A Proof of the Main Results

In this section we prove the main results. We first prove that the proposed estimators achieve near-optimal statistical rates of convergence. Then we prove the supporting lemma on our data-driven approach of truncation.

A.1 Proof of Theorem 3.2

Proof. We denote by $\widehat{W}$ the solution of the convex program in (3.3). Also, let $W^* = \beta^* \beta^{*\top}$. In the following, we establish an upper bound for $\|\widehat{W} - W^*\|_2$.

Since W^* is feasible for the optimization problem in (3.3), we have

$$\langle \widehat{W}, \widetilde{\Sigma} \rangle - \lambda \|\widehat{W}\|_1 \geq \langle W^*, \widetilde{\Sigma} \rangle - \lambda \|W^*\|_1. \tag{A.1}$$

We denote $\Sigma^* = \mathbb{E}[Y \cdot T(X)]$. Note that β^* is the leading eigenvector of Σ^*. Then (A.1) is equivalent to

$$\langle \widehat{W} - W^*, \widetilde{\Sigma} - \Sigma^* \rangle - \lambda \|\widehat{W}\|_1 + \lambda \|W^*\|_1 \geq \langle \Sigma^*, W^* - \widehat{W} \rangle. \tag{A.2}$$

The following Lemma in [33] establishes an upper bound for the first term on the left-hand side of (A.2).

Lemma A.1. Let $\Omega \in \mathbb{R}^{d \times d}$ be a symmetric matrix and let $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_d$ the eigenvalues of Ω in descending order. For any $\ell \in [d-1]$ such that $\lambda_\ell - \lambda_{\ell+1} > 0$, let $\Pi_\ell \in \mathbb{R}^{d \times d}$ be the projection matrix for the subspace spanned by the eigenvectors of Ω corresponding to $\lambda_1, \ldots, \lambda_\ell$. Then for any $\Lambda \in \mathbb{R}^{d \times d}$ satisfying $0 \preceq \lambda \preceq I_d$ and $\mathrm{Trace}(\Lambda) = k$, we have

$$(\lambda_\ell - \lambda_{\ell+1}) \cdot \|\Pi_k - \Lambda\|_{\mathrm{F}}^2 \leq 2 \langle \Omega, \Pi_k - \Lambda \rangle.$$

Note that W^* is the projection matrix for the subspace spanned by β^*. Applying Lemma A.1 to Σ^* with $\ell = 1$, we have

$$\langle \Sigma^*, W^* - \widehat{W} \rangle \geq C_0/2 \cdot \|\widehat{W} - W^*\|_{\mathrm{F}}^2, \tag{A.3}$$

where $C_0 > 0$ is defined in (3.2). In addition, by Hölder's inequality, we have

$$\langle \widehat{W} - W^*, \widetilde{\Sigma} - \Sigma^* \rangle \leq \|\widetilde{\Sigma} - \Sigma^*\|_\infty \cdot \|\widehat{W} - W^*\|_1. \tag{A.4}$$

In what follows, we bound $\|\widetilde{\Sigma} - \Sigma^*\|_\infty$.

Lemma A.2. Let $\widetilde{\Sigma}$ be defined in (3.5) and we define $\Sigma^* = \mathbb{E}[Y \cdot T(X)]$. Under Assumption 3.1, for any truncation level $\tau > 0$ in (3.4), with probablity at least $1 - d^{-2}$, we have

$$\|\widetilde{\Sigma} - \Sigma^*\|_\infty \leq 9M \cdot \tau^{-3} + 2\tau^3 \cdot \log d/n + 2\sqrt{5M \cdot \log d/n}. \tag{A.5}$$

Proof. See §A.3 for a detailed proof. □

By this lemma, if we set $\tau = (1.5Mn/\log d)^{1/6}$, then with probability at least $1 - d^{-1}$,

$$\|\widetilde{\Sigma} - \Sigma^*\|_\infty \leq (2\sqrt{5} + 2\sqrt{6}) \cdot \sqrt{M \log d/n} \leq 10\sqrt{M \log d/n}. \tag{A.6}$$

Thus by setting $\lambda = 10\sqrt{M \log d/n}$ we have $\|\widetilde{\Sigma} - \Sigma^*\|_\infty \leq \lambda$ with probability at least $1 - d^{-2}$.

Then combining (A.2), (A.3), and (A.4) we have

$$\lambda \Big(\|\widehat{W} - W^*\|_1 - \|\widehat{W}\|_1 + \|W^*\|_1 \Big) \geq C_0/2 \cdot \|\widehat{W} - W^*\|_{\mathrm{F}}^2. \tag{A.7}$$

Note that $W^* = \beta^* \beta^{*\top}$ and that β^* is s^*-sparse. We denote the support of W^* by $\mathcal{J}$, which is given by

$$\mathcal{J} = \big\{ (j,k) \in [d] \times [d] \colon \beta_j^* \cdot \beta_k^* \neq 0 \big\}.$$

Then by seperation of the ℓ_1-norm, we have

$$\|\widehat{W}\|_1 = \|\widehat{W}_{\mathcal{J}}\|_1 + \|\widehat{W}_{\mathcal{J}^c}\|_1 \ \text{ and } \ \|\widehat{W} - W^*\|_1 = \|\widehat{W}_{\mathcal{J}} - W^*_{\mathcal{J}}\|_1 + \|\widehat{W}_{\mathcal{J}^c}\|_1,$$