Appendix

A. Proof of Theorem 1 and Theorem 2

The proofs of two theorems are almost identical with a single difference selecting initial parameter on which the soft-thresholding is performed. In the proof, we denote this initial parameter, i.e., $(X^\top X + \epsilon I)^{-1} X^\top y$ or $\big[T_\nu\big(\frac{X^\top X}{n}\big)\big]^{-1} \frac{X^\top y}{n}$ by $\bar{\theta}$.

Let Δ be the error vector, $\widehat{\theta} - \theta^*$. Since we choose λ_n greater than $\mathcal{R}^*(|\theta^* - \bar{\theta})$,

$$\begin{aligned}\mathcal{R}^*(\Delta) &= \mathcal{R}^*(\widehat{\theta} - \bar{\theta} + \bar{\theta} - \theta^*) \\ &\le \mathcal{R}^*(\widehat{\theta} - \bar{\theta}) + \mathcal{R}^*(\theta^* - \bar{\theta}) \le 2\lambda_n \end{aligned} \tag{9}$$

where we utilize the fact that $\widehat{\theta}$ is feasible.

For notational simplicity, we use (S, S^c) instead of an arbitrary subspace pair $(\mathcal{M}, \overline{\mathcal{M}}^\perp)$. Additionally, we use the notion Δ_S to represent the ℓ_2 projection onto the model space $\mathcal{M}$. Then, by the assumption of the statement that $\theta^*_{S^c} = \mathbf{0}$, and the decomposability of $\mathcal{R}(\cdot)$ with respect to (S, S^c),

$$\begin{aligned}\mathcal{R}(\theta^*) &= \mathcal{R}(\theta^*) + \mathcal{R}(\Delta_{S^c}) - \mathcal{R}(\Delta_{S^c}) \\ &= \mathcal{R}(\theta^* + \Delta_{S^c}) - \mathcal{R}(\Delta_{S^c}) \\ &\overset{(i)}{\le} \mathcal{R}(\theta^* + \Delta_{S^c} + \Delta_S) + \mathcal{R}(\Delta_S) - \mathcal{R}(\Delta_{S^c}) \\ &= \mathcal{R}(\theta^* + \Delta) + \mathcal{R}(\Delta_S) - \mathcal{R}(\Delta_{S^c}) \end{aligned} \tag{10}$$

where the equality (i) holds by the triangle inequality, which is the basic property of norms. Since we minimize the objective $\mathcal{R}(\theta)$ in (4) or (6), we obtain the inequality of $\mathcal{R}(\theta^* + \Delta) = \mathcal{R}(\widehat{\theta}) \le \mathcal{R}(\theta^*)$. Combining this inequality with (10), we have

$$0 \le \mathcal{R}(\Delta_S) - \mathcal{R}(\Delta_{S^c}) \tag{11}$$

Armed with inequalities (9) and (11), we utilize the Hölder's inequality and the decomposability of our regularizer $\mathcal{R}(\cdot)$ in order to derive the error bounds in terms of ℓ_2 norm:

$$\begin{aligned}\|\Delta\|_2^2 &= \langle \Delta, \Delta \rangle \le \mathcal{R}^*(\Delta)\mathcal{R}(\Delta) \\ &\le \mathcal{R}^*(\Delta)\big(\mathcal{R}(\Delta_S) + \mathcal{R}(\Delta_{S^c})\big).\end{aligned}$$

Since the error vector Δ satisfies the inequality (11),

$$\|\Delta\|_2^2 \le 2\,\mathcal{R}^*(\Delta)\,\mathcal{R}(\Delta_S).$$

Combining all the pieces together yields

$$\|\Delta\|_2^2 \le \; 4\Psi(S)\lambda_n \|\Delta_S\|_2 \tag{12}$$

where $\Psi(\mathcal{M})$ is the abbreviation for $\Psi(S, \|\cdot\|_2)$.

