Identifying groups of strongly correlated variables through Smoothed Ordered Weighted ℓ_1-norms

Raman Sankaran
Indian Institute of Science
ramans@csa.iisc.ernet.in

Francis Bach
INRIA - École Normale Supérieure, Paris
francis.bach@inria.fr

Chiranjib Bhattacharyya
Indian Institute of Science
chiru@csa.iisc.ernet.in

Abstract

In this article, we provide additional statements and proofs complementing the main paper. We present here the proofs of the statements given in the main paper. The section numbers in this document are arranged in correspondence to the respective sections in the main paper.

3 Related Work: OWL, OSCAR, and SLOPE

3.1 Proof of Proposition 3.1

Proof. 1. Let $a = |w|$. Let us assume WLOG that $a_1 \geq \cdots \geq a_d \geq 0$. Then the Lovász extension $p(a) = \sum_{i=1}^{d} a_i c_i$, where $c_i = f(i) - f(i-1)$ (See [1]). We get the result.

2. The derived penalty is non-decreasing since $c \geq 0$. And since c forms a decreasing sequence, P is submodular. Hence the result.

□

4 SOWL - Definition and Properties

The below Lemma states that $\Omega_{\mathcal{S}}$ is a valid norm.

Lemma 4.A. *Let $w \in \mathbb{R}^d$. $\Omega_{\mathcal{S}}(w)$ defined in* (SOWL) *is a valid norm if $c_1 + \cdots + c_d \geq 0$.*

Proof. From Proposition 3.1, we see that we can derive a submodular function P such that $P(\emptyset) = 0$ and $P(A) > 0$ for $A \subseteq 1, \ldots, d$. Now, $\Omega_{\mathcal{S}}$ is a special case of norms proposed in [2, Section 2], which are indeed valid norms. □

The below statements show that for every c satisfying $c_1 \geq \cdots \geq c_d$, there exists $\tilde{c}$ satisfying $\tilde{c}_1 \geq \cdots \geq \tilde{c}_d \geq 0$, such that $\Omega_{\mathcal{S}}$ is same for both c and $\tilde{c}$.

Lemma 4.B. *Let $w \in \mathbb{R}^d$. Given $c \in \mathbb{R}^d$ such that $c_1 \geq \cdots \geq c_d$. Let k be the minimum integer such that $c_k + \cdots + c_d \geq 0$. Then define $\tilde{c} \in \mathbb{R}^d$ such that $\tilde{c}_i = c_i$ for $i = 1, \ldots, k-1$, $\tilde{c}_k = c_k + \cdots + c_d$, and $\tilde{c}_i = 0$ for $i = k+1, \ldots, d$. Explicitly mentioning the dependency of $\Omega_{\mathcal{S}}$ on c as $\Omega_{\mathcal{S}}(w; c)$, we have $\Omega_{\mathcal{S}}(w; c) = \Omega_{\mathcal{S}}(w; \tilde{c})$.*

Proof. Let P be the submodular function constructed from c, and $\tilde{P}$ be the corresponding function constructed from $\tilde{c}$. From [2, Lemma 3], we see that both have the same Lower Combinatorial Envelope, which implies that $\Omega_{\mathcal{S}}(w; c) = \Omega_{\mathcal{S}}(w; \tilde{c})$. □

4.1 Proof of Proposition 4.3

Proof. 1. Once we make the assumption on the lattice, the objective in (SOWL) is separable in terms of variables within each group $\mathcal{G}_j$. And the result follows.

2. This candidate δ_w is optimal only if for small perturbations around η_w, the objective function increases. Let us denote by $\Gamma(\eta) = \sum_{i=1}^{d} c_i \eta_{(i)}$, the Lovász extension of P. From [1], we see that around η_w, we have the decomposition of Γ as

$$\Gamma(\eta + d\eta) = \Gamma(\eta) + \sum_{j=1}^{k} \Gamma_j(d\eta_j),$$

where Γ_j is the lovasz extension of the function P_j, which is defined over all $C_j \subseteq \mathcal{G}_j$ as $P_j(C_j) = P(\mathcal{G}_1 \cup \cdots \cup \mathcal{G}_{j-1} \cup C_j) - P(\mathcal{G}_1 \cup \cdots \cup \mathcal{G}_{j-1})$. For all $d\eta$ sufficiently small, we require that

$$\sum_{i=1}^{d} \frac{w_i^2}{\eta_i + d\eta_i} + \Gamma(\eta + d\eta) > \sum_{i=1}^{d} \frac{w_i^2}{\eta_i} + \Gamma(\eta)$$

Proceedings of the 20^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2017, Fort Lauderdale, Florida, USA. JMLR: W&CP volume 54. Copyright 2017 by the author(s).

