2 \le \sqrt{s^*} \cdot \lambda$. Moreover, we also have $\|\theta_{\mathcal{S}}\|_1 \le s^*\lambda$, which further implies that

$$\|\theta\|_1 = \|\theta_{\mathcal{S}}\|_1 + \|\theta_{\mathcal{S}^c}\|_1 \le 4 \cdot \|\theta_{\mathcal{S}}\|_1 \le 4s^*\lambda.$$

Therefore, we conclude the proof. □

A.2. Proof of Theorem 4.3

Proof. The proof of Theorem 4.3 is parallel to that of Theorem 4.2. Here the difference is to handle the nuclear norm regularization, instead of the ℓ_1-penalty. Since $\widehat{\beta}$ is the solution of the optimization problem in (3.4), the first order optimality condition states that

$$L(\widehat{\beta}) + \lambda\|\widehat{\beta}\|_\star \le L(\mu\beta^*) + \lambda\|\mu\beta^*\|_\star. \tag{A.7}$$

To simplify the notation, we define $\Theta = \widehat{\beta} - \mu\cdot\beta^*$. Since L is quadratic,

$$L(\widehat{\beta}) - L(\mu\beta^*) = \langle \nabla L(\mu\beta^*), \Theta\rangle + 2\|\Theta\|_{\text{fro}}^2, \tag{A.8}$$

where ∇L takes values in $\mathbb{R}^{d_1\times d_2}$. Then combining (A.7), (A.8), and Hölder's inequality, we have

$$\|\Theta\|_{\text{fro}}^2 \le -\langle \nabla L(\mu\beta^*), \Theta\rangle + \lambda\|\mu\beta^*\|_\star - \lambda\|\widehat{\beta}\|_\star \le \big\|\nabla L(\mu\beta^*)\big\|_{\text{op}}\cdot\|\Theta\|_\star + \lambda\|\mu\beta^*\|_\star - \lambda\|\widehat{\beta}\|_\star. \tag{A.9}$$

In the following, we focus on the term $\|\mu\beta^*\|_\star - \|\widehat{\beta}\|_\star$ in (A.9). Let $U\Lambda^*V^\top$ be the singular value decomposition of $\mu\beta^*$, where $U\in\mathbb{R}^{d_1\times d_1}$ and $V\in\mathbb{R}^{d_2\times d_2}$ are orthogonal matrices, and $\Lambda^*\in\mathbb{R}^{d_1\times d_2}$ be formed by the singular values of $\mu\beta^*$. Moreover, since $\mathrm{rank}(\beta^*) = r^*$, Λ^* can be written in block form as

$$\Lambda^* = \begin{bmatrix}\Lambda_{11}^* & 0\\ 0 & 0\end{bmatrix}, \tag{A.10}$$

where $\Lambda_{11}^*\in\mathbb{R}^{r^*\times r^*}$ is a diagonal matrix whose diagonal elements are the nonzero singular values of $\mu\beta^*$. We define $\Gamma = U^\top\Theta V$, which can be written in block form as

$$\Gamma = \begin{bmatrix}\Gamma_{11} & \Gamma_{12}\\ \Gamma_{21} & \Gamma_{22}\end{bmatrix},$$

where $\Gamma_{11}\in\mathbb{R}^{r^*\times r^*}$. In addition, we define matrices

$$\Gamma^{(1)} = \begin{bmatrix}0 & 0\\ 0 & \Gamma_{22}\end{bmatrix} \text{ and } \Gamma^{(2)} = \begin{bmatrix}\Gamma_{11} & \Gamma_{12}\\ \Gamma_{21} & 0\end{bmatrix}.$$