Notice that the projection operator is non-expansive, $\|\Delta_S\|_2^2 \le \|\Delta\|_2^2$. Hence, we obtain $\|\Delta_S\|_2 \le 4\Psi(S)\lambda_n$, and plugging it back into (12) yields the ℓ_2 error bounds.

Finally, the error bounds in terms of the regularizer itself are straightforward from the following reasoning:

$$\begin{aligned}&\mathcal{R}(\Delta) = \mathcal{R}(\Delta_S) + \mathcal{R}(\Delta_{S^c}) \le 2\mathcal{R}(\Delta_S) \\ &\le 2\Psi(S)\|\Delta_S\|_2 \le 8[\Psi(S)]^2\lambda_n.\end{aligned}$$

B. Useful lemma(s)

Lemma 1 (Lemma 1 of (Ravikumar et al., 2011))**.** *Let $\mathcal{A}$ be the event that*

$$\left\| \frac{X^\top X}{n} - \Sigma \right\|_\infty \le 8(\max_i \Sigma_{ii})\sqrt{\frac{10\tau \log p'}{n}}$$

where $p' := \max\{n, p\}$ and τ is any constant greater than 2. Suppose that the design matrix X is i.i.d. sampled from Σ-Gaussian ensemble *with $n \ge 40\max_i \Sigma_{ii}$. Then, the probability of event $\mathcal{A}$ occurring is at least $1 - 4/p'^{\tau-2}$.*

Lemma 2 (In the proof of Corollary 2 (Negahban et al., 2012))**.** *By the conditions of (C-OLS2), and the sub-Gaussian property of noise w,*

$$P\left(\frac{\|X^\top w\|_\infty}{n} \ge t\right) \le 2\exp(-\frac{nt^2}{2\sigma^2} + \log p)$$

C. Proof of Proposition 1

By Lemma 1, we have the event $\mathcal{A}$:

$$\left\| \frac{X^\top X}{n} - \Sigma \right\|_\infty \le 8(\max_i \Sigma_{ii})\sqrt{\frac{10\tau \log p'}{n}}$$

with high probability specified in the statement of lemma. Conditioned on $\mathcal{A}$, $T_\nu\big(\frac{X^\top X}{n}\big)$ with the specific choice of ν in the statement, has larger diagonal entries and smaller off-diagonal entries than Σ. Therefore, on the $\mathcal{A}$, $T_\nu\big(\frac{X^\top X}{n}\big)$ is diagonally dominant, and hence invertible.

D. Proof of Corollary 2

In order to utilize Theorem 2, we need to derive the upper bound of $\big\|\theta^* - \big[T_\nu\big(\frac{X^\top X}{n}\big)\big]^{-1}\frac{X^\top y}{n}\big\|_\infty$:

$$\begin{aligned}&\|\theta^* - \bar{\theta}\|_\infty \\ &= \left\| \left[T_\nu\left(\frac{X^\top X}{n}\right)\right]^{-1} T_\nu\left(\frac{X^\top X}{n}\right)\theta^* - \left[T_\nu\left(\frac{X^\top X}{n}\right)\right]^{-1}\frac{X^\top y}{n}\right\|_\infty \\ &\le \left\| \left[T_\nu\left(\frac{X^\top X}{n}\right)\right]^{-1} \right\|_\infty \left\| T_\nu\left(\frac{X^\top X}{n}\right)\theta^* - \frac{X^\top y}{n}\right\|_\infty\end{aligned}$$