This is equivalent to

$$\sum_{i=1}^{d}\left(\frac{w_i^2}{\eta_i+d\eta_i}-\frac{w_i^2}{\eta_i}\right)+\sum_{j=1}^{k}\Gamma_j(d\eta_j)>0$$

$$\sum_{j=1}^{k}\Gamma_j(d\eta_j)>\sum_{i=1}^{d}w_i^2\left(\frac{1}{\eta_i}-\frac{1}{\eta_i+d\eta_i}\right)$$

$$=\sum_{i=1}^{d}\frac{w_i^2}{\eta_i}\left(\frac{d\eta_i}{(\eta_i+d\eta_i)}\right).$$

Following [3], the above equation is satisfied for all $d\eta>0$ if and only if $\forall j$,

$$\Gamma_j(d\eta_j)>\sum_{i\in\mathcal{G}_j}\frac{w_i^2}{\eta_i^2}d\eta_i=s^\top d\eta_j, \qquad (1)$$

where $s_i=\frac{w_i^2}{\eta_i^2}$. Defining $\hat{s}_i=\frac{w_i^2}{\|w_{\mathcal{G}_j}\|_2^2}$, and (1) is equivalent to $\hat{s}(C_j)\leq\frac{P_j(C_j)}{P_j(\mathcal{G}_j)},\forall C_j\subseteq\mathcal{G}_j$. The statement follows.

□

4.2 Stability of the grouping variable η_w

Proposition 4.C. *Given $w\in\mathbb{R}^d$, and let the minimizer of* (SOWL) *$\eta_w\in\mathcal{D}^k$. define the following quantities*

1. $\beta(w)=\min_{j<h}\frac{\left(\frac{\|w_{\mathcal{G}_j}\|_2}{\sqrt{\mathcal{A}_j(\mathcal{G}_j)}}-\frac{\|w_{\mathcal{G}_h}\|_2}{\sqrt{\mathcal{A}_h(\mathcal{G}_h)}}\right)}{\left(\sqrt{\frac{|\mathcal{G}_j|}{\mathcal{A}_j(\mathcal{G}_j)}}+\sqrt{\frac{|\mathcal{G}_h|}{\mathcal{A}_h(\mathcal{G}_h)}}\right)}$,
2. $\gamma_j(w_{\mathcal{G}_j})=\min_{i\in\mathcal{G}_j}\frac{\left(\frac{\|w_{u_i}\|_2}{\sqrt{\mathcal{A}_j(|u_i|)}}-\frac{\|w_{v_i}\|_2}{\sqrt{\mathcal{A}_j(|\mathcal{G}_j|)-\mathcal{A}_j(|u_i|)}}\right)}{\left(\sqrt{\frac{|u_i|}{\mathcal{A}_j(|u_i|)}}+\sqrt{\frac{|v_i|}{\mathcal{A}_j(|\mathcal{G}_j|)-\mathcal{A}_j(|u_i|)}}\right)}$,

where $u_i=\{\hat{i}\in\mathcal{G}_j|\hat{i}\leq i\}$, and $v_i=\mathcal{G}_j\setminus u_i$. Define $\gamma(w)=\min_j\gamma_j(w_j)$.

Let $\tilde{w}=w+\epsilon$, and let $\eta_{\tilde{w}}$ be the minimizer of (SOWL) *for $\tilde{w}$. Then $\eta_{\tilde{w}}\in\mathcal{D}^k$ if $\|\epsilon\|_\infty\leq\min(\beta(w),\gamma(w))$.*

Proof. It is easy to see that Condition 1 in Proposition (4.3) is met when $\|\epsilon\|_\infty\leq\beta_{\mathcal{D}}(w)$. And Condition 2 is satisfied when $\|\epsilon\|_\infty\leq\gamma_{\mathcal{D}}(w)$. □

4.3 Proof of Theorem 4.4

Proof. We construct the proof by establishing the following results. We introduce the notion of group coherency, which is inspired from [4] which will help establishing the proof.

