Supplementary Material for "Efficient Decoding from Heterogeneous 1-Bit Compressive Measurements over Networks"

Canyi Chen, Liping Zhu, and Zhengtian Zhu

University of Michigan, Renmin University of China and Tongji University

This supplementary material provides all additional simulation results and the proof for the theoretical results in the main text. Throughout, we denote $\widehat{\mathbf{\Sigma}}_j = 1/n \sum_{i=1}^n \mathbf{x}_{i,j}\mathbf{x}_{i,j}^{\mathrm{T}}$ and $\widehat{\mathbf{\Sigma}} = 1/m \sum_{j=1}^m \widehat{\mathbf{\Sigma}}_j$.

Contents

A. Additional simulation results

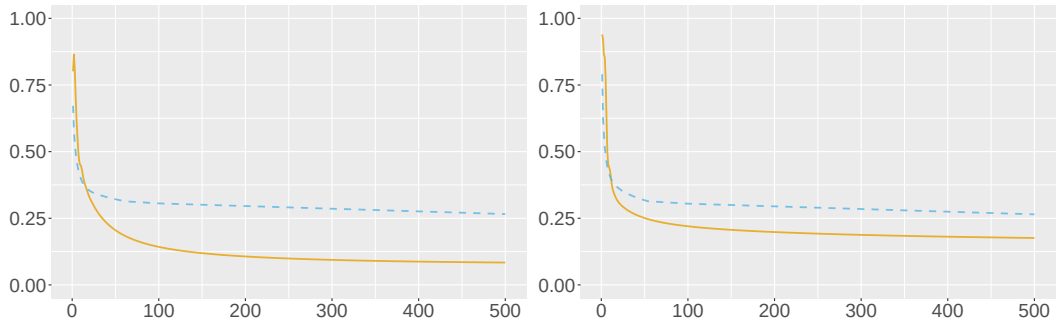
A.1 Effect of iteration

To illustrate the convergence of our method, we created convergence plots using Algorithm 1 to perform decentralized 1-bit CS with the default setting. As depicted in Figure A.1, we observe a rapid and linear decrease in the ℓ_2-error of our method. When examining the F_1-score, we observe that it exhibits a rapid initial increase, followed by a mild decline before stabilizing from Figure A.2.

We have also incorporated the subGD method into our comparative analysis. Regarding ℓ_2-error, the initial convergence rates of subGD and our method are comparable. However, after sufficient iteration cycles, our method ultimately yields lower ℓ_2-error than subGD. For the F_1-score, subGD maintains a constant F_1-score throughout the iterations as expected because it yields a dense estimate.

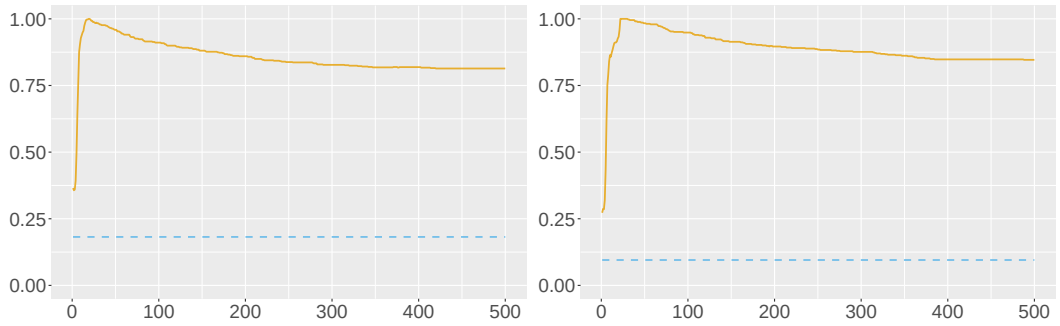
A.2 Effect of initial estimate

To investigate the impact of zero initialization on algorithm performance, we replicated the simulation settings from Section A.1 while replacing the initialization scheme with zero-initialized values. Corresponding results



(a) $p = 100$ (b) $p = 200$

Figure A.1: Convergence plot for solving decentralized 1-bit CS with validated tuning parameters using Algorithm 1 (solid line) and subGD (dashed line). The horizontal axis represents the number of iterations T, and the vertical axis denotes the ℓ_2-error.

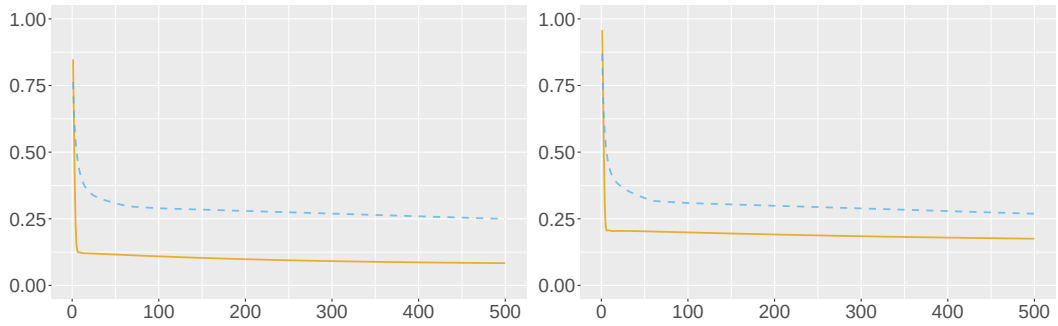


(a) $p = 100$ (b) $p = 200$

Figure A.2: Convergence plot for solving decentralized 1-bit CS with validated tuning parameters using Algorithm 1 (solid line) and subGD (dashed line). The horizontal axis represents the number of iterations T, and the vertical axis denotes the F_1-score.

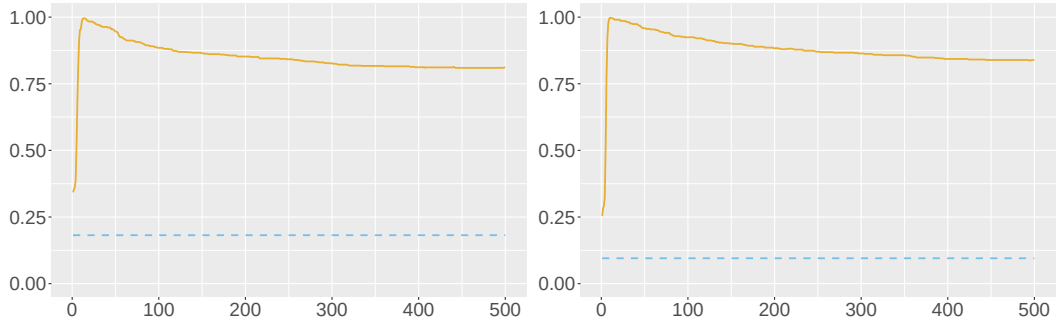
are presented in Figure A.3 and Figure A.4. The results demonstrate that our method still exhibits fast initial reduction in ℓ_2-error under zero initialization. However, after sufficient iterations, the default initialization strategy yields marginally lower ℓ_2-error compared to zero initialization. Considering F_1-score, zero initialization exerts minimal influence on F_1-score performance.

The impact of zero-mean initialization on the subGD method is relatively negligible. After sufficient iterations, the subGD approach with default initialization yields marginally lower ℓ_2-error than its zero-initialized counterpart.



(a) $p = 100$ (b) $p = 200$

Figure A.3: Convergence plot for solving decentralized 1-bit CS with validated tuning parameters using Algorithm 1 (solid line) and subGD (dashed line). Both algorithms are zero-initialized. The horizontal axis represents the number of iterations T, and the vertical axis denotes the ℓ_2-error.



(a) $p = 100$ (b) $p = 200$

Figure A.4: Convergence plot for solving decentralized 1-bit CS with validated tuning parameters using Algorithm 1 (solid line) and subGD (dashed line). Both algorithms are zero-initialized. The horizontal axis represents the number of iterations T, and the vertical axis denotes the F_1-score.

A.3 Effect of high-dimensional covariates

To further investigate the performance of various methods under high-dimensional settings, we conducted simulations with $p = 600$, $n = 100$, and computing node counts $m \in \{2, 3, 4, 5, 6\}$. The results are visualized in Figure A.5. The plots indicate that with fewer computing nodes, i.e., smaller sample sizes, distributed algorithms exhibit higher ℓ_2-error and lower F_1-score, demonstrating larger performance gaps compared to the pooled method. As the number of nodes increases, the ℓ_2-error of distributed algorithms approaches that of the pooled method. Although the F_1-score shows an upward trend, it remains differentiated from the pooled benchmark. Among all methods, the proposed approach maintains the best-performing results in this high-dimensional scenario. We note that the gap between our

estimate and the pooled estimate is possibly due to the optimization error since we fix the iteration budget $T = 500$.

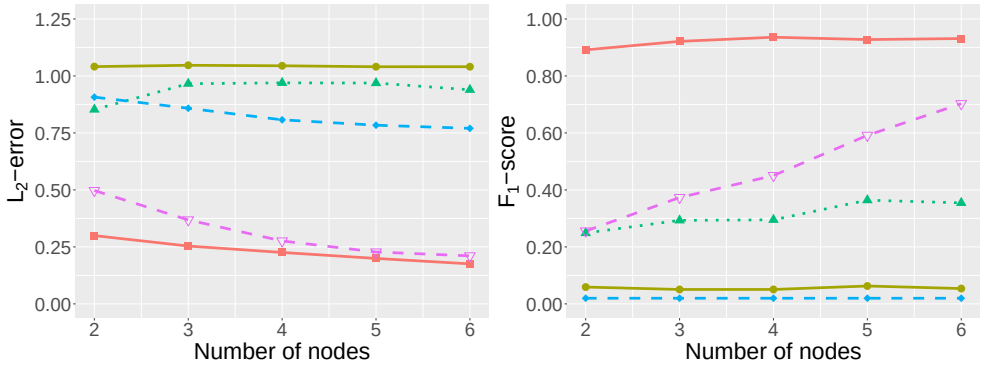


Figure A.5: The ℓ_2-errors (left panel) and F_1-scores (right panel) of the Pooled (—■—), Local (—●—), Avg (···▲···), subGD (- -◆- -), and our estimates (- -▽- -) under different number of nodes with local sample size $n = 100$.

A.4 Effect of non-elliptical distributed covariates

In this subsection, we examine the performance of our proposal when the covariates are non-elliptically distributed and heavy-tailed. We employ the identical simulation settings from Section 4.2, except for the distribution of covariates. Let $\mathbf{x}_{i,j} = (X_{i,j,1}, \ldots, X_{i,j,p})^{\mathrm{T}}$. The first two entries of $\mathbf{x}_{i,j}$ are generated from t copula with one degree of freedom and Pearson correlation coefficient being 0.1. For other entries of $\mathbf{x}_{i,j}$, $(X_{i,j,2k+1}, X_{i,j,2k+2})$ are all independent copies of $(X_{i,j,1}, X_{i,j,2})$, for $k = 1, \ldots, (p/2) - 1$.

All corresponding results are shown in Figure A.6. When covariates follow non-elliptical distributions, all methods experience performance degra-

dation. As the sample size increases, the proposed method exhibits marked improvement and becomes the best-performing approach among all decentralized algorithms. This shows the robustness of our method against the violence of the linearity and sub-Gaussian conditions.

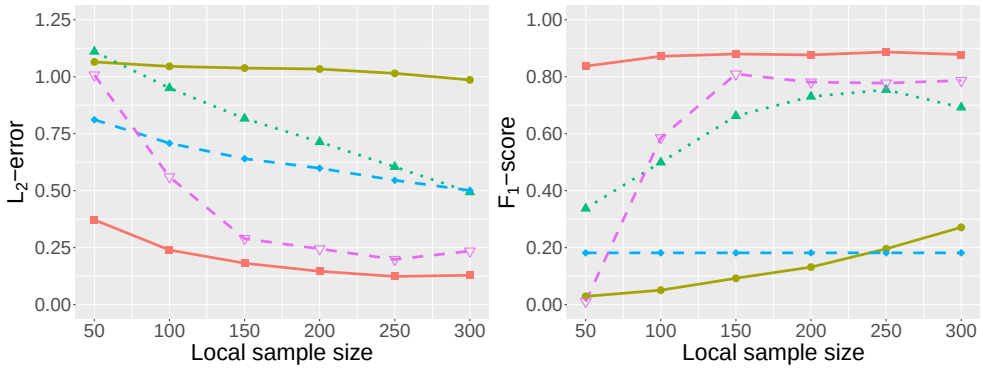


Figure A.6: The ℓ_2-errors (left panel) and F_1-scores (right panel) of the Pooled (—■—), Local (—●—), Avg (···▲···), subGD (- -◆- -), and our estimates (- -▽- -) under non-elliptical distributed covariates.

A.5 Performance of the worst node

In this subsection, we report the performance of the worst node of all distributed algorithms. Note that the worst performance is identical to the average performance for the Avg method. Therefore, we do not report the worst node corresponding to the Avg method. All simulation settings align with those in Section 4.2 of the main text. Figure A.7 reveals that for each method, its maximum ℓ_2-error and minimum F_1-score closely approximate the method's original ℓ_2-error and F_1-score, indicating that the overall

performance is not substantially compromised by the worst node's results. It should be noted that subGD produces dense outputs, resulting in identical F_1-scores across all nodes. Consequently, only subGD's inherent F_1-score is displayed in the figure.

