# Supplementary Materials

## A. Sparse graphical models

We derive a corollary for the trimmed Graphical lasso:

$$\underset{\boldsymbol{\Theta}\in\mathcal{S}^p_{++}}{\text{minimize}}\ \text{trace}\big(\widehat{\Sigma}\boldsymbol{\Theta}\big) - \log\det\big(\boldsymbol{\Theta}\big) + \lambda_n\mathcal{R}(\boldsymbol{\Theta}_{\text{off}};h). \tag{12}$$

Following the strategy of (Loh & Wainwright, 2017), we assume that the sample size scales with the row sparsity $d$ of true inverse covariance $\boldsymbol{\Theta}^* = (Cov(X))^{-1}$, which is a milder condition than other works ($n$ scaling with $k$, the number of non zero entries of $\boldsymbol{\Theta}^*$):

**Corollary 2.** *Consider the program (12) where the $x_i$'s are drawn from a sub-Gaussian and sample size $n > c_0 d^2 \log p$ with the selection of*

*(a) $\lambda_n \geq c_\ell\sqrt{\frac{\log p}{n}}$ for some constant $c_\ell$ depending only on $\boldsymbol{\Theta}^*$*

*(b) $h$ satisfying: for any selection of $T \subseteq \{1,2,\ldots,p\}\times\{1,2,\ldots,p\}$ s.t. $|T| = h$,*

$$\begin{aligned}&\left\|\!\left|\big(\boldsymbol{\Theta}^*\otimes\boldsymbol{\Theta}^*\big)_{UU}\right|\!\right\|_\infty \leq c_\infty, \\ &\max\left\{\|\!|\widehat{\Gamma}_{U^cU^c}|\!\|_\infty, \|\!|(\widehat{\Gamma}_{UU})^{-1}|\!\|_\infty\right\} \leq c_u \quad \textit{and} \\ &\left\|\!\left|\big(\boldsymbol{\Theta}^{*-1}\otimes\boldsymbol{\Theta}^{*-1}\big)_{U^cU}\Big(\big(\boldsymbol{\Theta}^{*-1}\otimes\boldsymbol{\Theta}^{*-1}\big)_{UU}\Big)^{-1}\right|\!\right\|_\infty \leq \eta.\end{aligned} \tag{13}$$

*Further suppose that $\frac{1}{2}\boldsymbol{\Theta}^*_{\min}$ is lower bounded by $c_1\sqrt{\frac{\log p}{n}} + 2\lambda_n c_\infty$ for some constant $c_1$. Then with high probability at least $1 - c_2\exp(-c_3\log p)$, any local minimum $\widetilde{\boldsymbol{\Theta}}$ of (4) has the following property:*

*(a) For every pair $j_1 \in S, j_2 \in S^c$, $|\widetilde{\boldsymbol{\Theta}}_{j_1}| > |\widetilde{\boldsymbol{\Theta}}_{j_2}|$,*

*(b) If $h < k$, all $j \in S^c$ are successfully estimated as zero and we have*

$$\|\widetilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_\infty \leq c_1\sqrt{\frac{\log p}{n}} + 2\lambda_n c_\infty \tag{14}$$

*(c) If $h \geq k$, at least the smallest $p - h$ entries in $S^c$ have exactly zero and we have*

$$\|\widetilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_\infty \leq c_1\sqrt{\frac{\log p}{n}}. \tag{15}$$

Note that condition (13) is the incoherence condition studied in (Ravikumar et al., 2011), and the same remarks as those for sparse linear models (see Section 3.1) can be made.

## B. Proofs

### B.1. Proof of Theorem 1

We extend the standard PDW technique (Wainwright, 2009c; Yang et al., 2015; Loh & Wainwright, 2017) for the trimmed regularizers. For any **fixed** $T$, we construct a primal and dual witness pair with the strict dual feasibility. Specifically, given the fixed $T$, consider the following program:

$$\underset{\boldsymbol{\theta}\in\Omega}{\text{minimize}}\ \mathcal{L}(\boldsymbol{\theta};\mathcal{D}) + \lambda_n\sum_{j\in T^c}|\theta_j|. \tag{16}$$

Note that the program (16) is convex (under (C-1)) where the regularizer is only effective over entries in (fixed) $T^c$. We construct the primal and dual pair $(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{z}})$ by the following restricted program

$$\widehat{\boldsymbol{\theta}} \in \underset{\boldsymbol{\theta}\in\mathbb{R}^U:\, \boldsymbol{\theta}\in\Omega}{\operatorname{argmin}} \quad \mathcal{L}(\boldsymbol{\theta}) + \lambda_n \mathcal{R}(\boldsymbol{\theta}; h) \tag{17}$$

and (5). The following lemma can guarantee under the strict dual feasibility that any solution of (16) has the same sparsity structure on $T^c$ with $\widehat{\boldsymbol{\theta}}$. Moreover, since the restricted program (5) is strictly convex as shown in the lemma below, we can conclude that $\widehat{\boldsymbol{\theta}}$ is the unique minimum point of the restricted program (16) given $T$.