Definition 4.D. *A group $\mathcal{G}_j$ having indices $[s,e]$, $0\leq s\leq e\leq d$, is defined to be Coherent if there does not exist an index i such that $\frac{w_s^2+\cdots+w_i^2}{c_s+\cdots+c_i}>\frac{w_{i+1}^2+\cdots+w_e^2}{c_{i+1}+\cdots+c_e}$.*

Proposition 4.E. *A group $\mathcal{G}_j$ is coherent if and only if the inequality* (3) *in Proposition 4.3 is satisfied.*

Proof. Trivial to see. □

The following Lemma gives the conditions under which two adjacent groups can be merged while maintaining coherency.

Lemma 4.F. *Consider a lattice $\mathcal{D}^{(k)}$ and assume $|w_1|\geq\cdots\geq|w_d|$. Let us assume that the groups $\mathcal{G}_j$ and $\mathcal{G}_{j+1}$ are coherent, and let $\max(\{i|i\in\mathcal{G}_j\})>\min(\{\hat{i}|\hat{i}\in\mathcal{G}_{j+1}\})$. The group $\mathcal{G}_j\cup\mathcal{G}_{j+1}$ is also coherent if and only if $\frac{\|w_{\mathcal{G}_j}\|_2}{\sqrt{\mathcal{A}_j(|\mathcal{G}_j|)}}<\frac{\|w_{\mathcal{G}_{j+1}}\|_2}{\sqrt{\mathcal{A}_{j+1}(|\mathcal{G}_{j+1}|)}}$.*

Proof. It the condition is not satisfied, obviously the merged group is not coherent. We now show that this is a sufficient condition too. W.l.o.g, let us assume $j=1$ and denote the indices in $\mathcal{G}_1$ and $\mathcal{G}_2$ are $[1:h]$ and $[h+1:d]$ respectively. Choose an index $m\in[1:d]$. Let us consider the case $m\leq h$. The following relations hold because $\mathcal{G}_1$ and $\mathcal{G}_2$ are coherent.

$$\frac{\sqrt{w_1^2+\cdots+w_m^2}}{\sqrt{\mathcal{A}_1(m)}}<\frac{\sqrt{w_{m+1}^2+\cdots+w_h^2}}{\sqrt{\mathcal{A}_1(h)-\mathcal{A}_1(m)}},$$

$$\frac{\sqrt{w_1^2+\cdots+w_m^2}}{\sqrt{\mathcal{A}_1(m)}}<\frac{\sqrt{w_{h+1}^2+\cdots+w_d^2}}{\sqrt{\mathcal{A}_2(d-h)}}.$$

Since $\frac{\sqrt{w_{m+1}^2+\cdots+w_d^2}}{\sqrt{\mathcal{A}_2(d-h)+\mathcal{A}_1(h)-\mathcal{A}_1(m)}}$ is sandwiched between the rhs of the above two expressions, we have $\frac{\sqrt{w_1^2+\cdots+w_m^2}}{\sqrt{\mathcal{A}_1(m)}}<\frac{\sqrt{w_{m+1}^2+\cdots+w_d^2}}{\sqrt{\mathcal{A}_2(d-h)+\mathcal{A}_1(h)-\mathcal{A}_1(m)}}$. A similar reasoning as above holds for $m>h$ case also. Thus $\mathcal{G}_1\cup\mathcal{G}_2$ is coherent. □

Lemma 4.G. *Let $w\in\mathbb{R}^d$ and η_w be the minimizer of* (SOWL). *If $|w_1|\geq\cdots\geq|w_d|$, then $(\eta_w)_1\geq\cdots\geq(\eta_w)_d$.*