We first control $\big\|\big[T_\nu\big(\frac{X^\top X}{n}\big)\big]^{-1}\big\|_\infty$ term. We are going to show that $T_\nu\left(\frac{X^\top X}{n}\right)$ is diagonally dominant with high probability hence the term we care about will be bound. By Lemma 1, if $n > 40\max_i \Sigma_{ii}$, the event $\mathcal{A}$ occurs with probability at least $1 - 4/p'^{\tau-2}$ for $p' := \max\{n, p\}$ and any constant $\tau > 2$. Conditioned on $\mathcal{A}$, for all row index i,

$$\left[T_\nu\Big(\frac{X^\top X}{n}\Big)\right]_{ii} - \sum_{j\neq i}\left|\left[T_\nu\Big(\frac{X^\top X}{n}\Big)\right]_{ij}\right|$$
$$\geq \left(\Sigma_{ii} - a\sqrt{\frac{\log p'}{n}} + \nu\right) - \sum_{j\neq i}\left(|\Sigma_{ij}| + a\sqrt{\frac{\log p'}{n}} - \nu\right).$$

where $a := 8(\max_i \Sigma_{ii})\sqrt{10\tau}$.

Therefore, provided $\nu := a\sqrt{\frac{\log p'}{n}}$,

$$\left[T_\nu\Big(\frac{X^\top X}{n}\Big)\right]_{ii} - \sum_{j\neq i}\left|\left[T_\nu\Big(\frac{X^\top X}{n}\Big)\right]_{ij}\right|$$
$$\geq \Sigma_{ii} - \sum_{j\neq i}|\Sigma_{ij}| \geq \delta_i \geq \delta_{\min}.$$

Note that conditioned on $\mathcal{A}$, the matrix $T_\nu\big(\frac{X^\top X}{n}\big)$ is invertible since it is strictly diagonally dominant matrix, and $\left\|\left[T_\nu\big(\frac{X^\top X}{n}\big)\right]^{-1}\right\|_\infty \leq \frac{1}{\delta_{\min}}$ by Varah (1975).

Now consider the second term $\left\|T_\nu\left(\frac{X^\top X}{n}\right)\theta^* - \frac{X^\top y}{n}\right\|_\infty$ in the equality:

$$\left\|T_\nu\Big(\frac{X^\top X}{n}\Big)\theta^* - \frac{X^\top y}{n}\right\|_\infty$$
$$= \left\|T_\nu\Big(\frac{X^\top X}{n}\Big)\theta^* - \frac{X^\top X}{n}\theta^* + \frac{X^\top X}{n}\theta^* - \frac{X^\top y}{n}\right\|_\infty$$
$$\leq \left\|T_\nu\Big(\frac{X^\top X}{n}\Big)\theta^* - \frac{X^\top X}{n}\theta^* - \frac{X^\top w}{n}\right\|_\infty$$
$$\leq \left\|T_\nu\Big(\frac{X^\top X}{n}\Big)\theta^* - \frac{X^\top X}{n}\theta^*\right\|_\infty + \left\|\frac{X^\top w}{n}\right\|_\infty.$$

Since $\frac{\|X^\top w\|_\infty}{n}$ can be upper-bounded by $2\sigma\sqrt{\frac{\log p}{n}}$ as stated in Lemma 2, the only remaining term to control is $\left\|\left(T_\nu\big(\frac{X^\top X}{n}\big) - \frac{X^\top X}{n}\right)\theta^*\right\|_\infty$. Each element of $T_\nu\big(\frac{X^\top X}{n}\big) - \frac{X^\top X}{n}$ is upper-bounded by ν by construction, which is set $a\sqrt{\frac{\log p'}{n}}$. Therefore, for every entry of $\left(T_\nu\big(\frac{X^\top X}{n}\big) - \frac{X^\top X}{n}\right)\theta^*$, we can apply Hölder inequality so that it is bound by $a\sqrt{\frac{\log p}{n}}\|\theta^*\|_1$.

Therefore, if we select λ_n as

$$\frac{1}{\delta_{\min}}\left(2\sigma\sqrt{\frac{\log p'}{n}} + a\sqrt{\frac{\log p'}{n}}\|\theta^*\|_1\right),$$

the constraint $\|\theta^* - \bar{\theta}\|_\infty \leq \lambda_n$ with high probability, which completes the proof.