**Lemma 1.** *Suppose that there exists a primal optimal solution $\widehat{\boldsymbol{\theta}}$ for* (16) *with associated sub-gradient (or dual) $\widehat{\boldsymbol{z}}$ such that $\|\widehat{\boldsymbol{z}}_{U^c}\|_\infty < 1$. Then any optimal solution $\widetilde{\boldsymbol{\theta}}$ of* (16) *will satisfy $\widetilde{\boldsymbol{\theta}}_j = 0$ for all $j \in U^c$.*

*Proof.* The lemma can be directly achieved by the basic property of convex optimization problem, as developed in existing works using PDW (Wainwright, 2009c; Yang et al., 2015). Note that even though the original problem with the trimmed regularizer is not convex, (16) given $T$ is convex. Therefore, by complementary slackness, we have $\sum_{j\in T^c} |\widetilde{\boldsymbol{\theta}}_j| = \langle \widehat{\boldsymbol{z}}_{T^c}, \widetilde{\boldsymbol{\theta}}_{T^c}\rangle$. Therefore, any optimal solution of (16) will satisfy $\widetilde{\boldsymbol{\theta}}_j = 0$ for all $j \in U^c$ since the associated (absolute) sub-gradient is strictly smaller than 1 by the assumption in the statement. □

**Lemma 2** (Section A.2 of (Loh & Wainwright, 2017))**.** *Under* (C-2)*, the loss function $\mathcal{L}(\boldsymbol{\theta})$ is strictly convex on $\boldsymbol{\theta} \in \mathbb{R}^U$ and hence $\big(\nabla^2\mathcal{L}(\boldsymbol{\theta})\big)_{UU}$ is invertible if $n \geq \frac{2\tau_1}{\kappa_l}(k+h)\log p$.*

Now from the definition of $\widehat{Q}$, we have

$$\widehat{Q}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) = \nabla\mathcal{L}(\widehat{\boldsymbol{\theta}}) - \nabla\mathcal{L}(\boldsymbol{\theta}^*) \tag{18}$$

where $\widehat{Q}$ is decomposed as $\begin{bmatrix} \widehat{Q}_{UU} & \widehat{Q}_{UU^c} \\ \widehat{Q}_{U^cU} & \widehat{Q}_{U^cU^c} \end{bmatrix}$. Then by the invertibility of $\big(\nabla^2\mathcal{L}(\boldsymbol{\theta})\big)_{UU}$ in Lemma 2 and the zero sub-gradient condition in (5) we have

$$\widehat{\boldsymbol{\theta}}_U - \boldsymbol{\theta}^*_U = \Big(\widehat{Q}_{UU}\Big)^{-1}\Big( - \nabla\mathcal{L}(\boldsymbol{\theta}^*)_U - \lambda_n \widehat{\boldsymbol{z}}_U\Big). \tag{19}$$

Since both $\widehat{\boldsymbol{\theta}}_{U^c}$ and $\boldsymbol{\theta}^*_{U^c}$ are zero vectors, we obtain

$$\begin{aligned} \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_\infty &= \Big\|\Big(\widehat{Q}_{UU}\Big)^{-1}\Big( - \nabla\mathcal{L}(\boldsymbol{\theta}^*)_U - \lambda_n \widehat{\boldsymbol{z}}_U\Big)\Big\|_\infty \\ &\leq \Big\|\Big(\widehat{Q}_{UU}\Big)^{-1}\nabla\mathcal{L}(\boldsymbol{\theta}^*)_U\Big\|_\infty + \lambda_n \big|\!\big|\!\big|\big(\widehat{Q}_{UU}\big)^{-1}\big|\!\big|\!\big|_\infty. \end{aligned} \tag{20}$$

Therefore, under the assumption on $\boldsymbol{\theta}^*_{\min}$ in the statement, the selection of $T$ in which there exists some $(j, j')$ s.t. $j \in S$, $j \in T^c$, $j' \in S^c$ and $j' \in T$, yields contradictory solution with (2). Under the strict dual feasibility condition for this specific choice of $T$ (along with Lemma 2) can guarantee that there is no local minimum for that choice of $T$. Hence, (21) can guarantee that for every pair $(j_1, j_2)$ such that $j_1 \in S$ and $j_2 \notin S$, we have $|\widetilde{\boldsymbol{\theta}}_{j_1}| > |\widetilde{\boldsymbol{\theta}}_{j_2}|$ (since $\widetilde{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}$). Note that for any *valid* selection of $T$, this statement holds. This immediately implies that any local minimum of (2) satisfies this property as well, as in the statement.