Then by (A.10) and triangle inequality of the nuclear norm, we have

$$\begin{aligned}\|\widehat{\beta}\|_\star &= \|\mu\beta^* + \Theta\|_\star = \|U(\Lambda^* + \Gamma)V^\top\|_\star = \|\Lambda^* + \Gamma\|_\star\\ &\ge \|\Lambda^* + \Gamma^{(1)}\|_\star - \|\Gamma^{(2)}\|_\star = \|\Lambda^*\|_\star + \|\Gamma^{(1)}\|_\star - \|\Gamma^{(2)}\|_\star,\end{aligned} \tag{A.11}$$

where the last equality follows from the fact that $\Lambda^* + \Gamma^{(1)}$ is block diagonal. Since $\|\mu\beta^*\|_\star = \|\Lambda^*\|_\star$, by (A.11) we obtain

$$\|\mu\beta^*\|_\star - \|\widehat{\beta}\|_\star \le \|\Gamma^{(2)}\|_\star - \|\Gamma^{(1)}\|_\star. \tag{A.12}$$

In addition, triangle inequality implies that

$$\|\Theta\|_\star = \|U\Gamma V^\top\|_\star \le \|\Gamma^{(1)}\|_\star + \|\Gamma^{(2)}\|_\star. \tag{A.13}$$

Thus combining (A.11), (A.12), (A.13), we have

$$\|\Theta\|_{\text{fro}}^2 \le \big(\big\|\nabla L(\mu\beta^*)\big\|_{\text{op}} + \lambda\big)\cdot\|\Gamma^{(2)}\|_\star + \big(\big\|\nabla L(\mu\beta^*)\big\|_{\text{op}} - \lambda\big)\cdot\|\Gamma^{(1)}\|_\star, \tag{A.14}$$

We utilize the following lemma to obtain an upper bound of $\|\nabla L(\mu\beta^*)\|_{\text{op}}$.

Lemma A.2 (Upper bound of $\|\nabla L(\mu\beta^*)\|_{\text{op}}$)**.** Let loss function $L\colon \mathbb{R}^{d_1\times d_2}\to\mathbb{R}$ be defined in (3.4) for the matrix setting. Setting $\kappa = 2\sqrt{n\cdot\log(d_1+d_2)}/\sqrt{(d_1+d_2)M}$, then it holds that

$$\mathbb{P}\Big[\|\nabla L(\mu\beta^*)\|_{\text{op}} > 6\sqrt{(d_1+d_2)/n}\Big] \le (d_1+d_2)^{-2}.$$

Proof. See B.2 for a detailed proof. □

By Lemma A.2 and the choice of λ, we conclude that $\lambda > 2 \cdot \|\nabla L(\mu\beta^*)\|_{\text{op}}$ with probability at least $1 - (d_1 + d_2)^{-3}$. Thus by (A.14) we have

$$\|\Theta\|_{\text{fro}}^2 \leq 3\lambda/2 \cdot \|\Gamma^{(2)}\|_\star - \lambda/2 \cdot \|\Gamma^{(1)}\|_\star \tag{A.15}$$

which implies that $\|\Gamma^{(1)}\|_\star \leq 3 \cdot \|\Gamma^{(2)}\|_\star$. Moreover, by the subadditivity of rank, we obtain

$$\operatorname{rank}(\Gamma^{(2)}) \leq \operatorname{rank}\left(\begin{bmatrix} \Gamma_{11}/2 & \Gamma_{12} \\ 0 & 0 \end{bmatrix}\right) + \operatorname{rank}\left(\begin{bmatrix} \Gamma_{11}/2 & 0 \\ \Gamma_{21} & 0 \end{bmatrix}\right) = 2r^*,$$

which implies that $\|\Gamma^{(2)}\|_\star \leq \sqrt{2r^*} \cdot \|\Gamma^{(2)}\|_{\text{fro}}$ Then by (A.15) we obtain that $\|\Theta\|_{\text{fro}} \leq 3/\sqrt{2} \cdot \sqrt{r^*} \cdot \lambda$. Finally, by triangle inequality for the nuclear norm,

$$\|\Theta\|_\star = \|\Gamma\|_\star \leq \|\Gamma^{(1)}\|_\star + \|\Gamma^{(2)}\|_\star \leq 4 \cdot \|\Gamma^{(2)}\|_\star \leq 4\sqrt{2r^*}\|\Gamma^{(2)}\|_{\text{fro}} = 12r^*\lambda.$$

Thus we conclude the proof of Theorem 4.3.