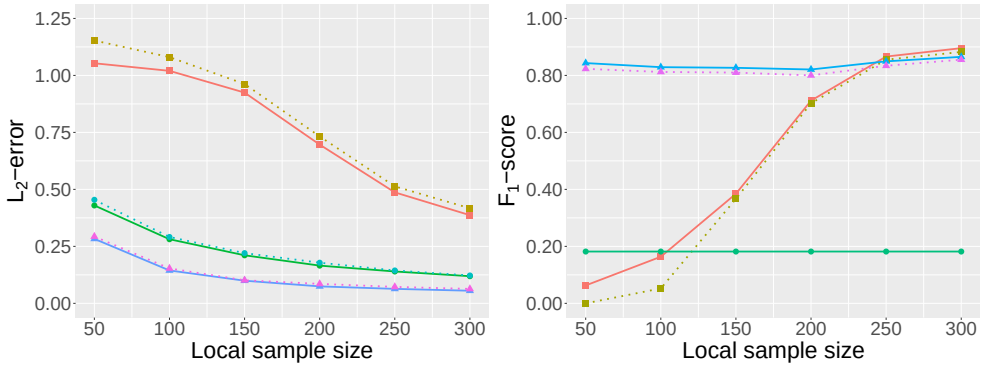


Figure A.7: The left panel shows the ℓ_2-errors (solid line) and the maximum ℓ_2-errors (dotted line) of the Local, subGD, our estimates. The right panel shows the F_1-errors (solid line) and the minimum F_1-errors (dotted line) of the Local, subGD, our estimates. Square points represent the local method, circle points represent the subGD method, and triangle points represent the proposed method.

A.6 Effect of heterogeneity

We investigate the effect of heterogeneity. In particular, we consider three types of heterogeneity. In the first type, each node randomly sets σ_j^2 and ρ_j from $\{1, 3\}$ and $\{0.1, 0.3\}$. For $\sigma_j^2 = 3$, this typically corresponds to a SNR $= 0.557$. For $\sigma_j^2 = 1$, this typically corresponds to a SNR $= 0.869$. In the second type, we randomly select the noise distribution from the standard normal distribution, Student's t distribution with two degrees of freedom,

and the standard Cauchy distribution. In the third type, the probability of sign flips at each local node is randomly chosen from $\{0.05, 0.1\}$.

We report the results in Figures A.8 to A.10. We can see that the pooled estimate attains the smallest ℓ_2-error and largest F_1-score. However, it requires pooling all the data into a single node, which may not be practical in a decentralized network and could incur huge communication costs and privacy concerns. Our proposed decentralized estimate performs the best of all the decentralized estimates in terms of both the ℓ_2-error and the F_1-score. Various heterogeneities did not affect our proposal much.

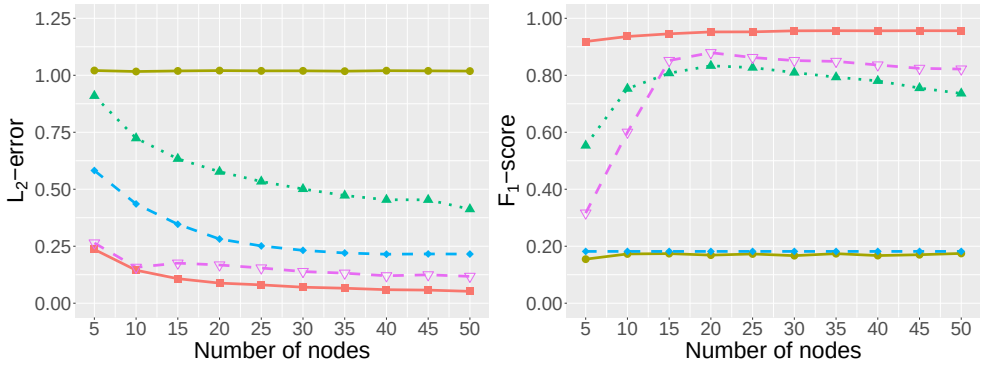


Figure A.8: The ℓ_2-errors (left panel) and F_1-scores (right panel) of the Pooled (—■—), Local (—●—), Avg (···▲···), subGD (- - ◆ - -), and our estimates (- - ▽ - -) under heterogeneous covariates.

B. An Application to Reconstruct EEG Signals

In this section, we evaluate the empirical performance of the proposed method using the EEG dataset SEED (Zheng and Lu, 2015). SEED is a

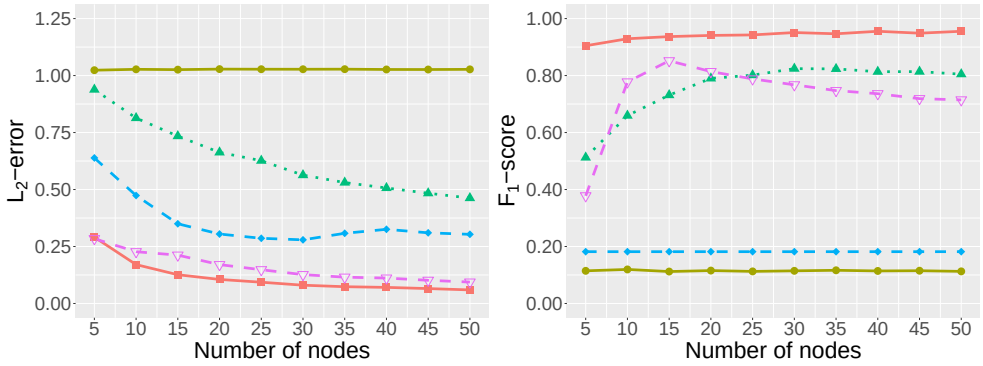


Figure A.9: The ℓ_2-errors (left panel) and F_1-scores (right panel) of the Pooled (—■—), Local (—●—), Avg (···▲···), subGD (- -◆- -), and our estimates (- -▽- -) under heterogeneous noises.

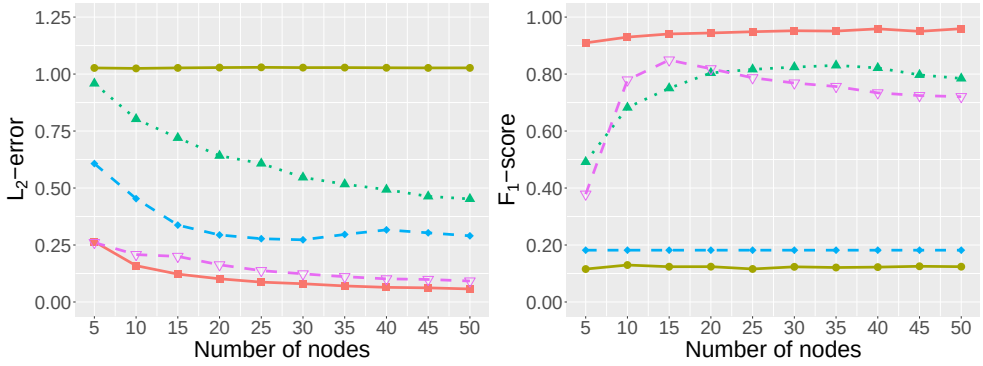


Figure A.10: The ℓ_2-errors (left panel) and F_1-scores (right panel) of the Pooled (—■—), Local (—●—), Avg (···▲···), subGD (- -◆- -), and our estimates (- -▽- -) under heterogeneous probabilities of sign flips.

dataset dedicated to emotion recognition research, collected, curated, and maintained by the BCMI Laboratory at Shanghai Jiao Tong University. The complete dataset is publicly accessible at: `https://bcmi.sjtu.edu.cn/home/seed/`. The SEED dataset consists of EEG signals acquired from 15 participants during exposure to curated film clips eliciting specific emotional states (positive, negative, and neutral). Subjects' neural activities are captured via a 62-channel EEG system.

We consider constructing the following two experiments to demonstrate the effectiveness of our method in reconstructing the EEG signal:

- *Experiment 1.* Considering film clips representing three distinct emotional categories, we perform signal recovery for data from all 62 electrodes on a single test subject. In this setup, the number of network nodes in the distributed system is $m = 62$. To maintain a consistent signal length, we extract EEG signals every 1 second. Since the sampling frequency is 200 Hz, the dimension of every signal is $p = 200$. The total sample size is $N = 93000$, with each node allocated $n = 1500$ samples.

- *Experiment 2.* Considering the same film clip, we perform signal recovery for data from a single electrode across all 15 subjects. Based on previous research (Zheng and Lu, 2015), we selected a film clip eliciting positive emotions, with electrode channels AF3, F7, and T7 chosen to best reflect signal recovery outcomes. In this setup, the number of nodes in the distributed system is $m = 15$. Similarly, we extract EEG signals every 0.2 second to maintain a consistent signal length. Here

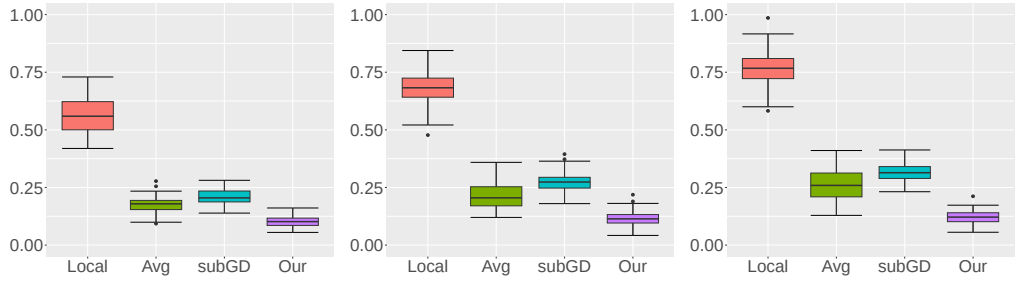
$p = 40$, $N = 4500$, and each node is allocated $n = 300$ samples.

At each node, the preprocessed EEG signals are transformed into 1-bit measurements for distributed reconstruction. The connection probability of the network is $p_c = 0.3$. The sensing vectors are sampled from $\mathcal{N}(0, \mathbf{\Sigma})$ where $\mathbf{\Sigma}_{\ell k} = 0.3^{|\ell - k|}$, and the noise terms $\varepsilon_{i,j}$ are drawn from $\mathcal{N}(0, 1)$. The probability of sign-flips is characterized by $\mathbb{P}(\xi_{i,j} = 1) = 0.9$. The final 1-bit measurements are generated based on (1.1). The true signal is defined as the average of these preprocessed EEG signals across all nodes (Yao, 2001). We repeat each experiment 100 times to provide a reliable assessment.

Figure B.11 summarizes the results of Experiment 1. As shown, under the three film categories representing different emotional states, the Local method demonstrates the poorest signal reconstruction performance. The Avg and subGD methods show comparable effectiveness, with Avg slightly outperforming subGD. The proposed method in this study achieves the best reconstruction results. Consistent with prior research, EEG signals corresponding to positive emotion film clips are more amenable to reconstruction, with all methods yielding lower errors compared to those observed in neutral and negative emotion conditions (Zheng and Lu, 2015).

Compared to Experiment 1, Experiment 2 was a more idealized setup as we selected a positive film clip with more prominent EEG signal activity and

three electrode channels (AF3, F7, T7) previously identified as significantly contributing to emotional judgment (Zheng and Lu, 2015). Figure B.12 presents the results of Experiment 2, where the performance of both the Local and Avg methods showed substantial improvement, while the effectiveness of the subGD method changed minimally. Our method remained the best-performing approach.



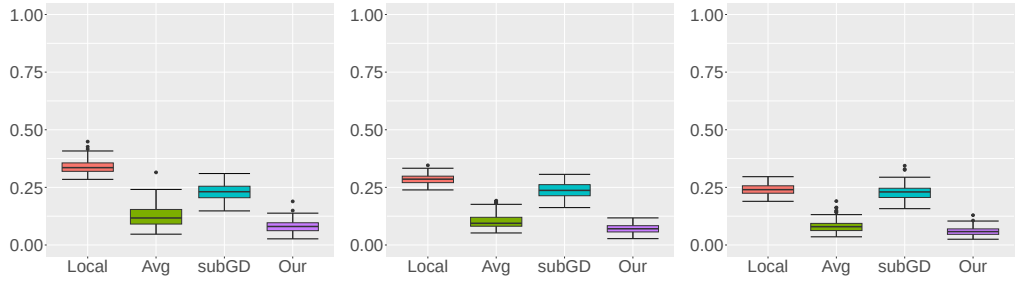
(a) Positive (b) Neutral (c) Negative

Figure B.11: ℓ_2-error of the Local (red), Avg (green), subGD (blue), and Our (purple) under different film clips positive (a), neutral (b) and negative (c), with $m = 62$, $p = 200$ and $N = 93000$. The real signal is an averaged 1s segmented EEG data from the same subject under each film clip, recorded by the 62-channel EEG electrode system.

C. Implication of the generalized ADMM Algorithm and Implementation Recommendation

C.1 Implication of the generalized ADMM algorithm

The local sample size n will affect the convergence factor γ of the generalized ADMM algorithm in Proposition 1, which in turn influences the convergence rate of the algorithm. We discuss the impact below.



(a) Channel AF3 (b) Channel F7 (c) Channel T7

Figure B.12: ℓ_2-error of the Local (red), Avg (green), subGD (blue), and Our (purple) under different channels AF3 (a), F7 (b) and T7 (c), with $m = 15$, $p = 40$ and $N = 4500$. The real signal is an average of 0.2 seconds of segmented EEG data recorded by each channel above from 15 subjects under the same positive film clips.