Finally turning to the bound when $h \geq k$, we have $U = T$ since all entries in $S$ are not penalized as shown above. In this case, $\widehat{\boldsymbol{z}}_U$ becomes zero vector (since $V$ is empty in the construction of $\widehat{\boldsymbol{z}}$), and the bound in (21) will be tighter as

$$\begin{aligned} \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_\infty &= \Big\|\Big(\widehat{Q}_{UU}\Big)^{-1}\Big( - \nabla\mathcal{L}(\boldsymbol{\theta}^*)_U - \lambda_n \widehat{\boldsymbol{z}}_U\Big)\Big\|_\infty \\ &\leq \Big\|\Big(\widehat{Q}_{UU}\Big)^{-1}\nabla\mathcal{L}(\boldsymbol{\theta}^*)_U\Big\|_\infty, \end{aligned} \tag{21}$$

as claimed.

### B.2. Proof of Theorem 2

Here we adopt the strategy developed in Yang et al. (2018) for analyzing local optima of trimmed loss function. Since our loss function $\mathcal{L}$ is convex, the story derived in this subsection can also be applied to results of Negahban et al. (2012). However, in order to simplify the procedure, we will not utilize the convexity of $\mathcal{L}$ and instead place the side constraint $\|\boldsymbol{\theta}\|_1 \leq R$ and some additional assumptions (see (Loh & Wainwright, 2013) for details). As in Yang et al. (2018), we introduce the the shorthand to denote local optimal error vector: $\widetilde{\Delta} := \widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*$ given an *arbitrary* local minimum $(\widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{w}})$ of (2). We additionally define $H$ to denote the set of indices not penalized by $\widetilde{\boldsymbol{w}}$ (that is, $\widetilde{\boldsymbol{w}}_j = 0$ for $j \in H$, $\widetilde{\boldsymbol{w}}_j = 1$ for $j \in H^c$ and $|H| = h$). Utilizing the fact that $(\widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{w}})$ is a local minimum of (2), we have an inequality

$$\big\langle \nabla_{\boldsymbol{\theta}} \mathcal{L}\big(\boldsymbol{\theta}^* + \widetilde{\Delta}\big),\, \widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta} \big\rangle \leq -\big\langle \partial \lambda \mathcal{R}(\boldsymbol{\theta}^* + \widetilde{\Delta}; h),\, \widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta} \big\rangle \quad \text{for any feasible } \boldsymbol{\theta}.$$

This inequality comes from the first order stationary condition (see Loh & Wainwright (2015) for details) in terms of $\boldsymbol{\theta}$ fixing $\boldsymbol{w}$ at $\widetilde{\boldsymbol{w}}$. Here, if we take $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ above, we have

$$\big\langle \nabla \mathcal{L}\big(\boldsymbol{\theta}^* + \widetilde{\Delta}\big),\, \widetilde{\Delta} \big\rangle \leq -\big\langle \partial \lambda \mathcal{R}(\boldsymbol{\theta}^* + \widetilde{\Delta}; h),\, \widetilde{\Delta} \big\rangle \overset{(i)}{\leq} \lambda(\|\boldsymbol{\theta}^*_{H^c}\|_1 - \|\widetilde{\boldsymbol{\theta}}_{H^c}\|_1)$$

where $S$ is true support set of $\boldsymbol{\theta}^*$ and the inequality $(i)$ holds due to the convexity of $\ell_1$ norm.

**i)** $h < k$**:** By Theorem 1, we can guarantee with high probability that $H \subset S$. Then, by triangular inequality (in inequality $(ii)$ below) and the fact that $\boldsymbol{\theta}^*$ is $S$-sparse vector, we have

$$\begin{aligned} \big\langle \nabla \mathcal{L}\big(\boldsymbol{\theta}^* + \widetilde{\Delta}\big),\, \widetilde{\Delta} \big\rangle &\leq \lambda(\|\boldsymbol{\theta}^*_{H^c}\|_1 + \|\widetilde{\Delta}_{S^c}\|_1 - \|\widetilde{\Delta}_{S^c}\|_1 - \|\widetilde{\boldsymbol{\theta}}_{H^c}\|_1) = \lambda(\|\boldsymbol{\theta}^*_{H^c} + \widetilde{\Delta}_{S^c}\|_1 - \|\widetilde{\Delta}_{S^c}\|_1 - \|\widetilde{\boldsymbol{\theta}}_{H^c}\|_1) \\ &\overset{(ii)}{\leq} \lambda\big(\|\boldsymbol{\theta}^*_{H^c} + \widetilde{\Delta}_{S^c} + \widetilde{\Delta}_{S-H}\|_1 + \|\widetilde{\Delta}_{S-H}\|_1 - \|\widetilde{\Delta}_{S^c}\|_1 - \|\widetilde{\boldsymbol{\theta}}_{H^c}\|_1\big) = \lambda(\|\widetilde{\Delta}_{S-H}\|_1 - \|\widetilde{\Delta}_{S^c}\|_1)\,. \end{aligned} \tag{22}$$