Proof. Let $\eta_w\in\mathcal{D}^k$ with unique values $\delta_j, j=1,\ldots,k$. W.l.o.g., let us assume that $k=2$. As a contradiction, let $\delta_2>\delta_1$ corresponding to adjacent groups $\mathcal{G}_1$ and $\mathcal{G}_2$. We denote $d_1=|\mathcal{G}_1|, d_2=|\mathcal{G}_2|$. By optimality conditions from Proposition (4.3), we have $\frac{\|w_{\mathcal{G}_2}\|_2}{\sqrt{c_1+\cdots+c_{d_2}}}>\frac{\|w_{\mathcal{G}_1}\|_2}{\sqrt{c_{d_2+1}+\cdots+c_{d_1+d_2}}}$, and $c_{d_2+1}>0$. Since $c_1>\cdots>c_d$, we have

$$\frac{|w_{d_1+1}|}{\sqrt{c_{d_2}}}\geq\frac{\|w_{\mathcal{G}_2}\|_2}{\sqrt{c_1+\cdots+c_{d_2}}}$$

$$>\frac{\|w_{\mathcal{G}_1}\|_2}{\sqrt{c_{d_2+1}+\cdots+c_{d_1+d_2}}}\geq\frac{|w_{d_1}|}{\sqrt{c_{d_2+1}}}.$$

This is impossible since $|w_{d_1}|>|w_{d_1+1}|$ and $c_{d_2}>c_{d_2+1}$, and hence the result. □

Proof of Theorem 4.4 continued. Lemma 4.G shows that the ordering of η_w follows that of $|w|$. Hence when $|w|$ is sorted, the coherent groups $\mathcal{G}_j$ in η are contiguous. The sorting of $|w|$ takes $O(d \log d)$ time. Algorithm 1 starts with coherent groups of size 1, and Lemma 4.F guarantees that the groups evolving are all coherent. Thus Algorithm 1 computes η_w which satisfy the conditions of Proposition 4.3. The outer loop is executed exactly $d-1$ times, and the inner loop is executed as many times a merge is made which is also upper bounded by $d-1$. Hence the result. □

4.4 Proof of Theorem 4.5

Proof. Before analyzing the solutions of the proximal operator of $\Omega_{\mathcal{S}}$, let us now analyze the solutions $\hat{\eta}$ obtained from solving the regression problem (5). Recalling the equivalence to submodular penalties, we denote denoting $\Gamma(\eta) = \sum_{i=1}^{d} c_i \eta_{(i)}$. Now eliminating w, we arrive at the following equivalent problem only in η.

$$\min_{\eta \geq 0} J(\eta) + \Gamma(\eta) \tag{2}$$

Where $J(\eta) = y^\top M^{-1}(\eta) y$, $M(\eta) = \left(X \mathcal{D}\left(\eta\right) X^\top + n\lambda I\right)$. The following proposition gives necessary and sufficient conditions for $\hat{\eta}$ belonging to a particular lattice $\mathcal{D}^{(k)}$.

Proposition 4.H. *Consider a lattice $\mathcal{D}^{(k)}$ along with its partition $\mathcal{G}_1, \ldots, \mathcal{G}_k$. For the problem (2), the solution $\hat{\eta} \in \mathcal{D}^k$ with unique values $\hat{\delta}_1 > \cdots > \hat{\delta}_k$ in the respective partitions if and only if the following statements hold true, where we denote by $\bar{\nu}$ the subgradient at $\hat{\delta}$ with respect to the positivity constraint $\delta \geq 0$.*

1. *$\hat{\eta}$ satisfies the order constraints for the lattice $\mathcal{D}^{(k)}$.*

2. *$\hat{\eta}$ satisfies $\sum_{i \in \mathcal{G}_j} \|x_i^\top M^{-1}(\hat{\eta}) y\|_2^2 + \bar{\nu}_j = \mathcal{A}_j(|\mathcal{G}_j|), \forall j = 1, \ldots, k$.*

3. *$s = \frac{\partial \Gamma}{\partial \hat{\eta}}$ satisfies $s(C) \leq c_1 + \cdots + c_{|C|}, \forall C \subseteq V$.*

Proof. Point 1 is the requirement of the lattice assumption. Define $H \in \mathbb{R}^{d \times k}$ with each $H_{ij} = 1$ if $i \in \mathcal{G}_j$ and 0 otherwise. The lattice assumption included in (2) leads to the following reduced problem, where we denote $t_j = \mathcal{A}_j(|\mathcal{G}_j|)$.

$$\min_{\delta \geq 0} J(H\delta) + \delta^\top t. \tag{3}$$

The first order conditions for optimality for the above equation leads to the following gradient equation, which proves the point 2.