We introduce several notations and definitions from graph theory. Define the degree matrix $\mathbf{D} \overset{\text{def}}{=} \text{diag}(|\mathcal{N}(1)|, \ldots, |\mathcal{N}(m)|) = \text{diag}(\sum_k |\mathbf{W}_{1k}|, \ldots, \sum_k |\mathbf{W}_{mk}|) \in \mathbb{R}^{m\times m}$. Further, let $\mathbf{G} \overset{\text{def}}{=} \text{diag}(\rho_\ell) \otimes \mathbf{I}_p + \tau(\mathbf{D} + \mathbf{W}) \otimes \mathbf{I}_p \in \mathbb{R}^{mp\times mp}$. The quantity $\sigma_{\min}(\boldsymbol{\Upsilon}) > 0$ denotes the minimum nonzero singular value of the linear mapping matrix $\boldsymbol{\Upsilon}$, which is induced by the adjacency matrix $\mathbf{W}$, as defined by

$$\begin{bmatrix} \sum_{j\in\mathcal{N}(1)} \left(\mathbf{u}^{(1j)} - \mathbf{u}^{(j1)}\right) \\ \vdots \\ \sum_{j\in\mathcal{N}(m)} \left(\mathbf{u}^{(mj)} - \mathbf{u}^{(jm)}\right) \end{bmatrix} = \boldsymbol{\Upsilon}\mathbf{u},$$

where $\mathbf{u} = ((\mathbf{u}^{(1j)\text{T}} : j \in \mathcal{N}(1))^\text{T}, \ldots, (\mathbf{u}^{(mj)\text{T}} : j \in \mathcal{N}(m))^\text{T})^\text{T} \in \mathbb{R}^{p\sum_{j=1}^m |\mathcal{N}(j)|}$. The minimum nonzero singular value of $\boldsymbol{\Upsilon}$ is the square root of the minimum nonzero singular value of the Laplacian matrix $\mathbf{L} = \mathbf{D} - \mathbf{W}$, i.e., $\sigma_{\min}(\boldsymbol{\Upsilon}) = \{\sigma_{\min}(\mathbf{L})\}^{1/2}$ (Fiedler, 1973; Chung, 1997). The value of $\sigma_{\min}(\mathbf{L})$ measures the network connectivity with larger $\sigma_{\min}(\mathbf{L})$ corresponding to stronger

connectivity.

According to the proof of Proposition 1, the convergence factor $\gamma = 1/(1+\delta)$, where

$$\begin{aligned}\delta \;\stackrel{\text{def}}{=}\; & \min\Big[\tau\sigma^2_{\min}(\boldsymbol{\Upsilon})/4\lambda_{\min}\{(\mathbf{G}^{\mathrm{T}}\mathbf{G})^{-1}(\mathbf{G}-\widetilde{\boldsymbol{\Sigma}})\},\\ & 1/2\lambda_{\min}\{(\mathbf{G}+1/2\widetilde{\boldsymbol{\Sigma}})^{-1}\widetilde{\boldsymbol{\Sigma}}\}, \tau\sigma^2_{\min}(\boldsymbol{\Upsilon})/\{4\lambda_{\max}(\widetilde{\boldsymbol{\Sigma}})\}\Big] > 0.\end{aligned}$$

The parameter δ directly affects γ, as larger values of δ yield a smaller convergence factor, $\gamma = 1/(1+\delta)$.

The formulation of δ has three key implications. First, as the network becomes sparser, $\sigma_{\min}(\boldsymbol{\Upsilon}) = \{\sigma_{\min}(\mathbf{L})\}^{1/2}$ decreases, which in turn reduces δ, leading to a larger convergence factor γ. In our analysis, we treat the minimum nonzero singular value of the Laplacian matrix, $\sigma_{\min}(\mathbf{L})$, as a constant that depends on the underlying network structure. Second, as $\lambda_{\max}(\widetilde{\boldsymbol{\Sigma}})$ increases, δ decreases, thereby increasing γ. Third, as $\lambda_{\min}(\widetilde{\boldsymbol{\Sigma}})$ increases, δ also increases, reducing γ.

To investigate the effect of local sample size n on δ, given the network topology fixed, it suffices to study how $\lambda_{\min}(\widetilde{\boldsymbol{\Sigma}})$ and $\lambda_{\max}(\widetilde{\boldsymbol{\Sigma}})$ depend on n. The local covariance matrix $\widehat{\boldsymbol{\Sigma}}_j$ is defined as $\widehat{\boldsymbol{\Sigma}}_j = (\mathbf{X}^{(j)})^{\mathrm{T}}\mathbf{X}^{(j)}/n$, where $\mathbf{X}^{(j)}$ is the local data matrix at node j. As the local sample size n increases,

the eigenvalues of $\widetilde{\mathbf{\Sigma}}$ converge to their population counterparts, which are bounded away from below and above by assumptions. In particular, by Theorem 5.11 in (Bai and Silverstein, 2010), assuming $p/n \to \mu^2 \in (0, 1)$, we have

$$\lambda_{\min}(\widehat{\mathbf{\Sigma}}_j) \xrightarrow{a.s.} \lambda_{\min}(\mathbf{\Sigma})(1-\mu)^2, \quad \lambda_{\max}(\widehat{\mathbf{\Sigma}}_j) \xrightarrow{a.s.} \lambda_{\max}(\mathbf{\Sigma})(1+\mu)^2.$$

Consequently, as n increases (decreased μ), the minimum nonzero eigenvalue $\lambda_{\min}(\widetilde{\mathbf{\Sigma}})$ could increase, while the maximum eigenvalue $\lambda_{\max}(\widetilde{\mathbf{\Sigma}})$ could decrease. This corresponds to an increase in δ, which leads to a decrease in the convergence factor γ.

To verify this theoretical investigation, let us consider the default data generation process for generating $\mathbf{x}_i$ with $p = 10$, and vary the local sample size $n \in \{100, 200, 400, 800\}$ to see how empirically the minimum and maximum eigenvalues of $\widehat{\mathbf{\Sigma}}^{(1)}$ change over 100 replications. Results are summarized in Figure C.13.

The practical guidance for setting up sensing vectors is to set the minimum eigenvalue of the population covariance matrix $\mathbf{\Sigma}$ larger and the maximum eigenvalue smaller. For a fixed local sample size n, this will lead to a larger $\lambda_{\min}(\widetilde{\mathbf{\Sigma}})$ and a smaller $\lambda_{\max}(\widetilde{\mathbf{\Sigma}})$, which in turn decreases the

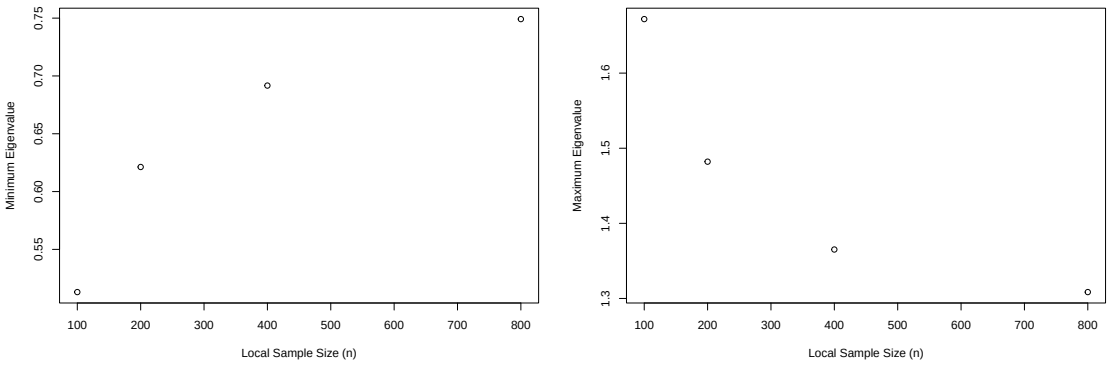


Figure C.13: The minimum (left panel) and maximum (right panel) eigenvalues of the local covariance matrix $\widehat{\boldsymbol{\Sigma}}_1$ under different local sample sizes n.

convergence factor γ and accelerates the convergence of the generalized ADMM algorithm.

C.2 Initial points

We recommend using local estimates as initialization when the local sample size n is sufficiently large, the system is relatively homogeneous, and computational overhead is not a limiting factor. In such settings, local estimates can provide a closer approximation to the global solution and thereby accelerate convergence. Conversely, under heterogeneous systems—where local estimates may be biased or inconsistent with respect to the true global parameter $\boldsymbol{\beta}^*$—we advocate initializing with zero. This approach can reduce the computational burden.

C.3 Choice of the maximum number of iterations T

In our simulations, we set the communication budget to $T = 500$ iterations, which is sufficient for the generalized ADMM and other decentralized algorithms to converge in most cases. Our numerical experiments indicate that the generalized ADMM algorithm typically converges within roughly 100 iterations under the default configuration. The precise number of iterations, however, depends on factors such as the network topology, local sample size, and underlying data distribution. For instance, larger or more complex networks generally require additional iterations to achieve full convergence. Accordingly, we recommend choosing the maximum iteration budget T in light of the application-specific network structure and the available communication resources. That said, providing a universal guideline for the choice of T is challenging. A practical approach is to monitor convergence by periodically evaluating the primal and dual residuals through the network gossip protocol. Such periodic checks help ensure convergence while limiting unnecessary communication.

C.4 Effect of the Lagrangian parameter τ

For the Lagrangian parameter τ, we adopt the choice $\tau = 1$, which works well in all numerical studies, following the convention in Ma and Huang (2017);

Liu et al. (2024). By Proposition 1, the generalized ADMM algorithm will converge linearly for any $\tau > 0$. With sufficient ADMM iterations, the optimization error will be absorbed into the statistical error, as shown by Theorem 1. The Lagrangian parameter τ can sometimes affect the convergence rate of the generalized ADMM algorithm. Using a larger τ slightly slows the convergence rate (i.e., with a larger convergence factor $\gamma \in (0, 1)$), possibly because it overly emphasizes the previous iterate, making the updates more conservative.

There are adaptive strategies for adjusting the parameter τ by monitoring changes in the primal and dual residuals, as discussed in Section 3.4.1 of Boyd (2010). However, such approaches may incur significant additional communication costs in decentralized networks. Instead, based on Theorem 1 of the revised manuscript, we advocate a data-driven approach to selecting τ, if needed. For example, one may apply information criteria or cross-validation by partitioning the local dataset into multiple folds. This strategy avoids the need for gossip consensus at each iteration of the optimization process and instead requires only a single round of consensus after the decentralized optimization is complete. This line of research—developing communication-efficient, adaptive parameter selection methods—has received considerable attention in the optimization community, and we identify it as a promising

direction for future work.

C.5 Implementation for multiple penalty parameters

In our numerical experiments, we implement the generalized ADMM algorithm for multiple penalty parameters by running the algorithm separately for each penalty parameter. The multiple penalty parameters are equally spaced logarithmically over the lower and upper bounds. The optimal λ_N is then selected by cross-validation or Bayesian information criterion via network gossip. This approach allows us to evaluate the performance of the proposed method under different levels of sparsity and regularization. To the best of our knowledge, this is the practice in the decentralized distributed learning literature. This is quite different from the way used in the centralized distributed learning, where the penalty selection can be fully conducted on the central node without further communication overhead. Nevertheless, we use a standard warm up technique to reduce the communication cost. That is to use the last solution of the generalized ADMM algorithm as the initial estimate for the next penalty parameter.

D. Additional Practical Application Scenarios

We discuss additional potential practical applications in areas like *wireless sensor networks (WSNs)*, *multi-agent robotic systems*, and *smart grids*. These applications highlight the need for decentralized, peer-to-peer communication where central aggregation is not feasible due to communication, privacy, or operational constraints.

- *Wireless sensor networks*: WSNs are often deployed for large-scale environmental monitoring (e.g., air quality, forest fire detection) where sensors are distributed across vast areas (Ling and Tian, 2010). In such networks, each node collects data locally and processes it without central aggregation, which is critical for energy conservation. Our proposed decentralized 1bit CS helps nodes predict efficiently under low and high noise environmental conditions while minimizing communication costs (Khedo et al., 2010; Yi et al., 2015). For example, for forest fire detection, it is important to deploy low-cost systems to detect early signs of fire risk without sending raw data to a centralized location (Aslan et al., 2012).

- *Multi-agent robotic systems*: In multi-agent systems, autonomous robots coordinate to perform tasks such as mapping, exploration, or

search-and-rescue missions in complex environments. These robots often operate in decentralized networks where they gather local data and communicate only with nearby agents due to bandwidth constraints or operational limitations (Jiménez et al., 2018). Our decentralized approach is particularly useful in handling uncertainties, including sensor noise or system failures. For example, in disaster zones, where robots may encounter unpredictable terrain or sensor noise, our method may provide robust predictions that account for heavy-tailed noises, enhancing the robots' ability to plan and respond under uncertainty (Lu and Amato, 2019; Rasheed et al., 2022).

- *Smart grids*: In a decentralized smart grid, multiple nodes (such as local microgrids, solar panels, wind turbines, or energy storage systems) operate autonomously, making decisions based on local conditions and information (Talat et al., 2020). Using our decentralized 1bit CS, each node can estimate future energy demand or generation, remaining robust even if other nodes fail. This allows for better preparedness in managing extreme cases, such as peak demand or supply shortages, without relying on a centralized coordinator (Schäfer et al., 2015).