Combining (22) and (C-2) yields

$$\kappa_l \|\widetilde{\Delta}\|_2^2 - \tau_1 \frac{\log p}{n} \|\widetilde{\Delta}\|_1^2 \leq \big\langle \nabla \mathcal{L}\big(\boldsymbol{\theta}^* + \widetilde{\Delta}\big) - \nabla \mathcal{L}\big(\boldsymbol{\theta}^*\big),\, \widetilde{\Delta} \big\rangle \leq -\big\langle \nabla \mathcal{L}\big(\boldsymbol{\theta}^*\big),\, \widetilde{\Delta} \big\rangle + \lambda\left(\|\widetilde{\Delta}_{S-H}\|_1 - \|\widetilde{\Delta}_{S^c}\|_1\right).$$

If we assume $\max\left\{\|\nabla \mathcal{L}\big(\boldsymbol{\theta}^*\big)\|_\infty, 2\rho\tau_1 \frac{\log p}{n}\right\} \leq \frac{\lambda}{4}$ (which are slightly different to assumptions in the statement, however they are purely for simplicity and can be relaxed if we use the convexity of $\mathcal{L}$, as we mentioned in the beginning of the proof), we can conclude that

$$\begin{aligned} 0 \leq \kappa_l \|\widetilde{\Delta}\|_2^2 &\leq \big\|\nabla \mathcal{L}\big(\boldsymbol{\theta}^*\big)\big\|_\infty \|\widetilde{\Delta}\|_1 + \lambda\left(\|\widetilde{\Delta}_{S-H}\|_1 - \|\widetilde{\Delta}_{S^c}\|_1\right) + 2\rho\tau_1 \frac{\log p}{n} \|\widetilde{\Delta}\|_1 \\ &\leq \frac{\lambda}{2} \|\widetilde{\Delta}\|_1 - \lambda \|\widetilde{\Delta}_{S^c}\|_1 + \lambda \|\widetilde{\Delta}_{S-H}\|_1 \leq \frac{\lambda}{2} \|\widetilde{\Delta}_S\|_1 - \frac{\lambda}{2} \|\widetilde{\Delta}_{S^c}\|_1 + \lambda \|\widetilde{\Delta}_{S-H}\|_2 \leq \frac{\lambda}{2} \|\widetilde{\Delta}_S\|_1 + \lambda \|\widetilde{\Delta}_{S-H}\|_2. \end{aligned} \tag{23}$$

As a result, we can finally have an $\ell_2$ error bound as follows:

$$\kappa_l \|\widetilde{\Delta}\|_2^2 \leq \frac{\lambda\sqrt{k}}{2} \|\widetilde{\Delta}_S\|_2 + \lambda\sqrt{k-h} \|\widetilde{\Delta}_{S-H}\|_2 \leq \left(\frac{\lambda\sqrt{k}}{2} + \lambda\sqrt{k-h}\right) \|\widetilde{\Delta}\|_2$$

implying that

$$\|\widetilde{\Delta}\|_2 \leq \frac{1}{\kappa_l} \left(\frac{\lambda\sqrt{k}}{2} + \lambda\sqrt{k-h}\right).$$

**ii)** $h \geq k$**:** As in the previous case, Theorem 1 can guarantee $S \subseteq H$ where equality holds if $h = k$. Instead of (22), now we have

$$\begin{aligned} &\big\langle \nabla \mathcal{L}\big(\boldsymbol{\theta}^* + \widetilde{\Delta}\big),\, \widetilde{\Delta} \big\rangle \leq \lambda(\|\boldsymbol{\theta}^*_{H^c}\|_1 + \|\widetilde{\Delta}_{H^c}\|_1 - \|\widetilde{\Delta}_{H^c}\|_1 - \|\widetilde{\boldsymbol{\theta}}_{H^c}\|_1) \\ &= \lambda(\|\boldsymbol{\theta}^*_{H^c} + \widetilde{\Delta}_{H^c}\|_1 - \|\widetilde{\Delta}_{H^c}\|_1 - \|\widetilde{\boldsymbol{\theta}}_{H^c}\|_1) = -\|\widetilde{\Delta}_{H^c}\|_1. \end{aligned} \tag{24}$$