$$\sum_{i \in A_j} \|x_i^\top M^{-1}(H\hat{\delta}) y\|_2^2 + \bar{\nu}_j = t_j, \forall j \tag{4}$$

This means at optimality, $\hat{\delta}$ solves the above non-linear system of equations. Now, from optimality conditions of (2), we have

$$\|x_i^\top M^{-1}(\hat{\eta}) y\|_2^2 + \nu_i = s_i, \tag{5}$$

with s is an element of the subgradients of Γ at $\hat{\eta}$, and ν_i the subgradient of the positivity constraint on $\hat{\eta}_i$. Invoking the properties of Lovász extensions [1], we denote by P the submodular function derived from c. For any $\eta \in \mathbb{R}^d$, we characterize the subgradient as follows.

Claim 4.I. *Let s be the subgradient of Γ at η. Then subgradient s should satisfy (a) $s \in \mathcal{B}(P)$(Base polytope) and (b) $s^\top \eta = \Gamma(\eta)$.*

Proof. Follows from the definition of the lovasz extension. □

Point (a) in the above claim is thus shown to be necessary. Whereas the following claim shows that, since $\hat{\eta}$ satisfies (5) and (4), the point (b) in the above claim is redundant.

Claim 4.J. *Let $\hat{\eta} = H\hat{\delta}$ be consistent with (4). Then in Claim 4.I, condition (a) implies the condition (b).*

Proof. For $s \in \mathcal{B}(P)$, $s^\top \eta = \Gamma(\eta)$ is satisfied as soon as $s(\mathcal{G}_1 \cup \cdots \cup \mathcal{G}_i) = P(\mathcal{G}_1 \cup \cdots \cup \mathcal{G}_i)$. Now, from (5) and (4) $s(\mathcal{G}_1 \cup \cdots \cup \mathcal{G}_i) = t_1 + \cdots + t_i = P(\mathcal{G}_1 \cup \cdots \cup \mathcal{G}_i)$. □

Now the condition $s \in \mathcal{B}(P)$ is equivalent to the condition $s(C) \leq c_1 + \cdots + c_{|C|}, \forall C \subseteq V$. This completes the proof. □

Proof of Theorem 4.5 continued.

1. When $\hat{\eta}_i^{(\mu)} > 0, \forall i$, the optimality conditions 4 translate to

$$\sum_{i \in \mathcal{G}_j} \frac{y_i^2}{\left(\hat{\delta}_j^{(\mu)} + n\mu\right)^2} = \sum_{i \in \mathcal{G}_j} \frac{y_i^2}{\left(\eta_i^{(\mu)} + n\mu\right)^2} = \mathcal{A}_j(|\mathcal{G}_j|). \tag{6}$$

This implies $\left(\hat{\delta}_j^{(\mu)} + n\mu\right)^2 = \dfrac{\sum_{i \in \mathcal{G}_j} y_i^2}{\mathcal{A}_j(|\mathcal{G}_j|)}$, (7) which is independent of μ.

$$\text{Let } C_j \subseteq \mathcal{G}_j,\ \text{then, } s(C_j) = \sum_{i\in C_j} \frac{y_i^2}{\left(\hat{\delta}_j^{(\mu)} + n\mu\right)^2} \tag{8}$$

$$= \frac{\sum_{i\in C_j} y_i^2}{\sum_{i\in \mathcal{G}_j} y_i^2} \mathcal{A}_j(|\mathcal{G}_j|). \tag{9}$$

$$\text{Then } \forall C \subseteq \{1,\dots,d\}, s(C) = \sum_{j=1}^{m} s(C \cup \mathcal{G}_j). \tag{10}$$

This implies that $s(C)$ does not depend on μ, and we arrive at the statement.

2. From the optimality condition (4), we see that

$$\hat{\delta}_j^{(\lambda)} = 0, \text{when } \sqrt{\frac{\sum_{i\in\mathcal{G}_j} y_i^2}{\mathcal{A}_j(|\mathcal{G}_j|)}} < n\lambda,$$
$$\bar{\nu}_j = t_j - \frac{\sum_{i\in\mathcal{G}_j} y_i^2}{n^2\lambda^2}.$$

The first part of the theorem guarantees that $\delta_k^{(\lambda)} = 0$, for all $k > j$. Also, the problem (3) can now be reduced in terms of the positive values of η and the first part of this theorem applies.