E. Derivations for the generalized ADMM Updates

Similar to Mateos et al. (2010), we develop the updating rules for $(\{\boldsymbol{\beta}^{(j)}\}, \{\mathbf{t}^{(jk)}\}, \{\mathbf{u}^{(jk)}\}, \{\mathbf{v}^{(jk)}\})$ in what follows. We first give the derivations for $(\{\boldsymbol{\beta}^{(j)}\}, \{\mathbf{t}^{(jk)}\}, \{\mathbf{u}^{(jk)}\}, \{\mathbf{v}^{(jk)}\})$. For tth iteration in the loop, given primal variables $(\{\boldsymbol{\beta}^{(j)}\}, \{\mathbf{t}^{(jk)}\})$, the dual variables can be updated with

$$\begin{aligned} \mathbf{u}_{t+1}^{(jk)} &= \mathbf{u}_t^{(jk)} + \tau(\boldsymbol{\beta}_t^{(j)} - \mathbf{t}_t^{(jk)}), \\ \mathbf{v}_{t+1}^{(jk)} &= \mathbf{v}_t^{(jk)} + \tau(\boldsymbol{\beta}_t^{(k)} - \mathbf{t}_t^{(jk)}). \end{aligned} \tag{E.1}$$

For primal variables, we can update them by solving their corresponding subproblems,

$$\begin{aligned} \boldsymbol{\beta}_{t+1}^{(j)} &= \arg\min_{\boldsymbol{\beta}^{(j)}} \frac{1}{2n}|\mathbf{y}^{(j)} - \mathbf{X}^{(j)}\boldsymbol{\beta}^{(j)}|_2^2 + \lambda_N|\boldsymbol{\beta}^{(j)}|_1 + \sum_{k\in\mathcal{N}(j)} \Big(\langle \mathbf{u}_{t+1}^{(jk)} + \mathbf{v}_{t+1}^{(jk)}, \boldsymbol{\beta}^{(j)} \rangle \\ &\quad + \tau/2|\boldsymbol{\beta}^{(j)} - \mathbf{t}_t^{(jk)}|_2^2 + \tau/2|\boldsymbol{\beta}^{(j)} - \mathbf{t}_t^{(kj)}|_2^2 \Big), \\ &= \arg\min_{\boldsymbol{\beta}^{(j)}} \frac{1}{2n}|\mathbf{y}^{(j)} - \mathbf{X}^{(j)}\boldsymbol{\beta}^{(j)}|_2^2 + \lambda_N|\boldsymbol{\beta}^{(j)}|_1 + \sum_{k\in\mathcal{N}(j)} \Big(\langle \mathbf{u}_{t+1}^{(jk)} + \mathbf{v}_{t+1}^{(jk)}, \boldsymbol{\beta}^{(j)} \rangle \\ &\quad + \tau|\boldsymbol{\beta}^{(j)} - (\mathbf{t}_t^{(jk)} + \mathbf{t}_t^{(kj)})/2|_2^2 \Big), \end{aligned} \tag{E.2}$$

$$\begin{aligned} \mathbf{t}_{t+1}^{(jk)} &= \arg\min_{\mathbf{t}^{(jk)}\in\mathbb{R}^p} -\langle \mathbf{u}_{t+1}^{(jk)} + \mathbf{v}_{t+1}^{(jk)}, \mathbf{t}^{(jk)} \rangle + \tau/2|\boldsymbol{\beta}_{t+1}^{(j)} - \mathbf{t}^{(jk)}|_2^2 + \tau/2|\boldsymbol{\beta}_{t+1}^{(k)} - \mathbf{t}^{(jk)}|_2^2 \\ &= (\mathbf{u}_{t+1}^{(jk)} + \mathbf{v}_{t+1}^{(jk)})/(2\tau) + (\boldsymbol{\beta}_{t+1}^{(j)} + \boldsymbol{\beta}_{t+1}^{(k)})/2. \end{aligned} \tag{E.3}$$

If $\mathbf{u}_t^{(jk)} + \mathbf{v}_t^{(jk)} = \mathbf{0}\ \forall j,k$, we have $\mathbf{t}_{t+1}^{(jk)} = (\boldsymbol{\beta}_{t+1}^{(j)} + \boldsymbol{\beta}_{t+1}^{(k)})/2$. This yields that $\mathbf{u}_{t+1}^{(jk)} + \mathbf{v}_{t+1}^{(jk)} = \tau(\boldsymbol{\beta}_t^{(j)} + \boldsymbol{\beta}_t^{(k)} - 2\mathbf{t}_t^{(jk)}) = \mathbf{0}$. By induction, if we setting $\mathbf{u}_0^{(jk)} + \mathbf{v}_0^{(jk)} = \mathbf{0},\ \forall j,k$, the updates (E.1)-(E.3) reduce to

$$\begin{aligned}
\mathbf{u}_{t+1}^{(jk)} &= \mathbf{u}_t^{(jk)} + \tau/2(\boldsymbol{\beta}_t^{(j)} - \boldsymbol{\beta}_t^{(k)}), \\
\mathbf{v}_{t+1}^{(jk)} &= \mathbf{v}_t^{(jk)} + \tau/2(\boldsymbol{\beta}_t^{(k)} - \boldsymbol{\beta}_t^{(j)}), \qquad \text{(E.4)} \\
\boldsymbol{\beta}_{t+1}^{(j)} &= \arg\min_{\boldsymbol{\beta}^{(j)}} \frac{1}{2n}|\mathbf{y}^{(j)} - \mathbf{X}^{(j)}\boldsymbol{\beta}^{(j)}|_2^2 + \lambda_N|\boldsymbol{\beta}^{(j)}|_1 + \sum_{k\in\mathcal{N}(j)} \Big(\langle \mathbf{u}_{t+1}^{(jk)} + \mathbf{v}_{t+1}^{(jk)}, \boldsymbol{\beta}^{(j)}\rangle \\
&\quad +\tau|\boldsymbol{\beta}^{(j)} - (\boldsymbol{\beta}_t^{(j)} + \boldsymbol{\beta}_t^{(k)})/2|_2^2\Big). \qquad \text{(E.5)}
\end{aligned}$$

We can simplify the update (E.4) by defining $\mathbf{p}_t^{(j)} \stackrel{\text{def}}{=} \sum_{k\in\mathcal{N}(j)}(\mathbf{u}_t^{(jk)} + \mathbf{v}_t^{(kj)})$ as

$$\mathbf{p}_{t+1}^{(j)} = \mathbf{p}_t^{(j)} + \tau \sum_{k\in\mathcal{N}(j)} (\boldsymbol{\beta}_t^{(j)} - \boldsymbol{\beta}_t^{(k)}), \qquad \text{(E.6)}$$

with $\mathbf{p}_0^{(j)} = 0$, and the update (E.5) as

$$\begin{aligned}
\boldsymbol{\beta}_{t+1}^{(j)} &= \arg\min_{\boldsymbol{\beta}^{(j)}} \frac{1}{2n}|\mathbf{y}^{(j)} - \mathbf{X}^{(j)}\boldsymbol{\beta}^{(j)}|_2^2 + \lambda_N|\boldsymbol{\beta}^{(j)}|_1 + \langle \mathbf{p}_{t+1}^{(j)}, \boldsymbol{\beta}^{(j)}\rangle \\
&\quad +\tau \sum_{k\in\mathcal{N}(j)} |\boldsymbol{\beta}^{(j)} - (\boldsymbol{\beta}_t^{(j)} + \boldsymbol{\beta}_t^{(k)})/2|_2^2. \qquad \text{(E.7)}
\end{aligned}$$

Then, we can solve the augmented Lagrangian by recursively performing (E.6) and (E.7).

To further reduce the computational cost for updating $\boldsymbol{\beta}_{t+1}^{(j)}$, in spirit of the Majorize-Minimization algorithm (Sun et al., 2017), we upper bound the Hessian matrix in $\frac{1}{2mn}|\mathbf{y}^{(j)} - \mathbf{X}^{(j)}\boldsymbol{\beta}^{(j)}|_2^2$ by $\rho_j \mathbf{I}$, and then (E.7) simplifies to

$$
\begin{aligned}
\boldsymbol{\beta}_{t+1}^{(j)} &= \arg\min_{\boldsymbol{\beta}^{(j)}} \frac{1}{2n}|\mathbf{y}^{(j)} - \mathbf{X}^{(j)}\boldsymbol{\beta}^{(j)}|_2^2 + \lambda_N|\boldsymbol{\beta}^{(j)}|_1 + \langle \mathbf{p}_{t+1}^{(j)}, \boldsymbol{\beta}^{(j)} \rangle \\
&\quad + \tau \sum_{k \in \mathcal{N}(j)} |\boldsymbol{\beta}^{(j)} - (\boldsymbol{\beta}_{t+1}^{(j)} + \boldsymbol{\beta}_{t+1}^{(k)})/2|_2^2 \\
&\approx \arg\min_{\boldsymbol{\beta}^{(j)}} \frac{1}{2\omega_j} \Bigg| \boldsymbol{\beta}^{(j)} - \omega_j \Big\{ \rho_j \boldsymbol{\beta}_t^{(j)} - \frac{1}{n}\mathbf{X}^{(j)\mathrm{T}}(\mathbf{X}^{(j)}\boldsymbol{\beta}_t^{(j)} - \mathbf{y}^{(j)}) - \mathbf{p}_{t+1}^{(j)} \\
&\quad + \tau \sum_{k \in \mathcal{N}(j)} (\boldsymbol{\beta}_t^{(j)} + \boldsymbol{\beta}_t^{(k)}) \Big\} \Bigg|_2^2 + \lambda_N |\boldsymbol{\beta}^{(j)}|_1 \\
&= \mathcal{S}_{\lambda_N \omega_j} \Bigg[\omega_j \Big\{ \rho_j \boldsymbol{\beta}_t^{(j)} - \frac{1}{n}\mathbf{X}^{(j)\mathrm{T}}(\mathbf{X}^{(j)}\boldsymbol{\beta}_t^{(j)} - \mathbf{y}^{(j)}) - \mathbf{p}_{t+1}^{(j)} \\
&\qquad\qquad + \tau \sum_{k \in \mathcal{N}(j)} (\boldsymbol{\beta}_t^{(j)} + \boldsymbol{\beta}_t^{(k)}) \Big\} \Bigg],
\end{aligned}
\tag{E.8}
$$

where $\omega_j = 1/(2\tau|\mathcal{N}(j)| + \rho_j)$, $\mathcal{S}_c(\mathbf{x}) \stackrel{\text{def}}{=} (\mathbf{x} - c\mathbf{1})_+ - (-\mathbf{x} - c\mathbf{1})_+$ is the coordinate-wise soft-thresholding operator, where $(c)_+ \stackrel{\text{def}}{=} \max(t, 0)$.

F. Proof of Lemma 1

Proof of Lemma 1: Despite its nondifferentiability, the sign function in model (1.1) is monotonically increasing, indicating $\mathbf{x}_j^{\mathrm{T}}\boldsymbol{\beta}^{\dagger}$ and Y_j are correlated. To verify this, we invoke the independence assumption between ξ_j and $(\mathbf{x}_j, \varepsilon_j)$. Let F_j denote the cumulative distribution function of ε_j in model (1.1). Using the law of total expectation, $E\left\{\mathbf{x}_j \mathrm{sign}(\mathbf{x}_j^{\mathrm{T}}\boldsymbol{\beta}^{\dagger} + \varepsilon_j)\right\} = -2E\left\{\mathbf{x}_j F_j(-\mathbf{x}_j^{\mathrm{T}}\boldsymbol{\beta}^{\dagger})\right\}$.