By similar reasoning in the case of i), we combine (24) and (C-2) to obtain

$$0 \leq \kappa_l \|\widetilde{\Delta}\|_2^2 \leq \frac{\lambda}{2}\|\widetilde{\Delta}\|_1 - \lambda\|\widetilde{\Delta}_{H^c}\|_1 \leq \frac{\lambda}{2}\|\widetilde{\Delta}_H\|_1 - \frac{\lambda}{2}\|\widetilde{\Delta}_{H^c}\|_1 \leq \frac{\lambda}{2}\|\widetilde{\Delta}_H\|_1 \leq \frac{\lambda\sqrt{h}}{2}\|\widetilde{\Delta}\|_2 \tag{25}$$

implying that

$$\|\widetilde{\Delta}\|_2 \leq \frac{1}{\kappa_l}\frac{\lambda\sqrt{h}}{2}\,.$$

### B.3. Proof of Corollary 1

The proof our corollary is similar to that of Corollary 1 of (Loh & Wainwright, 2017), who derive the result for $(\mu, \gamma)$-amenable regularizers. Here we only describe the parts that need to be modified from (Loh & Wainwright, 2017).

In order to utilize theorems in the main paper, we need to establish the RSC condition (C-2) and the strict dual feasibility: $\|\widehat{\boldsymbol{z}}_{U^c}\|_\infty \leq 1 - \delta$

First, the RSC is known to hold w.h.p as shown in several previous works such as Lemma 3.

**Lemma 3** (Corollary 1 of (Loh & Wainwright, 2015))**.** *The RSC condition in (C-2) for linear models holds with high probability with $\kappa_l = \frac{1}{2}\lambda_{\min}(\Sigma_x)$ and $\tau_1 \asymp 1$, under sub-Gaussian assumptions in the statement.*

In order to show the remaining strict dual feasibility condition of our PDW construction, we consider (18) (by the zero-subgradient and the definition of $\widehat{Q}$) in the block form:

$$\begin{bmatrix} \widehat{Q}_{TT} & \widehat{Q}_{TV} & \widehat{Q}_{TU^c} \\ \widehat{Q}_{VT} & \widehat{Q}_{VV} & \widehat{Q}_{VU^c} \\ \widehat{Q}_{U^cT} & \widehat{Q}_{U^cV} & \widehat{Q}_{U^cU^c} \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*_T \\ \widehat{\boldsymbol{\theta}}_V - \boldsymbol{\theta}^*_V \\ \mathbf{0} \end{bmatrix} + \begin{bmatrix} \nabla\mathcal{L}(\boldsymbol{\theta}^*)_T \\ \nabla\mathcal{L}(\boldsymbol{\theta}^*)_V \\ \nabla\mathcal{L}(\boldsymbol{\theta}^*)_{U^c} \end{bmatrix} + \lambda_n \begin{bmatrix} \mathbf{0} \\ \widehat{\boldsymbol{z}}_V \\ \widehat{\boldsymbol{z}}_{U^c} \end{bmatrix} = \mathbf{0}. \tag{26}$$

By simple manipulation, we can obtain

$$\widehat{\boldsymbol{z}}_{U^c} = \frac{1}{\lambda_n}\left\{-\nabla\mathcal{L}(\boldsymbol{\theta}^*)_{U^c} + \widehat{Q}_{U^cU}\left(\widehat{Q}_{UU}\right)^{-1}\left(-\nabla\mathcal{L}(\boldsymbol{\theta}^*)_U - \lambda_n\widehat{\boldsymbol{z}}_V\right)\right\}. \tag{27}$$

Here note that our construction of PDW can guarantee the $\ell_\infty$ bound in (21). In case of (4), since we have $\nabla\mathcal{L}(\boldsymbol{\theta}) = \widehat{\Gamma}\boldsymbol{\theta} - \widehat{\gamma}$ and $\nabla^2\mathcal{L}(\boldsymbol{\theta}) = \widehat{\Gamma}$ where $(\widehat{\Gamma}, \widehat{\gamma}) = \left(\frac{X^\top X}{n}, \frac{X^\top \boldsymbol{y}}{n}\right)$, we need to show below that

$$\begin{aligned} \widehat{\boldsymbol{z}}_{U^c} &\leq \frac{1}{\lambda_n}\left\{-\widehat{\Gamma}_{U^cU}\boldsymbol{\theta}^*_U + \widehat{\gamma}_{U^c} + \widehat{\Gamma}_{U^cU}\boldsymbol{\theta}^*_U - \widehat{\Gamma}_{U^cU}\left(\widehat{\Gamma}_{UU}\right)^{-1}\widehat{\gamma}_U\right\} + \left|\left|\left|\widehat{\Gamma}_{U^cU}\left(\widehat{\Gamma}_{UU}\right)^{-1}\right|\right|\right|_\infty \\ &\leq \frac{1}{\lambda_n}\left\{\widehat{\gamma}_{U^c} - \widehat{\Gamma}_{U^cU}\left(\widehat{\Gamma}_{UU}\right)^{-1}\widehat{\gamma}_U\right\} + \eta \end{aligned} \tag{28}$$