□

4.5 Proof of Corollary 4.6

Proof. From Theorem 4.5 it is clear that η_x can be computed in $O(d \log d)$ time. And the first order optimality condition for (5) gives

$$x_i = z_i \frac{(\eta_x)_i}{(\eta_x)_i + \lambda}.$$

Hence the result. □

5 Regularization with SOWL

5.1 Proof of Proposition 5.2

Proof. Consider the following proposition.

Proposition 5.A. *Consider $w \in \mathbb{R}^d$, and Let the minimizer of* (SOWL) *$\eta_w \in \mathcal{D}^k$. Then $\Omega_{\mathcal{S}}(w) = \sum_{j=1}^{k} \sqrt{\mathcal{A}_j(|\mathcal{G}_j|)}\|w_{\mathcal{G}_j}\|_2$.*

Proof. Once the lattice assumption is made, Γ is linear function on η. The problem (SOWL) is then separable in terms of the groups $\mathcal{G}_j$ and we get the result. □

Proof of Proposition 5.2 continued. Given $\hat{\eta}$, Corollary 5.A shows that the norm is equivalent to the group lasso penalty, and the conditions 1, 2 are the optimality conditions of the group lasso [5]. And Given $\hat{w}$, $\hat{\eta}$ is optimal if and only if it satisfied the conditions given in Proposition 4.3. The convexity of the problem (5) guarantees that $(\hat{w}, \hat{\eta})$ is indeed optimal. □

5.2 Proof of Theorem 5.3

Proof. First we note that if $\delta_k^* = 0$, then $\mathcal{G}_k = \mathcal{J}^c$.Now let us consider the restricted problem only on the support $\mathcal{J}$, and denote the minimizer as $w_{\mathcal{J}}$.

$$\min_{w_{\mathcal{J}}} \frac{1}{2}\|X_{\mathcal{J}} w_{\mathcal{J}} - y\|_2^2 + \lambda\Omega(w_{\mathcal{J}})$$

As $\lambda \to 0$, the objective in the above problem tends to $\frac{1}{2}y^\top y + \frac{1}{2}w_{\mathcal{J}}^\top \Sigma_{\mathcal{J},\mathcal{J}} w_{\mathcal{J}} - y^\top X w_{\mathcal{J}}$. It is easy to see that $w_{\mathcal{J}} \to w_{\mathcal{J}}^*$ due to the invertibility of $\Sigma_{\mathcal{J},\mathcal{J}}$. We now construct a candidate w by concatenating $w_{\mathcal{J}}$ with zeros for the remaining columns. We shall now show that the candidate solution is an optimal one. From the optimality conditions of the reduced problem, $\forall j = 1,\dots,k-1$, we have

$$X_{\mathcal{G}_j}^\top (X_{\mathcal{J}}(\hat{w}_{\mathcal{J}} - w_{\mathcal{J}}^*) + \epsilon_{\mathcal{J}}) = -\lambda n \sqrt{\mathcal{A}_j(|\mathcal{G}_j|)} \frac{\hat{w}_j}{\|\hat{w}_j\|_2}$$
$$\Rightarrow \hat{w}_{\mathcal{J}} - w_{\mathcal{J}}^* = -\lambda n (\Sigma_{\mathcal{J},\mathcal{J}})^{-1} \mathrm{D}\left(\frac{\sqrt{\mathcal{A}_j(|\mathcal{G}_j|)}}{\|\hat{w}_{\mathcal{G}_j}\|_2}\right) \hat{w}_{\mathcal{J}}$$
$$+ O_p(n^{-\frac{1}{2}}).$$