E. Proof of Corollary 3

For any $v \in \mathbb{R}^p$, the maximum absolute element of $\left[T_\nu\left(\frac{X^\top X}{n}\right)\right]^{-1} v$ is bounded by

$$\left\|\left[T_\nu\Big(\frac{X^\top X}{n}\Big)\right]^{-1} v\right\|_\infty \leq \left\|\left[T_\nu\Big(\frac{X^\top X}{n}\Big)\right]^{-1}\right\|_\infty \|v\|_\infty.$$

Moreover, since the maximum group cardinality is m, we have

$$\left\|\left[T_\nu\Big(\frac{X^\top X}{n}\Big)\right]^{-1} v\right\|^*_{\mathcal{G},\alpha} \leq \left\|\left[T_\nu\Big(\frac{X^\top X}{n}\Big)\right]^{-1} v\right\|_\infty m^{1/\alpha^*}$$
$$\leq \left\|\left[T_\nu\Big(\frac{X^\top X}{n}\Big)\right]^{-1}\right\|_\infty \|v\|_\infty m^{1/\alpha^*}$$

Now, we can derive the upper bound of $\|\theta^* - \bar{\theta}\|^*_{\mathcal{G},\alpha}$:

$$\|\theta^* - \bar{\theta}\|^*_{\mathcal{G},\alpha}$$
$$= \left\|\left[T_\nu\Big(\frac{X^\top X}{n}\Big)\right]^{-1} T_\nu\Big(\frac{X^\top X}{n}\Big)\theta^* - \left[T_\nu\Big(\frac{X^\top X}{n}\Big)\right]^{-1}\frac{X^\top y}{n}\right\|^*_{\mathcal{G},\alpha}$$
$$\leq \left\|\left[T_\nu\Big(\frac{X^\top X}{n}\Big)\right]^{-1}\right\|_\infty \left\|T_\nu\Big(\frac{X^\top X}{n}\Big)\theta^* - \frac{X^\top y}{n}\right\|_\infty m^{1/\alpha^*}.$$

Finally, by the same reasoning and conditions as in Section D, we have, conditioned on the event $\mathcal{A}$,

$$\|\theta^* - \bar{\theta}\|^*_{\mathcal{G},\alpha} \leq \frac{m^{1/\alpha^*}}{\delta_{\min}}\left(2\sigma\sqrt{\frac{\log p'}{n}} + a\sqrt{\frac{m\log p'}{n}}\|\theta^*\|_1\right).$$

Therefore, given the choice of λ_n as in the statement, we have $\|\theta^* - \bar{\theta}\|^*_{\mathcal{G},\alpha} \leq \lambda_n$ with high probability, and we can directly apply Theorem 2.

References

Bach, F. Consistency of trace norm minimization. *Journal of Machine Learning Research*, 9:1019–1048, June 2008.

Bach, F., Jenatton, R., Mairal, J., and Obozinski, G. Structured sparsity through convex optimization. *Statistical Science*, 27 (4):450–468, 2012.

Baraniuk, R. G., Cevher, V., Duarte, M. F., and Hegde, C. Model-based compressive sensing. Technical report, Rice University, 2008. Available at arxiv:0808.3572.

Bickel, P., Ritov, Y., and Tsybakov, A. Simultaneous analysis of lasso and dantzig selector. 37(4):1705–1732, 2009. Annals of Statistics.

Candes, E. and Tao, T. The Dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics*, 2006.

Candes, E. J. and Tao, T. The dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics*, 35(6): 2313–2351, 2007.

Donoho, D. For most large underdetermined systems of linear equations, the minimal ℓ_1-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics*, 59(6):797–829, June 2006.

Fan, J. and Lv, J. Sure independence screening for ultra-high dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) (JRSSB)*, 70:849–911, 2008.

Friedman, J., Hastie, T., Hofling, H., and Tibshirani, R. Pathwise coordinate optimization. *Annals of Applied Statistics*, 2007.