for the strict dual feasibility from (27). As derived in (Loh & Wainwright, 2017), we can write

$$\left\|\widehat{\gamma}_{U^c} - \widehat{\Gamma}_{U^cU}\left(\widehat{\Gamma}_{UU}\right)^{-1}\widehat{\gamma}_U\right\|_\infty = \left\|\frac{X_{U^c}^\top \Pi \boldsymbol{\epsilon}}{n}\right\|_\infty \tag{29}$$

where $\Pi$ is an orthogonal project matrix on $X_U$: $I - X_U(X_U^\top X_U)^{-1}X_U^\top$.

For any $j$, we define $u_j$ such that $e_j^\top \frac{X_{U^c}^\top \Pi \boldsymbol{\epsilon}}{n} := u_j^\top \boldsymbol{\epsilon}$. Then we have

$$\|u_j\|_2^2 = \left\|\frac{\Pi X_{U^c} e_j}{n}\right\|_2^2 \leq \left\|\frac{X_{U^c} e_j}{n}\right\|_2^2 \leq \frac{c_u}{n}. \tag{30}$$

Hence by the sub-Gaussian tail bounds followed by a union bound, we can conclude that

$$\left\|\widehat{\gamma}_{U^c} - \widehat{\Gamma}_{U^cU}\left(\widehat{\Gamma}_{UU}\right)^{-1}\widehat{\gamma}_U\right\|_\infty \leq C\sqrt{\frac{\log p}{n}} \tag{31}$$

with probability at least $1 - c\exp(-c' \log p)$ for *all* selections of $T$. We can establish have strict dual feasibility for any selection of $T$ w.h.p, provided $\lambda_n > \frac{C}{1-\eta}\sqrt{\frac{\log p}{n}}$, and now turn to $\ell_\infty$ bounds. From (6), we have

$$\left\| \widehat{\Gamma}_{UU} \left( \widehat{\Gamma}_{UU} \boldsymbol{\theta}^*_U - \widehat{\gamma}_U \right) \right\|_\infty = \left\| \left( \frac{X_U^\top X_U}{n} \right)^{-1} \left( \frac{X_U^\top \boldsymbol{\epsilon}}{n} \right) \right\|_\infty . \tag{32}$$

Then for $j \in U$, we define $v$ such that $e_j^\top \left( \frac{X_U^\top X_U}{n} \right)^{-1} \left( \frac{X_U^\top \boldsymbol{\epsilon}}{n} \right) := v_j^\top \boldsymbol{\epsilon}$. Since for any selection of $T$, $\|v_j\|_2^2$ is bounded as follows:

$$\|v_j\|_2^2 = \frac{1}{n^2} \left\| X_U \left( \frac{X_U^\top X_U}{n} \right)^{-1} e_j \right\|_2^2 = \frac{1}{n} \left| e_j^\top \left( \frac{X_U^\top X_U}{n} \right)^{-1} e_j \right|_2^2 \leq \frac{c_u}{n}. \tag{33}$$

Similarly by the sub-Gaussian tail bound and a union bound over $j$, we can obtain

$$\left\| \widehat{\Gamma}_{UU} \left( \widehat{\Gamma}_{UU} \boldsymbol{\theta}^*_U - \widehat{\gamma}_U \right) \right\|_\infty \leq C \sqrt{\frac{\log p}{n}} \tag{34}$$

with probability at least $1 - c\exp(-c' \log p)$.

### B.4. Proof of Corollary 2

As in the proof of Corollary 1, the proof procedure is quite similar to that of Corollary 4 of (Loh & Wainwright, 2017). Deriving upper bounds on $S$ in (Loh & Wainwright, 2017) can be seamlessly extendable to upper bounds on $U$ for any selection of $T \subseteq \{1, 2, \ldots, p\} \times \{1, 2, \ldots, p\}$ s.t. $|T| = h$. mainly because the required upper bounds are related to entry-wise maximum on the true support $S$ but entry-wise maximum in this case is uniformly upper bounded for all entries.