With Assumption (A1), using the law of total expectation again, we get:

$$E\left\{\mathbf{x}_j F_j(-\mathbf{x}_j^{\mathrm{T}}\boldsymbol{\beta}^{\dagger})\right\} = \boldsymbol{\Sigma}\boldsymbol{\beta}^{\dagger}\mathrm{cov}\left\{(\mathbf{x}_j^{\mathrm{T}}\boldsymbol{\beta}^{\dagger}), F_j(-\mathbf{x}_j^{\mathrm{T}}\boldsymbol{\beta}^{\dagger})\right\}.$$

Summarizing the above results, we have:

$$\boldsymbol{\Sigma}^{-1}\mathrm{cov}(\mathbf{x}_j, Y_j) \;=\; \boldsymbol{\beta}^{\dagger}\Big[(2-4q_j)\mathrm{cov}\left\{(\mathbf{x}_j^{\mathrm{T}}\boldsymbol{\beta}^{\dagger}), F_j(-\mathbf{x}_j^{\mathrm{T}}\boldsymbol{\beta}^{\dagger})\right\}\Big]. \tag{F.1}$$

In a decentralized system comprising m nodes, which may be heterogeneous, to seek the direction of $\boldsymbol{\beta}^{\dagger}$, we suggest minimizing the following

objective function:

$$\begin{aligned}\boldsymbol{\beta}^* &\stackrel{\text{def}}{=} (m\boldsymbol{\Sigma})^{-1}\sum_{j=1}^{m}\text{cov}(\mathbf{x}_j, Y_j) \\ &= \arg\min_{\boldsymbol{\beta}}\sum_{j=1}^{m}E(Y_j - \mathbf{x}_j^{\mathrm{T}}\boldsymbol{\beta})^2. \qquad \text{(F.2)}\end{aligned}$$

Assuming that the linearity Assumption (A1) holds for all $j = 1, \ldots, m$, an immediate consequence of (F.1) is that

$$\begin{aligned}\boldsymbol{\beta}^* &= \boldsymbol{\beta}^\dagger\left[m^{-1}\sum_{j=1}^{m}(2-4q_j)\text{cov}\left\{(\mathbf{x}_j^{\mathrm{T}}\boldsymbol{\beta}^\dagger), F_j(-\mathbf{x}_j^{\mathrm{T}}\boldsymbol{\beta}^\dagger)\right\}\right] \\ &\stackrel{\text{def}}{=} \boldsymbol{\beta}^\dagger\sum_{j=1}^{m}c_j/m.\end{aligned}$$

In particular, if ε_j follows $\mathcal{N}(0, \sigma_j^2)$ and $\mathbf{x}_{i,j}$ follows $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, then:

$$\boldsymbol{\beta}^* = \boldsymbol{\beta}^\dagger\left[m^{-1}\sum_{j=1}^{m}(2q_j-1)\left\{\pi(1+\sigma_j^2)/2\right\}^{-1/2}\right]. \qquad \text{(F.3)}$$

Next we show that when $\mathbf{x}_j$ and ε_j are normal, c_j can be further simplified. Let $u_j \stackrel{\text{def}}{=} \mathbf{x}_j^{\mathrm{T}}\boldsymbol{\beta}^\dagger$. Recall that we assume $\epsilon_j \sim \mathcal{N}(0, \sigma_j^2)$, and $\mathbf{x}_j \sim \mathcal{N}(0, \boldsymbol{\Sigma})$. Note that $\text{cov}\{u_j F_j(-u_j)\} = E\{u_j F_j(-u_j)\}$ by the fact that $E(\mathbf{x}_j) = 0$. Invoking the Stein's Lemma, $E\{u_j F_j(-u_j)\} = -E\{f_j(-u_j)\} = -1/2\{\pi(\sigma_j^2+1)/2\}^{-1/2}$ where $f_j(u)$ is the standard normal density function.

Then the proportionality constant c_j reduces to $(2q_j-1)\{\pi(\sigma_j^2+1)/2\}^{-1/2}$, which completes our proof.

□

G. Proof of Proposition 1

Proof of Proposition 1: We mainly exploit the results in Chang et al. (2014) to prove Proposition 1. We first introduce some notations. The generalized ADMM algorithm of decomposer targets solving the following reformulated network-consensus ℓ_1-penalized least squares problem,

$$\begin{aligned} \min \quad & \sum_{j=1}^{m} f^{(j)}(n^{-1/2}\mathbf{X}^{(j)}\boldsymbol{\beta}^{(j)}) + g_{\lambda_N}(\boldsymbol{\beta}^{(j)}) \\ & \text{s.t. } \boldsymbol{\beta}^{(j)} = \boldsymbol{\beta}^{(k)} = \mathbf{t}^{(jk)},\ \forall k \in \mathcal{N}(j),\ j = 1, \ldots, m, \end{aligned} \tag{G.1}$$

where $f^{(j)}(n^{-1/2}\mathbf{X}^{(j)}\boldsymbol{\beta}^{(j)}) \overset{\text{def}}{=} (2n)^{-1}|\mathbf{y}^{(j)} - \mathbf{X}^{(j)}\boldsymbol{\beta}^{(j)}|_2^2 = 1/2|n^{-1/2}\mathbf{y}^{(j)} - n^{-1/2}\mathbf{X}^{(j)}\boldsymbol{\beta}^{(j)}|_2^2$ and $g_{\lambda_N}(\boldsymbol{\beta}^{(j)}) = \lambda_N|\boldsymbol{\beta}^{(j)}|_1$. It follows that $f^{(j)}(\cdot)$ is 1-strongly convex and 1-Lipschitz smooth, and $g_{\lambda_N}(\cdot)$ is convex.

Denote by $\mathbf{B} \overset{\text{def}}{=} (\widehat{\boldsymbol{\beta}}^{(1)^{\mathrm{T}}}, \ldots, \widehat{\boldsymbol{\beta}}^{(m)^{\mathrm{T}}})^{\mathrm{T}}$ and $\mathbf{u} = \{(\mathbf{u}^{(1k)^{\mathrm{T}}}, k \in \mathcal{N}(1)), \ldots, (\mathbf{u}^{(mk)^{\mathrm{T}}}, k \in \mathcal{N}(m))\}^{\mathrm{T}}$ the optimal primal-dual pair to problem (G.1). And let $\mathbf{B}_t \overset{\text{def}}{=} (\widehat{\boldsymbol{\beta}}_t^{(1)^{\mathrm{T}}}, \ldots, \widehat{\boldsymbol{\beta}}_t^{(m)^{\mathrm{T}}})^{\mathrm{T}}$ and $\mathbf{u}_t = \{(\mathbf{v}_t^{(1k)^{\mathrm{T}}}, k \in \mathcal{N}(1)), \ldots, (\mathbf{v}_t^{(mk)^{\mathrm{T}}}, k \in \mathcal{N}(m))\}^{\mathrm{T}}$ be iterators at tth loop. Let $\widetilde{\mathbf{X}} \overset{\text{def}}{=} \text{diag}(n^{-1/2}\mathbf{X}^{(j)})$. By setting $L_{f_i} = \sigma_{f_i} =$

$\rho = 1$ in Theorem 1(b) in Chang et al. (2014), if the following conditions are satisfied for some $\delta > 0$,

$$(1-\alpha)/2\widetilde{\mathbf{\Sigma}} \succeq \delta(1/2\mathbf{G} + \alpha/2\widetilde{\mathbf{\Sigma}}), \tag{G.2a}$$

$$1/2\mathbf{G} - 1/2\widetilde{\mathbf{\Sigma}} \succeq \delta/\{\tau(1-1/\mu)\sigma^2_{\min}(\mathbf{\Upsilon})\}\mathbf{G}^{\mathrm{T}}\mathbf{G}, \tag{G.2b}$$

$$\alpha/2 \succeq \frac{\delta\{(\mu-1)\lambda_{\max}(\widetilde{\mathbf{\Sigma}})\}}{\{\tau(1-1/\mu)\sigma^2_{\min}(\mathbf{\Upsilon})\}}, \tag{G.2c}$$

then we have

$$\begin{aligned} & \|\mathbf{B}_{t+1} - \mathbf{B}\|^2_{1/2\mathbf{G}+\alpha/2\widetilde{\mathbf{\Sigma}}} + 1/\tau|\mathbf{u}_{t+1} - \mathbf{u}|_2^2 \\ \leq \;& 1/(1+\delta)^{t+1}(\|\mathbf{B}_0 - \mathbf{B}\|^2_{1/2\mathbf{G}+\alpha/2\widetilde{\mathbf{\Sigma}}} + 1/\tau|\mathbf{u}_0 - \mathbf{u}|_2^2), \end{aligned}$$

where $\alpha \in (0,1)$ and $\mu \in (1,\infty)$ are two arbitrary constants, $\|\mathbf{u}\|^2_{\mathbf{A}} \stackrel{\text{def}}{=} \mathbf{u}^{\mathrm{T}}\mathbf{A}\mathbf{u}$ for any vector $\mathbf{u}$ and symmetric matrix $\mathbf{A}$, $\widetilde{\mathbf{\Sigma}} \stackrel{\text{def}}{=} \widetilde{\mathbf{X}}^{\mathrm{T}}\widetilde{\mathbf{X}}$, $\mathbf{G} \stackrel{\text{def}}{=} \mathrm{diag}(\rho_j) \otimes \mathbf{I}_p + \tau(\mathbf{D}+\mathbf{W}) \otimes \mathbf{I}_p$, and $\sigma_{\min}(\mathbf{\Upsilon}) > 0$ is the minimum nonzero singular value of the linear mapping matrix $\mathbf{\Upsilon}$ defined in (A.31) in Chang et al. (2014). By setting $\mu = 2$ and $\alpha = 1/2$, the conditions in (G.2) are fulfilled if we have

$$\begin{aligned} \delta \;\stackrel{\text{def}}{=}\; & \min\Big[\tau\sigma^2_{\min}(\mathbf{\Upsilon})/4\lambda_{\min}\{(\mathbf{G}^{\mathrm{T}}\mathbf{G})^{-1}(\mathbf{G} - \widetilde{\mathbf{\Sigma}})\}, \\ & 1/2\lambda_{\min}\{(\mathbf{G} + 1/2\widetilde{\mathbf{\Sigma}})^{-1}\widetilde{\mathbf{\Sigma}}\}, \tau\sigma^2_{\min}(\mathbf{\Upsilon})/\{4\lambda_{\max}(\widetilde{\mathbf{\Sigma}})\}\Big] > 0. \end{aligned} \tag{G.3}$$

To ensure (G.3), we only need $\rho_j + \tau\lambda_{\min}(\mathbf{D}+\mathbf{W}) > \lambda_{\max}(\widehat{\mathbf{\Sigma}}_j)$ for all $j = 1, \ldots, m$ so that $\mathbf{G} \succ \widetilde{\mathbf{\Sigma}}$, where $\widehat{\mathbf{\Sigma}}_j = n^{-1}\mathbf{X}^{(j)\mathrm{T}}\mathbf{X}^{(j)}$ and $\widetilde{\mathbf{\Sigma}} = \mathrm{diag}(\widehat{\mathbf{\Sigma}}_j)$. Because of $\mathbf{D}_{jj} = \sum_k |\mathbf{W}_{jk}|$, we have $\lambda_{\min}(\mathbf{D}+\mathbf{W}) \geq 0$ by Gershgorin circle theorem. Thus, with $\rho_j > \lambda_{\max}(\widehat{\mathbf{\Sigma}}_j)$ for all $j = 1, \ldots, m$, we have $\delta > 0$ that

$$\|\mathbf{B}_{t+1} - \mathbf{B}\|_F^2 \leq \frac{\|\mathbf{B}_0 - \mathbf{B}\|^2_{1/2\mathbf{G}+1/4\widetilde{\mathbf{\Sigma}}} + 1/\tau|\mathbf{u}_0 - \mathbf{u}|_2^2}{(1+\delta)^{t+1}\lambda_{\min}(1/2\mathbf{G} + 1/4\widetilde{\mathbf{\Sigma}})}.$$

□

H. Some Useful Lemmas

We first provide some preliminaries which will be used in our main proof. Here, we assume that $\widehat{\mathbf{\Sigma}}_{\mathcal{S}_0\times\mathcal{S}_0}$ is invertible, which holds with probability tending to one under the conditions of Theorem 1. For simplicity of proof, we assume $\mathbf{x}_{i,j}$'s and $Y_{i,j}$'s are centered separately in the sample level hereafter.

Lemma H.1. Under Assumptions (A1)–(A5), if $N \geq 2C_2^2(s + \log p)$, we assert

$$\|\widehat{\mathbf{\Sigma}}_{\mathcal{S}_0\times\mathcal{S}_0} - \mathbf{\Sigma}_{\mathcal{S}_0\times\mathcal{S}_0}\|_{\mathrm{op}} \leq 2^{1/2}C_2c_0\{(s+\log p)/N\}^{1/2}, \tag{H.1}$$

and

$$\|\widehat{\mathbf{\Sigma}}^{-1}_{\mathcal{S}_0\times\mathcal{S}_0} - \mathbf{\Sigma}^{-1}_{\mathcal{S}_0\times\mathcal{S}_0}\|_{\mathrm{op}} \leq 2^{1/2}C_2c_0^3\{(s+\log p)/N\}^{1/2}, \tag{H.2}$$

with probability at least $1-2p^{-C_1C_2^2}$, where C_1, C_2 are generic positive constants depending on the sub-Gaussian norm of $\mathbf{x}_{i,1}$.

Lemma H.2. Under Assumptions (A1)–(A5), we can show that

$$\left|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\right|_\infty \leq 2C_1^{-1/2}(\log p/N)^{1/2}, \tag{H.3}$$

and

$$\left|\widehat{\mathbf{\Sigma}}\boldsymbol{\beta}^\dagger - \mathbf{\Sigma}\boldsymbol{\beta}^\dagger\right|_\infty \leq 2C_1^{-1/2}(\log p/N)^{1/2}, \tag{H.4}$$

hold with probability at least $1-2/p^2$, and

$$\left|\sum_{j=1}^m\sum_{i=1}^n(E\mathbf{x}_{i,j}Y_{i,j} - \mathbf{x}_{i,j}Y_{i,j})/N\right|_\infty \leq 2C_1^{-1/2}(\log p/N)^{1/2}, \tag{H.5}$$

holds with probability at least $1-2/p^3$.

I. Proof of Theorem 1

We first present four lemmas to prove our main theorems. The first lemma is a determinant result. Condition (I.2) is called compatibility condition providing the ℓ_2- and ℓ_1-consistency of the resulting estimator.