Genovese, C. R., Jin, J., Wasserman, L., and Yao, Z. A comparison of the lasso and marginal regression. *Journal of Machine Learning Research (JMLR)*, 13:2107–2143, 2012.

Hsieh, C. J., Sustik, M., Dhillon, I., and Ravikumar, P. Sparse inverse covariance matrix estimation using quadratic approximation. In *Neur. Info. Proc. Sys. (NIPS)*, 24, 2011.

Huang, J., Zhang, T., and Metaxas, D. Learning with structured sparsity. *Journal of Machine Learning Research (JMLR)*, 12: 3371–3412, 2011.

Jacob, L., Obozinski, G., and Vert, J. P. Group Lasso with Overlap and Graph Lasso. In *International Conference on Machine Learning (ICML)*, pp. 433–440, 2009.

Lounici, K., Pontil, M., Tsybakov, A. B., and van de Geer, S. Taking advantage of sparsity in multi-task learning. Technical Report arXiv:0903.1468, ETH Zurich, March 2009.

Lozano, A. C., Swirszcz, G., and Abe, N. Group orthogonal matching pursuit for variable selection and prediction. In *Neur. Info. Proc. Sys (NIPS)*, 2009.

Mallat, S. and Zhang, Z. Matching pursuits with time-frequency dictionaries. *IEEE Trans. Signal Processing*, December 1993.

Meinshausen, N. and Bühlmann, P. High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34: 1436–1462, 2006.

Meinshausen, N. and Yu, B. Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics*, 37 (1):246–270, 2009.

Negahban, S. and Wainwright, M. J. Simultaneous support recovery in high-dimensional regression: Benefits and perils of $\ell_{1,\infty}$-regularization. Technical report, Department of Statistics, UC Berkeley, April 2009.

Negahban, S., Ravikumar, P., Wainwright, M. J., and Yu, B. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

Obozinski, G., Wainwright, M. J., and Jordan, M. I. Union support recovery in high-dimensional multivariate regression. Technical report, Department of Statistics, UC Berkeley, August 2008.

Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5: 935–980, 2011.

Recht, B., Fazel, M., and Parrilo, P. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, Vol 52(3):471–501, 2010.

Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.

Tropp, J. A., Gilbert, A. C., and Strauss, M. J. Algorithms for simultaneous sparse approximation. *Signal Processing*, 86:572–602, April 2006. Special issue on "Sparse approximations in signal and image processing".

van de Geer, S. and Buhlmann, P. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3: 1360–1392, 2009.

van de Geer, S., Bühlmann, P., and Zhou, S. The adaptive and the thresholded lasso for potentially misspecified models (and a lower bound for the lasso). *Electronic Journal of Statistics*, 5:688–749, 2011.

Varah, J. M. A lower bound for the smallest singular value of a matrix. *Linear Algebra and its Applications*, 11(1):3–5, 1975.

Wainwright, M. J. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Information Theory*, 55:2183–2202, May 2009.

Wille, A., Zimmermann, P., and Vranova, E. [and others]. Sparse graphical gaussian modeling of the isoprenoid gene network in arabidopsis thaliana. *Genome Biology*, 5, 2004.

Yuan, M. and Lin, Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society B*, 1(68):49, 2006.

Zhang, J., Jeng, X. J., and Liu, H. Some two-step procedures for variable selection in high-dimensional linear regression. *Arxiv preprint arXiv:0810.1644*, 2008a.

Zhang, T. Sparse recovery with orthogonal matching pursuit under rip. Tech Report arXiv:1005.2249, May 2010.

Zhang, Z., Dolecek, L., Nikolic, B., Anantharam, V., and Wainwright, M. J. Lowering LDPC error floors by post-processing. In *Proc. IEEE GLOBECOM*, September 2008b.

Zhao, P. and Yu, B. On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7:2541–2567, 2006.

Zhao, P., Rocha, G., and Yu, B. Grouped and hierarchical model selection through composite absolute penalties. *Annals of Statistics*, 37(6A):3468–3497, 2009.