Specifically, it computes the upper bound of $\|\text{vec}(\widehat{\Sigma}_S - \Sigma^*_S)\|_\infty$ from the fact that $\|\text{vec}(\widehat{\Sigma} - \Sigma^*)\|_\infty \leq c\sqrt{\frac{\log p}{n}}$. This actually holds for any selection of $T$. Similarly, it computes the upper bound of $\max_{(j,k)\in S} |e_j^\top (\Sigma^* \Delta)^\ell \Sigma^* e_k|$ by Hölder's inequality and the definition of matrix induced norms: $|e_j^\top (\Sigma^* \Delta)^\ell \Sigma^* e_k| \leq \|e_j^\top (\Sigma^* \Delta)^{\ell-1}\|_1 \|\Delta \Sigma^* e_k\|_\infty \leq |||(\Sigma^* \Delta)^{\ell-1}|||_\infty \|\Delta\|_{\max} \|\Sigma^* e_k\|_1 \leq |||\Sigma^*|||_\infty^{\ell+1} |||\Delta|||_1^{\ell-1} \|\Delta\|_{\max}$, which clearly holds for any index $(j,k)$ beyond $S$. Finally, $\|\widehat{Q}_{SS} - \nabla^2 \mathcal{L}(\boldsymbol{\Theta}^*)_{SS}\|_\infty$ is shown to be upper bounded by the fact that $\|\widehat{Q}_{SS} - \nabla^2 \mathcal{L}(\boldsymbol{\Theta}^*)_{SS}\|_\infty \precsim d\sqrt{\frac{\log p}{n}}$.

The remaining proof of this result directly follows similar lines to the proof of Corollary 4 in (Loh & Wainwright, 2017).

### B.5. Proof of Theorem 3

*Proof.* From Algorithm 1, we obtain the relation

$$\begin{aligned} \frac{1}{\tau}(\boldsymbol{w}^k - \boldsymbol{w}^{k+1}) + r(\boldsymbol{\theta}^{k+1}) - r(\boldsymbol{\theta}^k) &\in \boldsymbol{r}(\boldsymbol{\theta}^{k+1}) + \partial\delta(\boldsymbol{w}^{k+1}|\mathcal{S}) \\ \frac{1}{\eta}(\boldsymbol{\theta}^k - \boldsymbol{\theta}^{k+1}) + \nabla\mathcal{L}(\boldsymbol{\theta}^{k+1}) - \nabla\mathcal{L}(\boldsymbol{\theta}^k) &\in \nabla\mathcal{L}(\boldsymbol{\theta}^{k+1}) + \lambda \sum_{i=1}^p w_i^{k+1} \partial r_i(\boldsymbol{\theta}^{k+1}) \end{aligned}$$

from the proximal gradient steps.

At $k$-th iteration, we have

$$\begin{aligned}
&\mathcal{L}(\boldsymbol{\theta}^{k+1}) + \lambda\langle \boldsymbol{w}^{k+1},\, \boldsymbol{r}(\boldsymbol{\theta}^{k+1})\rangle\\
\leq &\mathcal{L}(\boldsymbol{\theta}^{k}) + \langle \nabla\mathcal{L}(\boldsymbol{\theta}^{k}),\, \boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^{k}\rangle + \frac{L}{2}\|\boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^{k}\|^2 + \lambda\langle \boldsymbol{w}^{k+1},\, \boldsymbol{r}(\boldsymbol{\theta}^{k+1})\rangle\\
\leq &\mathcal{L}(\boldsymbol{\theta}^{k}) + \langle \nabla\mathcal{L}(\boldsymbol{\theta}^{k}) + \lambda\sum_{i=1}^{p} w_i^{k+1}\partial r_i(\boldsymbol{\theta}^{k+1}),\, \boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^{k}\rangle + \frac{L}{2}\|\boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^{k}\|^2 + \lambda\langle \boldsymbol{w}^{k+1},\, \boldsymbol{r}(\boldsymbol{\theta}^{k})\rangle\\
= &\mathcal{L}(\boldsymbol{\theta}^{k}) + \lambda\langle \boldsymbol{w}^{k},\, \boldsymbol{r}(\boldsymbol{\theta}^{k})\rangle - (\frac{1}{\eta} - \frac{L}{2})\|\boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^{k}\|^2 + \lambda\langle \boldsymbol{w}^{k+1} - \boldsymbol{w}^{k},\, \boldsymbol{r}(\boldsymbol{\theta}^{k})\rangle\\
\leq &\mathcal{L}(\boldsymbol{\theta}^{k}) + \lambda\langle \boldsymbol{w}^{k},\, \boldsymbol{r}(\boldsymbol{\theta}^{k})\rangle - (\frac{1}{\eta} - \frac{L}{2})\|\boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^{k}\|^2 + \lambda\langle \boldsymbol{w}^{k+1} - \boldsymbol{w}^{k},\, \frac{1}{\tau}(\boldsymbol{w}^{k} - \boldsymbol{w}^{k+1}) - \partial\delta(\boldsymbol{w}^{k+1})\rangle\\
\leq &f(\boldsymbol{\theta}^{k}) + \lambda\langle \boldsymbol{w}^{k},\, \boldsymbol{r}(\boldsymbol{\theta}^{k})\rangle - (\frac{1}{\eta} - \frac{L}{2})\|\boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^{k}\|^2 - \frac{\lambda}{\tau}\|\boldsymbol{w}^{k+1} - \boldsymbol{w}^{k}\|^2
\end{aligned}$$