Now,

$$X_{\mathcal{J}^c}^\top y - X_{\mathcal{J}^c}^\top X_{\mathcal{J}} \hat{w}_{\mathcal{J}} = \Sigma_{\mathcal{J}^c,\mathcal{J}}(w_{\mathcal{J}}^* - \hat{w}_{\mathcal{J}}) + O_p(n^{-\frac{1}{2}})$$
$$= \lambda n \Sigma_{\mathcal{J}^c,\mathcal{J}} \left(\Sigma_{\mathcal{J}\mathcal{J}}\right)^{-1} \mathrm{D}\left(\frac{\sqrt{\mathcal{A}_j(|\mathcal{G}_j|)}}{\|\hat{w}_{\mathcal{G}_j}\|_2}\right) \hat{w}_S + O_p(n^{-\frac{1}{2}})$$
$$\Rightarrow \|X_{\mathcal{J}^c}^\top (X_{\mathcal{J}} \hat{w}_{\mathcal{J}} - y)\|_2 \leq \lambda n,$$

where the last line used the irrepresentability condition. Thus $\hat{w}$ converges to w^* in probability and proves the theorem. □

5.3 Proof of Theorem 5.1

Proof. Let us denote by $\Gamma(\eta) = \sum_{i=1}^{d} c_i \eta_{(i)}$. The problem (5) is jointly convex in w, η and first order conditions are neccessary and sufficient. They are

$$x_i^\top (X\hat{w} - y) + \lambda \frac{\hat{w}_i}{\hat{\eta}_i} = 0 \tag{11}$$

$$-\frac{\hat{w}_i^2}{\hat{\eta}_i^2} + s_i + h_i = 0, \tag{12}$$

where s_i denotes the subgradient of Γ with respect to η_i, and h_i the subgradient with respect to the positivity condition on η_i. Obviously, when $\hat{\eta}_i > 0$, $h_i = 0$.

When $\hat{\eta}_i = 0$, we see that from (11) that $\hat{w}_i = 0$. Substituting this in (12), $h_i = -s_i$.

Let us assume $\hat{\eta}_i(\lambda) \neq \hat{\eta}_j(\lambda)$ and different from the rest. This implies that $s_i = c_i$. This leads to the following equations which follows from (11) and (12).

$$\begin{aligned}
\frac{\hat{w}_i}{\hat{\eta}_i} &= -\frac{1}{\lambda} x_i^\top \left(X\hat{w} - y\right) \\
c_i = \frac{\hat{w}_i^2}{\hat{\eta}_i^2} &= \frac{1}{\lambda^2} \left(X\hat{w} - y\right)^\top x_i x_i^\top \left(X\hat{w} - y\right) \\
|c_i - c_j| &= \frac{1}{\lambda^2} \left| \left(X\hat{w} - y\right)^\top \left(x_i x_i^\top - x_j x_j^\top\right) \left(X\hat{w} - y\right) \right| \\
&\leq \frac{1}{\lambda^2} \|X\hat{w} - y\|_2^2 \|x_i x_i^\top - x_j x_j^\top\|_2 \\
&\leq \frac{1}{\lambda^2} \|y\|_2^2 \; \|x_i + x_j\|_2 \; \|x_i - x_j\|_2 \\
&\leq \frac{1}{\lambda^2} \|y\|_2^2 \sqrt{2 + 2\rho_{ij}} \sqrt{2 - 2\rho_{ij}} \\
&= \frac{1}{\lambda^2} \|y\|_2^2 \sqrt{4 - 4\rho_{ij}^2}
\end{aligned}$$

We know that $|c_i - c_j| \geq C > 0$, where $C = \min_{k<d}(c_k - c_{k+1})$. This implies that

$$0 < C \leq |c_i - c_j| \leq \frac{1}{\lambda^2} \|y\|_2^2 \sqrt{4 - 4\rho_{ij}^2}$$

This is impossible to happen for all $\lambda > 0$ and leads to a contradiction for the assumption that $\hat{\eta}_i(\lambda) \neq \hat{\eta}_j(\lambda)$. We define $\lambda_0 = \frac{\|y\|_2}{\sqrt{C}} (4 - 4\rho_{ij}^2)^{\frac{1}{4}}$ and we have $\hat{\eta}_i(\lambda) = \hat{\eta}_j(\lambda)$ for all $\lambda > \lambda_0$. □

5.4 Proof of Theorem 5.4

Proof. The proof proceeds by proving the following lemmas. In this section we denote by Let $A = \text{supp}(\eta)$.