Lemma I.3. Let $\mathcal{C} \stackrel{\text{def}}{=} \{\delta \in \mathbb{R}^p : |\delta|_1 \leq 4s^{1/2}|\delta|_2\}$ and $\mathbf{z}_N = \sum_{i,j} \mathbf{x}_{i,j} Y_{i,j}/N$. Assume that the following conditions

$$|\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}^* - \mathbf{z}_N|_\infty \leq \lambda_N/2, \tag{I.1}$$

and

$$\delta^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}\delta \geq \gamma|\delta|_2^2, \quad \forall \delta \in \mathcal{C}, \tag{I.2}$$

hold for some constant $\gamma > 0$. Then we have

$$|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|_2 \leq 6s^{1/2}\gamma^{-1}\lambda_N, \tag{I.3}$$

and

$$|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|_1 \leq 24s\gamma^{-1}\lambda_N. \tag{I.4}$$

To prove Theorem 1, it remains for us to verify that the conditions (I.2) and (I.3) hold with high probability, which is completed by the following two lemmas.

Lemma I.4. In addition to Assumptions (A1)–(A5), we set

$$\lambda_N = 4C_0(1+|c|)C_1^{-1/2}(\log p/N)^{1/2}.$$

Then with probability at least $1-2/p^3-2/p^2$, one has $|\widehat{\mathbf{\Sigma}}\boldsymbol{\beta}^*-\mathbf{z}_N|_\infty \leq \lambda_N/2$.

At this time, the condition (I.1) is established with high probability under the condition of Theorem 1. It remains to verify the rest one condition (I.2), which is completed by the following lemma.

Lemma I.5. Under Assumptions (A1)–(A5), with probability at least $1-4/p^2$, we have $\widehat{\mathbf{\Sigma}}$ satisfies condition (I.2), e.g.

$$\delta^{\mathrm{T}}\widehat{\mathbf{\Sigma}}\delta \geq \gamma|\delta|_2^2, \quad \forall \delta \in \mathcal{C},$$

where $\gamma = 1/68\lambda_{\min}(\Sigma)\{4\kappa(\Sigma)+1\}^{-2}$.

Given two sequences $\{a_n\}$, $\{b_n\}$, we write $a_n = \Omega(b_n)$ if there exists a positive constant $c>0$ such that $a_n \geq cb_n$ as n goes to infinity.

Proof of Theorem 1: The first two equations of Theorem 1 is an

immediate result of Lemmas I.4, I.5 and I.3 and Proposition 1. We now turn to bound the angle between $\boldsymbol{\beta}_{T+1}^{(j)}$ and $\boldsymbol{\beta}^{\dagger}$. Note that $\boldsymbol{\beta}^{*}$ is proportional to $\boldsymbol{\beta}^{\dagger}$. It suffices to bound the angle between $\boldsymbol{\beta}_{T+1}^{(j)}$ and $\boldsymbol{\beta}^{*}$. By the identifiability condition $\boldsymbol{\beta}^{\dagger\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\beta}^{\dagger} = 1$ and $c_0^{-1} \leq \lambda_{\min}(\boldsymbol{\Sigma}) \leq \lambda_{\max}(\boldsymbol{\Sigma}) \leq c_0$, we have $|\boldsymbol{\beta}^{\dagger}|_2 = \Omega(1)$ and hence $|\boldsymbol{\beta}^{*}|_2 = \Omega(1)$. By the triangle inequality, we have $|\boldsymbol{\beta}_{T+1}^{(j)}|_2 \geq |\boldsymbol{\beta}^{*}|_2 - |\boldsymbol{\beta}_{T+1}^{(j)} - \boldsymbol{\beta}^{*}|_2$. This, in together with $|\boldsymbol{\beta}_{T+1}^{(j)} - \boldsymbol{\beta}^{*}|_2 = o_p(1)$, gives that $|\boldsymbol{\beta}_{T+1}^{(j)}|_2 = \Omega(1)$ with probability tending to one. By some algebraic calculations, we have

$$
\begin{aligned}
|\boldsymbol{\beta}_{T+1}^{(j)} - \boldsymbol{\beta}^{*}|_2^2/(|\boldsymbol{\beta}_{T+1}^{(j)}|_2|\boldsymbol{\beta}^{*}|_2) &= |\boldsymbol{\beta}_{T+1}^{(j)}|_2/|\boldsymbol{\beta}^{*}|_2 + |\boldsymbol{\beta}^{*}|_2/|\boldsymbol{\beta}_{T+1}^{(j)}|_2 \\
&\quad -2\boldsymbol{\beta}^{*\mathrm{T}}\boldsymbol{\beta}_{T+1}^{(j)}/(|\boldsymbol{\beta}_{T+1}^{(j)}|_2|\boldsymbol{\beta}^{*}|_2) \\
&\geq 2 - 2\boldsymbol{\beta}^{*\mathrm{T}}\boldsymbol{\beta}_{T+1}^{(j)}/(|\boldsymbol{\beta}_{T+1}^{(j)}|_2|\boldsymbol{\beta}^{*}|_2),
\end{aligned}
$$

where the second inequality is due to $a^2 + b^2 \geq 2ab$. Summarizing all above, we have $|\cos\theta_{T+1}^{(j)}| \geq 1 - O_p(|\boldsymbol{\beta}_{T+1}^{(j)} - \boldsymbol{\beta}^{*}|_2^2)$. Noting that $\theta_{T+1}^{(j)} \in [0, \pi/2]$, we have

$$
|\cos\theta_{T+1}^{(j)}| \geq 1 - O_p(|\boldsymbol{\beta}_{T+1}^{(j)} - \boldsymbol{\beta}^{*}|_2^2).
$$

We completed the proof of Theorem 1.

□

J. Proof of Theorem 2

We first summarize the primal-dual witness (PDW) construction borrowed from Wainwright (2009) that we use to prove these theorems. By Karush-Kuhn-Tucker's Theorem, it follows immediately that a vector $\widehat{\boldsymbol{\beta}}$ is optimal for the problem (2.2) if and only if there exists subgradient $\hat{\mathbf{z}} \in \partial|\widehat{\boldsymbol{\beta}}|_1$ such that $\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\beta}} - \mathbf{z}_N + \lambda_N \hat{\mathbf{z}} = 0$. In what follows, we will assume that the matrix $\widehat{\boldsymbol{\Sigma}}_{\mathcal{S}_0 \times \mathcal{S}_0}$ is invertible (which is with probability 1), even though this is not required by the PDW. The PDW method constructs a pair $(\breve{\boldsymbol{\beta}}, \check{\mathbf{z}}) \in \mathbb{R}^p \times \mathbb{R}^p$ by following steps:

- Define $\breve{\boldsymbol{\beta}}$ to be the solution of the following semidefinite programming:

$$\breve{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p, \boldsymbol{\beta}_{\mathcal{S}_0^c} = 0}{\arg\min} \frac{1}{2} \boldsymbol{\beta}^{\mathrm{T}} \widehat{\boldsymbol{\Sigma}} \boldsymbol{\beta} - \boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}_N + \lambda_N |\boldsymbol{\beta}|_1.$$

  This solution is unique under the invertibility of $\widehat{\boldsymbol{\Sigma}}_{\mathcal{S}_0 \times \mathcal{S}_0}$ since, in this case, the function is strictly convex.

- Choose a $\check{\mathbf{z}}_{\mathcal{S}_0} \in \partial|\breve{\boldsymbol{\beta}}_{\mathcal{S}_0}|_1$ such that

$$\widehat{\boldsymbol{\Sigma}}_{\mathcal{S}_0 \times \mathcal{S}_0} \breve{\boldsymbol{\beta}}_{\mathcal{S}_0} - (\mathbf{z}_N)_{\mathcal{S}_0} + \lambda_N \check{\mathbf{z}}_{\mathcal{S}_0} = 0. \tag{J.1}$$

- Set

$$\check{\mathbf{z}}_{\mathcal{S}_0^c} = -\lambda_N^{-1}\left[\left(\widehat{\boldsymbol{\Sigma}}\check{\boldsymbol{\beta}}\right)_{\mathcal{S}_0^c} - (\mathbf{z}_N)_{\mathcal{S}_0^c}\right]. \tag{J.2}$$

  For $j \in \mathcal{S}_0^c$, let $Z_j \stackrel{\text{def}}{=} (\check{\mathbf{z}}_{\mathcal{S}_0^c})_j$. Checking that $|Z_j| < 1$ for all $j \in \mathcal{S}_0^c$ ensures that there is a unique solution $\check{\boldsymbol{\beta}}$ satisfying $\mathcal{S}(\check{\boldsymbol{\beta}}) \in \mathcal{S}_0$.

The proof of Theorem 2 (i) boils down to the strict duality of Z_j for all $j \in \mathcal{S}_0^c$, which is completed by the following lemma.

Lemma J.6. Under the conditions of Theorem 2, we have, with probability at least $1 - 2/p^{C_1C_2^2} - 2/p^2 - 2/p^3$, $|Z_j| \leq v$ uniformly for $j \in \mathcal{S}_0^c$, for some $0 < v < 1$.

Proof of Theorem 2: By lemma J.6, uniformly for $j \in \mathcal{S}_0^c$ and $v < 1$, $|Z_j| \leq v < 1$. with probability tending to one. By the PDW construction, we have $\widehat{\boldsymbol{\beta}} = \check{\boldsymbol{\beta}}$ with probability tending to one. Thus,

$$\mathbb{P}\left\{\mathcal{S}(\widehat{\boldsymbol{\beta}}) \subseteq \mathcal{S}_0\right\} \to 1.$$

For theorem 2 (ii), note that $\mathbb{P}(\widehat{\boldsymbol{\beta}} = \check{\boldsymbol{\beta}}) \to 1$. So we can alternately verify

the results for $\check{\boldsymbol{\beta}}$. For brevity, let

$$G_1 \stackrel{\text{def}}{=} -\lambda_N \check{\mathbf{z}}_{\mathcal{S}_0}, \quad G_2 \stackrel{\text{def}}{=} -\left(\widehat{\boldsymbol{\Sigma}}_{\mathcal{S}_0\times\mathcal{S}_0} - \boldsymbol{\Sigma}_{\mathcal{S}_0\times\mathcal{S}_0}\right)\left(\check{\boldsymbol{\beta}}_{\mathcal{S}_0} - \boldsymbol{\beta}^*_{\mathcal{S}_0}\right)$$
$$G_4 \stackrel{\text{def}}{=} \left(\mathbf{z}_N - \widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}^*\right)_{\mathcal{S}_0}.$$

We rewrite (J.1) as $\check{\boldsymbol{\beta}}_{\mathcal{S}_0} - \boldsymbol{\beta}^*{}_{\mathcal{S}_0} = \boldsymbol{\Sigma}^{-1}_{\mathcal{S}_0\times\mathcal{S}_0}(G_1 + G_2 + G_4)$. By the definition of matrix norm and triangular inequality, we have $|\check{\boldsymbol{\beta}}_{\mathcal{S}_0} - \boldsymbol{\beta}^*{}_{\mathcal{S}_0}|_\infty \leq \|\boldsymbol{\Sigma}^{-1}_{\mathcal{S}_0\times\mathcal{S}_0}\|_\infty(|G_1|_\infty + |G_2|_\infty + |G_4|_\infty)$. It suffice to study the convergence rate for $|G_i|_\infty, i = 1, \ldots, 3$. It follows immediately that $|G_1|_\infty \leq \lambda_N$.

By the fact that $|\cdot|_\infty \leq |\cdot|_2$, we have $|G_2|_\infty \leq |G_2|_2$. Recalling the definition of the operator norm, we have $|G_2|_2 \leq \|\widehat{\boldsymbol{\Sigma}}_{\mathcal{S}_0\times\mathcal{S}_0} - \boldsymbol{\Sigma}_{\mathcal{S}_0\times\mathcal{S}_0}\|_{\text{op}}|\check{\boldsymbol{\beta}}_{\mathcal{S}_0} - \boldsymbol{\beta}^*{}_{\mathcal{S}_0}|_2$. By invoking Lemma H.1, we have $|G_2|_2 \leq 2^{1/2}C_2c_0\{s(s+\log p)/n\}^{1/2}$ with probability tending 1. This, together with the assumption $s^2\log p/N = o(1)$, immediately implies $\|\boldsymbol{\Sigma}^{-1}_{\mathcal{S}_0\times\mathcal{S}_0}\|_\infty|G_2|_\infty \leq 1/2|\check{\boldsymbol{\beta}}_{\mathcal{S}_0} - \boldsymbol{\beta}^*{}_{\mathcal{S}_0}|_\infty$ with probability tending to one.

By Lemma H.2, $|G_4|_\infty \leq 2(1+|c|)C_1^{-1/2}(\log p/N)^{1/2}$. With the choice of λ_N, we have that

$$\left|\check{\boldsymbol{\beta}}_{\mathcal{S}_0} - \boldsymbol{\beta}^*_{\mathcal{S}_0}\right|_\infty \leq C\|\boldsymbol{\Sigma}^{-1}_{\mathcal{S}_0\times\mathcal{S}_0}\|_\infty\left\{(\log p/N)^{1/2}\right\},$$

for some sufficiently large constant $C > 0$. Then Theorem 2 (ii) follows from the above arguments and together with the lower bound condition on $\boldsymbol{\beta}^{*\min}$.