If we choose $\eta = 1/L_f$, we have,

$$\frac{L_f}{2}\|\boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^{k}\|^2 + \frac{\lambda}{\tau}\|\boldsymbol{w}^{k+1} - \boldsymbol{w}^{k}\|^2 \leq F(\boldsymbol{\theta}^{k}) - F(\boldsymbol{\theta}^{k+1})$$

By telescoping both sides we get,

$$\frac{1}{K}\sum_{k=1}^{K}(\frac{L_f}{2}\|\boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^{k}\|^2 + \frac{\lambda}{\tau}\|\boldsymbol{w}^{k+1} - \boldsymbol{w}^{k}\|^2) \leq \frac{1}{K}(F(\boldsymbol{\theta}^{K}) - F^*).$$

Moreover we know that,

$$T(\boldsymbol{\theta}^{k+1}, \boldsymbol{w}^{k+1}) \leq (4 + 2\lambda L_r/L_f)\mathcal{G}_k.$$

□

## C. Simulations for Gaussian Graphical Models.

We now illustrate the usefulness trimmed regularization for sparse Gaussian Graphical Model estimation. We consider the “diamond” graph example described in (Ravikumar et al., 2011) (section 3.1.1). This graph $G = (V, E)$ has vertex set $V = \{1, 2, 3, 4\}$, with all edges except $(1, 4)$. We consider a family of true covariance matrices with diagonal entries $\Sigma^*_{ii} = 1$ for all $i \in V$; off-diagonal elements $\Sigma^*_{ij} = \rho$ for all edges $(i, j) \in E \setminus \{(2, 3)\}$; $\Sigma^*_{23} = 0$; and finally the entry corresponding to the non-edge $(1, 4)$ is set as $\Sigma^*_{14} = 2\rho^2$. We analyze the performance of Graphical Trimmed Lasso under two settings: $\rho \in \{0, 1, 0.3\}$. As discussed in (Ravikumar et al., 2011), if $\rho = 0.1$ the incoherence condition is satisfied ; if $\rho = 0.3$ it is violated. Under both settings, we report the probability of successful support recovery based on 100 replicate experiments for $n = 100$ and $\frac{p^2-h}{p^2} \in \{0.4, 0.5, ..., 1\}$ and compare it with Graphical Lasso, Graphical SCAD and Graphical MCP (The MCP and SCAD parameters were set to 2.5 and 3.0 as varying these did not affect the results significantly). For each method and replicate experiment, success is declared if the true support is recovered for at least one value of $\lambda_n$ along the solution path. We can see that for a wide range of values for the trimming parameter, Graphical Trimmed Lasso outperforms SCAD and MCP alternatives regardless of whether the incoherence condition holds or not. In addition its probability of success is always superior to that of vanilla Graphical Lasso, which fails to recover the true support when the incoherence condition is violated.

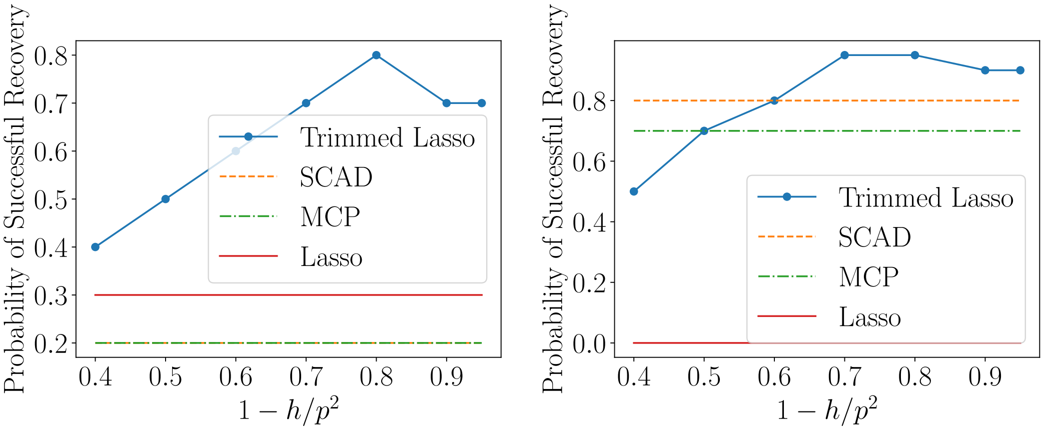

*Figure 6.* Probability of successful support recovery for Graphical Trimmed Lasso as $h$ vary, Graphical SCAD, Graphical MCP and Graphical Lasso. Left: incoherence condition holds. Right: incoherence condition is violated.