Lemma 5.B. *Consider the prob* (6)*. Let* $|A| = r$*. Then the following statements hold true.*

$$\frac{z_r^2}{c_r} > \lambda^2, \frac{z_{r+1}^2}{c_{r+1}} \leq \lambda^2. \tag{13}$$

Proof. Consider any $j < r$, and perturb η_i for $i = j, \ldots, r$ as $\tilde{\eta}_i = \eta_i - h$ for a fixed $h > 0$. Then the difference of the objective in (6) for η and $\tilde{\eta}$ is given as

$$\sum_{i=j}^{r} z_i^2 \left(\frac{1}{\eta_i - h + \lambda} - \frac{1}{\eta_i + \lambda} \right) - c_i h > 0,$$

because of the optimality of η. Rearranging the terms we have,

$$\sum_{i=j}^{r} \frac{z_i^2}{(\eta_i + \lambda)^2} > \sum_{i=j}^{r} c_i.$$

for $j = r$, we get the first condition. The proof of the next condition is similar by perturbing the η_i for $i > r$. □

Lemma 5.C. *Consider the following modified problem of* (6)*, where* $\tilde{z} = [z_2, \ldots, z_d]$*,* $\tilde{c} = [c_2, \ldots, c_d]$*.*

$$\min_{\tilde{\eta} \geq 0} \sum_{i=1}^{d-1} \left(\frac{\tilde{z}_i^2}{\tilde{\eta}_i + \lambda} + \tilde{c}_i \tilde{\eta}_{(i)} \right) \tag{14}$$

If problem (6) *has exactly* r *non-zero values in* η *at optimality, then problem* (14) *has exactly* $r - 1$ *non-zero entries in* $\tilde{\eta}$*.*

Proof. Trivial. □

Lemma 5.D. *Let* H_i *be the hypothesis that* $\eta_i = 0$*. Then* $\{z | H_i$ *is rejected and* $|A| = r\} = \{z | z_i^2 > \lambda c_r^2\}$*.*

Proof of Theorem 5.4 continued. We now choose $\sqrt{\lambda c_r} = \Phi^{-1}\left(1 - \frac{iq}{2d}\right)$, which leads to the following bound.

$$\begin{aligned}
P(H_i \text{ rejected and } |A| = r) \leq & P(z_i^2 > \lambda c_r \text{ and } |\tilde{A}| = r - 1) \\
\leq & P(z_i^2 > \lambda c_r) P(|\tilde{A}| = r - 1) \\
\leq & P(|z_i| > \sqrt{\lambda c_r}) P(|\tilde{A}| = r - 1) \\
\leq & \frac{qr}{d} P(|\tilde{A}| = r - 1)
\end{aligned}$$

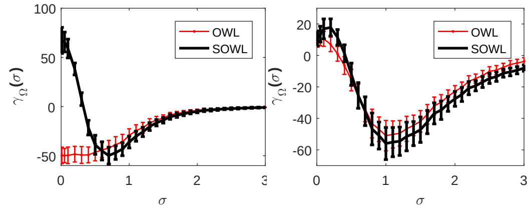
Hence the FDR is given as

$$\begin{aligned}
FDR = & \sum_{r=1}^{d} \frac{1}{r} \sum_{i=1}^{d_0} P(H_i \text{ rejected and } |A| = r) \\
\leq & \sum_{r=1}^{d} \frac{q d_0}{d} P(|\tilde{S}| = r - 1) = \frac{q d_0}{d}.
\end{aligned}$$

□

6 Additional Plots

We show in Figure 1 the proximal denoising plots given in the main paper with error bars.



(a) (b)

Figure 1: Proximal denoising plots (a), and (b) refer to examples in Figures 3a, , 3b respectively.

References

[1] F. Bach. Learning with submodular functions: A convex optimization perspective. *Foundations and Trends in Machine Learning*, 6-2-3:145–373, 2011.

[2] G. Obozinski and F. Bach. Convex relaxation for combinatorial penalties. Technical Report 00694765, HAL, 2012.

[3] F. Bach. Shaping level sets with submodular functions. In *Advances in Neural Information Processing Systems (NIPS)*, 2011.

[4] X. Zeng and M. Figueiredo. The ordered weighted ℓ_1 norm: Atomic formulation, projections, and algorithms. *ArXiv preprint:1409.4271v5*, 2015.

[5] F. Bach. Consistency of the group lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9:1179–1225, 2008.

‘