□

K. Proof of Technical Lemmas

Proof of Lemma H.1: Recall that $\widehat{\boldsymbol{\Sigma}} = 1/N \sum_{i,j} \mathbf{x}_{i,j}^{\mathrm{T}} \mathbf{x}_{i,j}$. By setting $t = C_2(\log p)^{1/2}$ in Lemma B.2 of Huang et al. (2018) and the assumption $N \geq 2C_2^2(s + \log p)$, we have

$$\|\widehat{\boldsymbol{\Sigma}}_{\mathcal{S}_0 \times \mathcal{S}_0} - \boldsymbol{\Sigma}_{\mathcal{S}_0 \times \mathcal{S}_0}\|_{\mathrm{op}} \leq C_2\{(s/N)^{1/2} + (\log p/N)^{1/2}\}\lambda_{\max}(\boldsymbol{\Sigma}_{\mathcal{S}_0 \times \mathcal{S}_0}), \quad \text{(K.1)}$$

with probability at least $1-2p^{-C_1C_2^2}$. By the basic inequality $(a+b)^2 \leq 2(a^2+b^2)$ for all $a, b \geq 0$, we have $\{(s/N)^{1/2}+(\log p/N)^{1/2}\} \leq 2^{1/2}\{(s+\log p)/N\}^{1/2}$. This, together with (K.1), entails (H.1).

To get (H.2), we make use of the following inequality: for any matrix $\mathbf{A}, \Delta\mathbf{A} \in \mathbb{R}^{s\times s}$, $\|(\mathbf{A} + \Delta\mathbf{A})^{-1} - \mathbf{A}^{-1}\|_{\mathrm{op}} \leq \|\mathbf{A}^{-1}\|_{\mathrm{op}}^2 \|\Delta\mathbf{A}\|_{\mathrm{op}}$. Choosing $\mathbf{A} = \boldsymbol{\Sigma}_{\mathcal{S}_0 \times \mathcal{S}_0}$ and $\Delta\mathbf{A} = \widehat{\boldsymbol{\Sigma}}_{\mathcal{S}_0 \times \mathcal{S}_0} - \boldsymbol{\Sigma}_{\mathcal{S}_0 \times \mathcal{S}_0}$, we obtain $\|\widehat{\boldsymbol{\Sigma}}_{\mathcal{S}_0 \times \mathcal{S}_0}^{-1} - \boldsymbol{\Sigma}_{\mathcal{S}_0 \times \mathcal{S}_0}^{-1}\|_{\mathrm{op}} \leq \|\boldsymbol{\Sigma}_{\mathcal{S}_0 \times \mathcal{S}_0}^{-1}\|_{\mathrm{op}}^2 \|\widehat{\boldsymbol{\Sigma}}_{\mathcal{S}_0 \times \mathcal{S}_0} - \boldsymbol{\Sigma}_{\mathcal{S}_0 \times \mathcal{S}_0}\|_{\mathrm{op}}$. This, together with (H.1), entails (H.2). The proof is now completed. □

Proof of Lemma H.2: Let $G_{\ell,k}^{i,j} \stackrel{\mathrm{def}}{=} (\mathbf{x}_{i,j})_\ell (\mathbf{x}_{i,j})_k - \boldsymbol{\Sigma}_{\ell,k} \in \mathbb{R}^1, i = 1, \ldots, n, j =$

$1, \ldots, m, \ell = 1, \ldots, p, k = 1, \ldots, p$, which is sub-Exponential by Lemma B.1 of Huang et al. (2018). By Lemma B.3 of Huang et al. (2018), we have

$$\mathbb{P}[\{|\sum_{i,j}^{n} G_{\ell,k}^{i}/N| \geq t\}] \leq 2\exp\{-\min(C_1 t^2, C_2 t)N\},$$

where C_1 and C_2 are generic constants independent of (ℓ, k). By a union bound over all (ℓ, k) pairs,

$$\mathbb{P}\left(\left|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\right|_{\infty} \geq t\right) \leq 2p^2\exp\{-\min(C_1 t^2, C_2 t)N\}.$$

By assumption that $N > 4C_1 C_2^{-1}\log p$, (H.3) follows by setting $t = 2C_1^{-1/2}(\log p/N)^{1/2}$.

Following similar arguments for proving (H.3), inequality (H.4) can be established by invoking the sub-exponential inequality and union bound.

Let $G_k^{i,j} \stackrel{\text{def}}{=} Y_{i,j}(\mathbf{x}_{i,j})_k - E\{Y_{i,j}(\mathbf{x}_{i,j})_k\}, i = 1, \ldots, n, j = 1, \ldots, m, k = 1, \ldots, p$, which is sub-Exponential by applying Lemma B.1 of Huang et al. (2018). By Lemma B.3 of Huang et al. (2018), $\mathbb{P}[\{|\sum_{j=1}^{m}\sum_{i=1}^{n} G_k^{i,j}/N| \geq t\}] \leq 2\exp\{-\min(C_1 t^2, C_2 t)N\}$, where C_1 and C_2 are generic constants

independent of k. By a union bound over k, we have

$$\mathbb{P}\left\{\left|\sum_{j=1}^{m}\sum_{i=1}^{n}(E\mathbf{x}_{i,j}Y_{i,j}-\mathbf{x}_{i,j}Y_{i,j})/N\right|_{\infty}\geq t\right\}\leq 2p\exp\{-\min(C_1t^2,C_2t)N\}.$$

(H.5) follows from the assumption $N > 4C_1C_2^{-1}\log p$ by setting $t = 2C_1^{-1/2}(\log p/N)^{1/2}$. We complete our proof. □

Proof of Lemma I.3: We first show that $|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*|_1 \leq 4s^{1/2}|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*|_2$, which implies that $\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*$ belongs to $\mathcal{C}$. The ℓ_2 and ℓ_1-consistency of $\widehat{\boldsymbol{\beta}}$ are then established by invoking (I.1) and (I.2).

For brevity, let $\mathbf{A} \overset{\text{def}}{=} \widehat{\boldsymbol{\Sigma}}$, $\mathbf{b} \overset{\text{def}}{=} \mathbf{z}_N$ and $M \overset{\text{def}}{=} 1/2\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{A}\widehat{\boldsymbol{\beta}} - (\widehat{\boldsymbol{\beta}})^{\mathrm{T}}\mathbf{b} - (1/2\boldsymbol{\beta}^{*\mathrm{T}}\mathbf{A}\boldsymbol{\beta}^* - \boldsymbol{\beta}^{*\mathrm{T}}\mathbf{b})$. By the definition of $\widehat{\boldsymbol{\beta}}$, we have $M \leq \lambda_N(|\boldsymbol{\beta}^*|_1 - |\widehat{\boldsymbol{\beta}}|_1)$. Because ${\boldsymbol{\beta}^*}_{\mathcal{S}_0^c} = 0$, we have $|(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}})_{\mathcal{S}_0}|_1 - |(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}})_{\mathcal{S}_0^c}|_1 \geq |{\boldsymbol{\beta}^*}_{\mathcal{S}_0}|_1 - |\widehat{\boldsymbol{\beta}}|_1$. Combining the last two inequalities, we can further bound M by

$$M \leq \lambda_N|(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}})_{\mathcal{S}_0}|_1 - \lambda_N|(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}})_{\mathcal{S}_0^c}|_1. \tag{K.2}$$

Because $\mathbf{A}$ is non-negative definite, we have, $(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}})^{\mathrm{T}}\mathbf{A}(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}) \geq 0$, which implies $M \geq (\mathbf{A}\boldsymbol{\beta}^* - \mathbf{b})^{\mathrm{T}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$. By invoking Hölder inequality and condition (I.1), we have $(\mathbf{A}\boldsymbol{\beta}^* - \mathbf{b})^{\mathrm{T}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \geq -\lambda_N|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|_1/2$. Combining

the last two inequalities, we have a lower bound

$$M \geq -\lambda_N |\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|_1/2. \tag{K.3}$$

Combining (K.2) and (K.3), we obtain $|(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}})_{\mathcal{S}_0^c}|_1 \leq 3|(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}})_{\mathcal{S}_0}|_1$, which implies $|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|_1 \leq 4|(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{\mathcal{S}_0}|_1 \leq 4s^{1/2}|(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{\mathcal{S}_0}|_2 \leq 4s^{1/2}|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|_2$. We finish our first part of the proof.

Recall that a vector $\mathbf{z}$ is a subgradient of the ℓ_1 norm $|\cdot|_1$ evaluated at a vector $\boldsymbol{\beta} \in \mathbb{R}^p$ (i.e., $\mathbf{z} \in \partial|\boldsymbol{\beta}|_1$) if we have $\mathbf{z}_j = \text{sign}(\boldsymbol{\beta}_j), \boldsymbol{\beta}_j \neq 0$ and $\mathbf{z}_j \in [-1, 1]$ otherwise. By the first order condition, we have $\mathbf{A}\widehat{\boldsymbol{\beta}} - \mathbf{b} \in \lambda_N \partial|\widehat{\boldsymbol{\beta}}|_1$, which implies $|\mathbf{A}\widehat{\boldsymbol{\beta}} - b|_\infty \leq \lambda_N$. This, together with (I.1), yields $|\mathbf{A}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)|_\infty \leq 3/2\lambda_N$. Since $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$ belongs to $\mathcal{C}$, invoking (I.2), we have $\gamma|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|_2^2 \leq (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^{\mathrm{T}}\mathbf{A}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$. Combining the last two inequalities, we have $|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|_2^2 \leq 6\lambda_N s^{1/2}\gamma^{-1}|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|_2$. Recalling $|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|_1 \leq 4s^{1/2}|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|_2$, (I.3) deduces (I.4). □

Proof of Lemma I.4: Recall that $\mathbf{z}_N = 1/N \sum_{i,j} \mathbf{x}_{i,j} Y_{i,j}$ and $\widehat{\boldsymbol{\Sigma}} = 1/N \sum_{i,j} \mathbf{x}_{i,j}\mathbf{x}_{i,j}^{\mathrm{T}}$. Let $Q_1 \stackrel{\text{def}}{=} |\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}^* - \boldsymbol{\Sigma}\boldsymbol{\beta}^*|_\infty$ and $Q_2 \stackrel{\text{def}}{=} |\boldsymbol{\Sigma}\boldsymbol{\beta}^* - \mathbf{z}_N|_\infty$. By the triangular inequality, $|\mathbf{z}_N - \widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}^*|_\infty \leq Q_1 + Q_2$. It follows immediately that

$$Q_2 = |\sum_{j=1}^{m}\sum_{i=1}^{n}(E\mathbf{x}_{i,j}Y_{i,j} - \mathbf{x}_{i,j}Y_{i,j})/N|_\infty,$$

by the definition of $\boldsymbol{\beta}^*$, which is bounded by $2C_1^{-1/2}(\log p/N)^{1/2}$ with probability at least $1-2/p^3$ by invoking Lemma H.2. We turn to Q_1. Recall that $\boldsymbol{\beta}^* = c\boldsymbol{\beta}^\dagger$. By Lemma H.2, we conclude that $Q_1 \leq 2C_1^{-1/2}|c|(\log p/N)^{1/2}$ holds with probability at least $1-2/p^2$. The proof of Lemma I.4 is completed by combining the bounds for Q_1 and Q_2. □

Proof of Lemma I.5: Under Assumptions (A3)–(A5), by a similar arguments of Lemma 5 in Huang et al. (2018), we can establish this result. □

Proof of Lemma J.6: Let

$$\mathbf{V} \stackrel{\text{def}}{=} -\lambda_N^{-1}\left(\mathbf{z}_N - \widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}^*\right)_{\mathcal{S}_0}, \quad M_1 \stackrel{\text{def}}{=} \widehat{\boldsymbol{\Sigma}}_{\mathcal{S}_0^c\times\mathcal{S}_0}\widehat{\boldsymbol{\Sigma}}_{\mathcal{S}_0\times\mathcal{S}_0}^{-1}\mathbf{V},$$
$$M_2 \stackrel{\text{def}}{=} \widehat{\boldsymbol{\Sigma}}_{\mathcal{S}_0^c\times\mathcal{S}_0}\widehat{\boldsymbol{\Sigma}}_{\mathcal{S}_0\times\mathcal{S}_0}^{-1}\check{\mathbf{z}}_{\mathcal{S}_0}, \quad M_3 \stackrel{\text{def}}{=} \lambda_N^{-1}\{\mathbf{z}_N - \boldsymbol{\Sigma}\boldsymbol{\beta}^*\}_{\mathcal{S}_0^c}.$$

Plugging (J.1) into (J.2), immediately leads to

$$\check{\mathbf{z}}_{\mathcal{S}_0^c} \quad = \quad M_1 + M_2 + M_3. \tag{K.4}$$

By the triangular inequality, $|\check{\mathbf{z}}_{\mathcal{S}_0^c}|_\infty \leq |M_1|_\infty + |M_2|_\infty + |M_3|_\infty$. It remains to bound the supremum norm of each M_i for $i = 1, \ldots, 3$.