□

B. Proof of Auxiliary Results

B.1. Proof of Lemma A.1

Proof. By definition of the loss function L in (3.4), we have

$$\nabla L(\mu\beta^*) = 2\mu\beta^* - \frac{2}{n}\sum_{i=1}^n \widetilde{Y}_i \cdot \widetilde{S}(X_i) = \mathbb{E}\big[2Y_i \cdot S(X_i)\big] - \frac{2}{n}\sum_{i=1}^n \widetilde{Y}_i \cdot \widetilde{S}(X_i).$$

By triangle inequality,

$$\|\nabla L(\mu\beta^*)\|_\infty \leq \Big\|\mathbb{E}\big[2Y \cdot S(X)\big] - \mathbb{E}\big[2\widetilde{Y} \cdot \widetilde{S}(X)\big]\Big\|_\infty + \Big\|\mathbb{E}\big[2\widetilde{Y} \cdot \widetilde{S}(X)\big] - \frac{2}{n}\sum_{i=1}^n \widetilde{Y}_i \cdot \widetilde{S}(X_i)\Big\|_\infty. \tag{B.1}$$

For any $j \in [d]$, by the definition of the truncated response $\widetilde{Y}$ and truncated score $\widetilde{S}$, we obtain

$$\begin{aligned}
\big|\mathbb{E}\big[\widetilde{Y} \cdot \widetilde{S}_j(X)\big] - \mathbb{E}\big[Y \cdot S_j(X)\big]\big| &\leq \Big|\mathbb{E}\Big\{\widetilde{Y} \cdot \big[\widetilde{S}_j(X) - S_j(X)\big]\Big\}\Big| + \big|\mathbb{E}\big[(\widetilde{Y} - Y) \cdot S_j(X)\big]\big| \\
&= \underbrace{\big|\mathbb{E}\big[\widetilde{Y} \cdot S_j(X) \cdot \mathbb{1}\{|S_j(X)| > \tau\}\big]\big|}_{a_1} + \underbrace{\big|\mathbb{E}\big[Y \cdot S_j(X) \cdot \mathbb{1}\{|Y| > \tau\}\big]\big|}_{a_2}.
\end{aligned} \tag{B.2}$$

By Cauchy-Schwarz inequality, we have

$$\begin{aligned}
a_1^2 &\leq \mathbb{E}\big[\widetilde{Y}^2 S_j^2(X)\big] \cdot \mathbb{P}\big[|S_j(X)| \geq \tau\big] \\
&\leq \sqrt{\mathbb{E}(\widetilde{Y}^4) \cdot \mathbb{E}\big[S_j^4(X)\big]} \cdot \mathbb{E}\big[S_j^4(X)\big] \cdot \tau^{-4} \\
&= M^2 \cdot \tau^{-4},
\end{aligned} \tag{B.3}$$

where the second inequality follows from Chebyshev's inequality. Similarly, for a_2 we have

$$\begin{aligned}
a_2^2 &\leq \mathbb{E}\big[Y^2 S_j^2(X)\big] \cdot \mathbb{P}\big(|Y| \geq \tau\big) \\
&\leq \sqrt{\mathbb{E}(\widetilde{Y}^4) \cdot \mathbb{E}\big[S_j^4(X)\big]} \cdot \mathbb{E}(Y^4) \cdot \tau^{-4} \\
&\leq M^2 \cdot \tau^{-4}.
\end{aligned} \tag{B.4}$$

Thus combining (B.2), (B.3), and (B.4), we conclude that

$$\Big|\mathbb{E}\big[\widetilde{Y}\cdot\widetilde{S}_j(X)\big]-\mathbb{E}\big[Y\cdot S_j(X)\big]\Big|\le a_1+a_2\le 2M\cdot\tau^{-2}$$

for all $j\in[d]$. Thus choosing $\tau=2(M\cdot n/\log d)^{1/4}$, we have

$$\Big\|\mathbb{E}\big[\widetilde{Y}\cdot\widetilde{S}_j(X)\big]-\mathbb{E}\big[Y\cdot S_j(X)\big]\Big\|_\infty\le 1/2\cdot\sqrt{M\cdot\log d/n}. \tag{B.5}$$