We first claim that

$$\|\widehat{\mathbf{\Sigma}}_{\mathcal{S}_0^c\times\mathcal{S}_0}\widehat{\mathbf{\Sigma}}_{\mathcal{S}_0\times\mathcal{S}_0}^{-1}\|_\infty = o_p(1) + \|\mathbf{\Sigma}_{\mathcal{S}_0^c\times\mathcal{S}_0}\mathbf{\Sigma}_{\mathcal{S}_0\times\mathcal{S}_0}^{-1}\|_\infty. \tag{K.5}$$

By invoking Lemma I.4 and setting C_0 in λ_N sufficient large, we have $|\mathbf{V}|_\infty \leq \alpha/8$ with high probability. By the definition of supremum norm and $\|\mathbf{\Sigma}_{\mathcal{S}_0^c\times\mathcal{S}_0}\mathbf{\Sigma}_{\mathcal{S}_0\times\mathcal{S}_0}^{-1}\|_\infty \leq 1-\alpha$, we have $|M_1|_\infty \leq \|\widehat{\mathbf{\Sigma}}_{\mathcal{S}_0^c\times\mathcal{S}_0}\widehat{\mathbf{\Sigma}}_{\mathcal{S}_0\times\mathcal{S}_0}^{-1}\|_\infty|\mathbf{V}|_\infty \leq \alpha/8$ with high probability. By $|\check{\mathbf{z}}_{\mathcal{S}_0}|_\infty \leq 1$ and the definition of $\|\cdot\|_\infty$, we have $|M_2|_\infty \leq 1-\alpha/2$ with high probability. By the similar arguments for bounding $\mathbf{V}$, we can get $|M_3|_\infty \leq \alpha/8$ with high probability. Combining the bounds for $|M_i|, i = 1,\ldots,3$, we conclude that uniformly for $j \in \mathcal{S}_0^c$ and some constant $v = 1-\alpha/4 < 1$, $|Z_j| \leq v < 1$ with high probability.

Now, it remains to show (K.5). Let

$$E_1 \stackrel{\text{def}}{=} \left(\widehat{\mathbf{\Sigma}}_{\mathcal{S}_0^c\times\mathcal{S}_0} - \mathbf{\Sigma}_{\mathcal{S}_0^c\times\mathcal{S}_0}\right)\left(\widehat{\mathbf{\Sigma}}_{\mathcal{S}_0\times\mathcal{S}_0}^{-1} - \mathbf{\Sigma}_{\mathcal{S}_0\times\mathcal{S}_0}^{-1}\right),\ E_2 \stackrel{\text{def}}{=} \mathbf{\Sigma}_{\mathcal{S}_0^c\times\mathcal{S}_0}\left(\widehat{\mathbf{\Sigma}}_{\mathcal{S}_0\times\mathcal{S}_0}^{-1} - \mathbf{\Sigma}_{\mathcal{S}_0\times\mathcal{S}_0}^{-1}\right)$$
$$E_3 \stackrel{\text{def}}{=} \left(\widehat{\mathbf{\Sigma}}_{\mathcal{S}_0^c\times\mathcal{S}_0} - \mathbf{\Sigma}_{\mathcal{S}_0^c\times\mathcal{S}_0}\right)\mathbf{\Sigma}_{\mathcal{S}_0\times\mathcal{S}_0}^{-1},\ \text{and}\ E_4 \stackrel{\text{def}}{=} \mathbf{\Sigma}_{\mathcal{S}_0^c\times\mathcal{S}_0}\mathbf{\Sigma}_{\mathcal{S}_0\times\mathcal{S}_0}^{-1}.$$

By the triangular inequality, we have $\|\widehat{\mathbf{\Sigma}}_{\mathcal{S}_0^c\times\mathcal{S}_0}\widehat{\mathbf{\Sigma}}_{\mathcal{S}_0\times\mathcal{S}_0}^{-1}\|_\infty \leq \|E_1\|_\infty + \|E_2\|_\infty + \|E_3\|_\infty + \|E_4\|_\infty$. Bounding $\|\widehat{\mathbf{\Sigma}}_{\mathcal{S}_0^c\times\mathcal{S}_0}\widehat{\mathbf{\Sigma}}_{\mathcal{S}_0\times\mathcal{S}_0}^{-1}\|_\infty$ boils down to bound $\|E_i\|_\infty$ for $i = 1,\ldots,3$.

Bound for $\|E_1\|_\infty$. Invoking Lemma H.1, we have

$$\|\widehat{\mathbf{\Sigma}}_{\mathcal{S}_0\times\mathcal{S}_0} - \mathbf{\Sigma}_{\mathcal{S}_0\times\mathcal{S}_0}\|_{\mathrm{op}} = O_p\left[\{(s+\log p)/N\}^{1/2}\right],$$

and

$$\|\widehat{\mathbf{\Sigma}}^{-1}_{\mathcal{S}_0\times\mathcal{S}_0} - \mathbf{\Sigma}^{-1}_{\mathcal{S}_0\times\mathcal{S}_0}\|_{\mathrm{op}} = O_p\left[\{(s+\log p)/N\}^{1/2}\right].$$

This, together with Lemma H.2 and the fact that $\|A\|_\infty \le n|A|_\infty$ and $\|A\|_\infty \le \sqrt{n}\|A\|_{\mathrm{op}}$ for any matrix $A \in \mathbb{R}^{m\times n}$, gives that

$$\begin{aligned}
\|E_1\|_\infty &\le \|\widehat{\mathbf{\Sigma}}_{\mathcal{S}_0^c\times\mathcal{S}_0} - \mathbf{\Sigma}_{\mathcal{S}_0^c\times\mathcal{S}_0}\|_\infty \|\widehat{\mathbf{\Sigma}}^{-1}_{\mathcal{S}_0\times\mathcal{S}_0} - \mathbf{\Sigma}^{-1}_{\mathcal{S}_0\times\mathcal{S}_0}\|_\infty \\
&\le s^{3/2}|\widehat{\mathbf{\Sigma}}_{\mathcal{S}_0^c\times\mathcal{S}_0} - \mathbf{\Sigma}_{\mathcal{S}_0^c\times\mathcal{S}_0}|_\infty \|\widehat{\mathbf{\Sigma}}^{-1}_{\mathcal{S}_0\times\mathcal{S}_0} - \mathbf{\Sigma}^{-1}_{\mathcal{S}_0\times\mathcal{S}_0}\|_{\mathrm{op}} \\
&= O_p(s^2\log p/N),
\end{aligned}$$

where we use the submultiplicative property of $\|\cdot\|_\infty$ in the first inequality.

Bound for $\|E_2\|_\infty$. For each $j \in \{1,\ldots,(p-s)\}$, let $\mathbf{e}_j$ be the vector in $\mathbb{R}^{p-s}$ whose components are all zero, except j-th term that equals 1. We

have

$$\begin{aligned}
\|E_2\|_\infty &= \max_{j\in\{1,\ldots,(p-s)\}} \sup_{|\mathbf{e}|_\infty=1,\ \mathbf{e}\in\mathbb{R}^s} \left|\mathbf{e}_j^{\mathrm{T}} \boldsymbol{\Sigma}_{\mathcal{S}_0^c\times\mathcal{S}_0} \left(\widehat{\boldsymbol{\Sigma}}_{\mathcal{S}_0\times\mathcal{S}_0}^{-1} - \boldsymbol{\Sigma}_{\mathcal{S}_0\times\mathcal{S}_0}^{-1}\right)\mathbf{e}\right| \\
&\leq c_0 \max_{j\in\{1,\ldots,(p-s)\}} \sup_{|\mathbf{e}|_\infty=1,\ \mathbf{e}\in\mathbb{R}^s} \left|\left(\widehat{\boldsymbol{\Sigma}}_{\mathcal{S}_0\times\mathcal{S}_0}^{-1} - \boldsymbol{\Sigma}_{\mathcal{S}_0\times\mathcal{S}_0}^{-1}\right)\mathbf{e}\right|_2 |\mathbf{e}_j|_2 \\
&\leq c_0 s^{1/2} \left\|\widehat{\boldsymbol{\Sigma}}_{\mathcal{S}_0\times\mathcal{S}_0}^{-1} - \boldsymbol{\Sigma}_{\mathcal{S}_0\times\mathcal{S}_0}^{-1}\right\|_{\mathrm{op}} \\
&= O_p\left[\left\{(s^2 + s\log p)/N\right\}^{1/2}\right],
\end{aligned}$$

where the first line is by the definition of $\|\cdot\|_\infty$, the second line owes to $c_0^{-1} \leq \lambda_{\min}(\boldsymbol{\Sigma})$ by Assumption (A4), the third line is again due to the definition of $\|\cdot\|_\infty$, and the last line is by Lemma H.1.

Bound for $\|E_3\|_\infty$. Similarly, we have

$$\begin{aligned}
\|E_3\|_\infty &= \max_{j\in\{1,\ldots,(p-s)\}} \sup_{|\mathbf{e}|_\infty=1,\ \mathbf{e}\in\mathbb{R}^s} \left|\mathbf{e}_j^{\mathrm{T}} \left(\widehat{\boldsymbol{\Sigma}}_{\mathcal{S}_0^c\times\mathcal{S}_0} - \boldsymbol{\Sigma}_{\mathcal{S}_0^c\times\mathcal{S}_0}\right)\boldsymbol{\Sigma}_{\mathcal{S}_0\times\mathcal{S}_0}^{-1}\mathbf{e}\right| \\
&\leq c_0 \max_{j\in\{1,\ldots,(p-s)\}} \sup_{|\mathbf{e}|_\infty=1,\ \mathbf{e}\in\mathbb{R}^s} \left|\mathbf{e}_j^{\mathrm{T}} \left(\widehat{\boldsymbol{\Sigma}}_{\mathcal{S}_0^c\times\mathcal{S}_0} - \boldsymbol{\Sigma}_{\mathcal{S}_0^c\times\mathcal{S}_0}\right)\right|_2 |\mathbf{e}|_2 \\
&\leq c_0 s^{1/2} \left\|\widehat{\boldsymbol{\Sigma}}_{\mathcal{S}_0^c\times\mathcal{S}_0} - \boldsymbol{\Sigma}_{\mathcal{S}_0^c\times\mathcal{S}_0}\right\|_{\mathrm{op}} \\
&= O_p\left[\left\{(s^2 + s\log p)/N\right\}^{1/2}\right].
\end{aligned}$$

□

References

Aslan, Y. E., I. Korpeoglu, and Ö. Ulusoy (2012). A framework for use of wireless sensor networks in forest fire detection and monitoring. *Computers, Environment and Urban Systems 36*(6), 614–625.

Bai, Z. and J. W. Silverstein (2010). *Spectral Analysis of Large Dimensional Random Matrices* (2nd ed ed.). Springer Series in Statistics. New York ; London: Springer.

Boyd, S. (2010). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning 3*(1), 1–122.

Chang, T.-H., M. Hong, and X. Wang (2014, February). Multi-Agent Distributed Optimization via Inexact Consensus ADMM. *IEEE Transactions on Signal Processing 63*.

Chung, F. R. K. (1997). *Spectral Graph Theory*. Number no. 92 in Regional Conference Series in Mathematics. Providence, R.I: Published for the Conference Board of the mathematical sciences by the American Mathematical Society.

Fiedler, M. (1973). Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal 23*(2), 298–305.

Huang, J., Y. Jiao, X. Lu, and L. Zhu (2018, January). Robust decoding from 1-bit compressive sampling with ordinary and regularized least squares. *SIAM Journal on Scientific Computing 40*(4), A2062–A2086.

Jiménez, A., V. García-Díaz, and S. Bolaños (2018, February). A Decentralized Framework for

Multi-Agent Robotic Systems. *Sensors 18*(2), 417.

Khedo, K. K., R. Perseedoss, A. Mungur, et al. (2010). A wireless sensor network air pollution monitoring system. *arXiv preprint arXiv:1005.1737*.

Ling, Q. and Z. Tian (2010). Decentralized sparse signal recovery for compressive sleeping wireless sensor networks. *IEEE Transactions on Signal Processing 58*(7), 3816–3827.

Liu, W., X. Mao, X. Zhang, and X. Zhang (2024, April). Robust personalized federated learning with sparse penalization. *Journal of the American Statistical Association 120*(549), 266–277.

Lu, X. and C. Amato (2019). Decentralized likelihood quantile networks for improving performance in deep multi-agent reinforcement learning. *arXiv preprint arXiv:1812.06319 v4*.

Ma, S. and J. Huang (2017). A concave pairwise fusion approach to subgroup analysis. *Journal of the American Statistical Association 112*(517), 410–423.

Mateos, G., J. Bazerque, and G. Giannakis (2010). Distributed sparse linear regression. *IEEE Transactions on Signal Processing 58*(10), 5262–5276.

Rasheed, A. A. A., M. N. Abdullah, and A. S. Al-Araji (2022). A review of multi-agent mobile robot systems applications. *International Journal of Electrical and Computer Engineering 12*(4), 3517–3529.

Schäfer, B., M. Matthiae, M. Timme, and D. Witthaut (2015). Decentral smart grid control. *New Journal of Physics 17*(1), 015002.

Sun, Y., P. Babu, and D. P. Palomar (2017, February). Majorization-Minimization Algorithms in Signal Processing, Communications, and Machine Learning. *IEEE Transactions on Signal Processing 65*(3), 794–816.

Talat, R., M. Muzammal, Q. Qu, W. Zhou, M. Najam-ul Islam, S. H. Bamakan, and J. Qiu (2020). A decentralized system for green energy distribution in a smart grid. *Journal of Energy Engineering 146*(1), 04019036.

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory 55*(5), 2183–2202.

Yao, D. (2001, November). A method to standardize a reference of scalp eeg recordings to a point at infinity. *Physiological Measurement 22*(4), 693–711.

Yi, W. Y., K. M. Lo, T. Mak, K. S. Leung, Y. Leung, and M. L. Meng (2015). A survey of wireless sensor network based air pollution monitoring systems. *Sensors 15*(12), 31392–31427.

Zheng, W.-L. and B.-L. Lu (2015). Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks. *IEEE Transactions on Autonomous Mental Development 7*(3), 162–175.