Furthermore, under Assumption 4.1, the variance of $\widetilde{Y}\cdot\cdot\widetilde{S}_j(X)$ is bounded by

$$\mathrm{Var}[\widetilde{Y}\cdot\widetilde{S}_j(X)]\le\mathbb{E}[\widetilde{Y}^2\cdot\widetilde{S}_j^2(X)]\le\mathbb{E}[Y^2\cdot S_j^2(X)]\le\sqrt{\mathbb{E}(Y^4)\cdot\mathbb{E}[S_j^4(X)]}\le M.$$

Thus for the second term in (B.1), since $|\widetilde{Y}\cdot\widetilde{S}_j(X)|\le\tau^2$, by the Bernstein inequality in (Boucheron et al., 2013) (Theorem 2.10), for any $j\in[d]$ and any $t>0$, we have

$$\mathbb{P}\bigg\{\Big|\frac{1}{n}\sum_{i=1}^n\widetilde{Y}_i\cdot\widetilde{S}_j(X_i)-\mathbb{E}\big[\widetilde{Y}\cdot\widetilde{S}_j(X)\big]\Big|\ge\sqrt{\frac{2M\cdot t}{n}}+\frac{\tau^2\cdot t}{3n}\bigg\}\le\exp(-t). \tag{B.6}$$

Taking union bound over $j\in[t]$ in (B.6) yields

$$\mathbb{P}\bigg\{\Big\|\frac{1}{n}\sum_{i=1}^n\widetilde{Y}_i\cdot\widetilde{S}_j(X_i)-\mathbb{E}\big[\widetilde{Y}\cdot\widetilde{S}_j(X)\big]\Big\|_\infty\ge\sqrt{\frac{2M\cdot t}{n}}+\frac{\tau^2\cdot t}{3n}\bigg\}\le\exp(-t+\log d). \tag{B.7}$$

Finally, we plug in $\tau=2(M\cdot n/\log d)^{1/4}$ and set $t=3\log d$ in (B.7) to obtain that

$$\Big\|\frac{1}{n}\sum_{i=1}^n\widetilde{Y}_i\cdot\widetilde{S}_j(X_i)-\mathbb{E}\big[\widetilde{Y}\cdot\widetilde{S}_j(X)\big]\Big\|_\infty\le(4+\sqrt{6})\sqrt{\frac{M\cdot\log d}{n}} \tag{B.8}$$

with probability at least $1-d^{-2}$. Finally, combining (B.1), (B.5), and (B.8), we conclude the proof. □

B.2. Proof of Lemma A.2

Proof. For loss function L defined in (3.4) in the matrix setting, we have

$$\nabla L(\mu\beta^*)=2\mu\beta^*-\frac{2}{\kappa\cdot n}\sum_{i=1}^n\psi\big[\kappa\cdot Y_i\cdot S(X_i)\big]=2\mathbb{E}[Y\cdot S(X)]-\frac{2}{\kappa\cdot n}\sum_{i=1}^n\psi\big[\kappa\cdot Y_i\cdot S(X_i)\big]. \tag{B.9}$$

Here the last equality follows from the generalized Stein's identity. In the sequel, we apply results in (Minsker, 2016) to bound $\|\nabla L(\mu\beta^*)\|_{\mathrm{op}}$. To begin with, we first consider the operator norm of $\mathbb{E}[Y^2\cdot S(X)S(X)^\top]\in\mathbb{R}^{d_1\times d_2}$ and $\mathbb{E}[Y^2\cdot S(X)^\top S(X)]\in\mathbb{R}^{d_2\times d_2}$. For notational simplicity, we denote by $S_{j,\cdot}(\cdot)\in\mathbb{R}^{d_2}$ $S_{\cdot,k}(\cdot)\in\mathbb{R}^{d_1}$ the j-th row and k-the column of the score function $S(\cdot)$, respectively. For any $u\in\mathcal{S}^{d_1-1}$, by Cauchy-Schwarz inequality we have

$$\mathbb{E}[Y^2\cdot u^\top S(X)S(X)^\top u]=\sum_{k=1}^{d_2}\mathbb{E}\big\{[Y^2\cdot S_{\cdot,k}(X)^\top u]^2\big\}\le d_2\cdot\sqrt{\mathbb{E}(Y^4)\cdot\mathbb{E}\big\{[S_{\cdot,1}(X)^\top u]^4\big\}}, \tag{B.10}$$

where we use the fact that the entries of $S(X)$ are i.i.d. Since $\mathbb{E}[S_{ij}(X)]=0$ and $\mathbb{E}[S_{ij}^4(X)]\le M$, by Cauchy-Schwarz inequality we obtain that

$$\begin{aligned}\mathbb{E}\big\{[S_{\cdot,1}(X)^\top u]^4\big\}&=\sum_{j_1=1}^d\sum_{j_2=1}^d\mathbb{E}[S_{j_1,1}(X)^2\cdot S_{j_2,1}^2(X)]\cdot u_{j_i}^2u_{j_2}^2\\&\le\sum_{j_1=1}^d\sum_{j_2=1}^d\sqrt{\mathbb{E}[S_{j_1,1}^4(X)]\cdot\mathbb{E}[S_{j_2,1}^4(X)]}\cdot u_{j_i}^2u_{j_2}^2\le M\sum_{j_1=1}^d\sum_{j_2=1}^d u_{j_i}^2u_{j_2}^2=M.\end{aligned} \tag{B.11}$$

Thus combining (B.10) and (B.11) we obtain that

$$\mathbb{E}[Y^2 \cdot u^\top S(X)S(X)^\top u] \le d_2 \cdot M,$$

which implies that $\|\mathbb{E}[Y^2 \cdot S(X)S(X)^\top]\|_{\mathrm{op}} \le d_2 \cdot M$. Similarly, we obtain $\|\mathbb{E}[Y^2 \cdot S(X)^\top S(X)]\|_{\mathrm{op}} \le d_1 \cdot M$. Thus by Corollary 3.1 in (Minsker, 2016), we have

$$\mathbb{P}\Big\{\Big\|\frac{1}{\kappa \cdot n}\sum_{i=1}^n \psi\big[\kappa \cdot Y_i \cdot S(X_i)\big] - \mathbb{E}[Y \cdot S(X)]\Big\|_{\mathrm{op}} \ge \frac{t}{\sqrt{n}}\Big\} \le 2(d_1+d_2)\exp\big[-\kappa t\sqrt{n} + \kappa^2(d_1+d_2)M/2\big] \quad \text{(B.12)}$$

for any $t > 0$ and $\kappa > 0$. We set $\kappa = 2\sqrt{n \cdot \log(d_1+d_2)}/\sqrt{(d_1+d_2)M}$ and $t = \sqrt{(d_1+d_2)M} \cdot s$ in (B.12), which implies that

$$\begin{aligned}&\mathbb{P}\Big\{\Big\|\frac{1}{\kappa \cdot n}\sum_{i=1}^n \psi\big[\kappa \cdot Y_i \cdot S(X_i)\big] - \mathbb{E}[Y \cdot S(X)]\Big\|_{\mathrm{op}} \ge \sqrt{\frac{(d_1+d_2)M}{n}} \cdot s\Big\} \\ &\quad \le 2(d_1+d_2)\exp\big[-2\sqrt{\log(d_1+d_2)} \cdot s + 2 \cdot \log(d_1+d_2)\big]. \quad \text{(B.13)}\end{aligned}$$

Now we set $s = 3 \cdot \sqrt{\log(d_1+d_2)}$, which implies that the right-hand side of (B.13) is less than

$$2(d_1+d_2)\exp\big[-6\log(d_1+d_2) + 2 \cdot \log(d_1+d_2)\big] \le (d_1+d_2)^2 \cdot \exp\big[-4 \cdot \log(d_1+d_2)\big] = (d_1+d_2)^{-2}.$$

Therefore, combining (B.9) and (B.13) we conclude that

$$\|\nabla L(\mu\beta^*)\|_{\mathrm{op}} \le 6\sqrt{(d_1+d_2) \cdot M/n} \cdot \sqrt{\log(d_1+d_2)}$$

with probability at least $1 - (d_1+d_2)^{-2}$, which concludes the proof.

□