Multi-stage convex relaxation for feature selection

TONG ZHANG

Hill Center, Rutgers University, Piscataway, NJ 08854, USA. E-mail: tzhang@stat.rutgers.edu

A number of recent work studied the effectiveness of feature selection using Lasso. It is known that under the restricted isometry properties (RIP), Lasso does not generally lead to the exact recovery of the set of nonzero coefficients, due to the looseness of convex relaxation. This paper considers the feature selection property of nonconvex regularization, where the solution is given by a multi-stage convex relaxation scheme. The nonconvex regularizer requires two tuning parameters (compared to one tuning parameter for Lasso). Although the method is more complex than Lasso, we show that under appropriate conditions including the dependence of a tuning parameter on the support set size, the local solution obtained by this procedure recovers the set of nonzero coefficients without suffering from the bias of Lasso relaxation, which complements parameter estimation results of this procedure in (*J. Mach. Learn. Res.* **11** (2011) 1087–1107).

Keywords: multi-stage convex relaxation; non-convex penalty; variable selection

1. Introduction

We consider the linear regression problem, where we observe a set of input vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n \in R^p$, with corresponding desired output variables $y_1, \ldots, y_n$. In a statistical linear model, it is common to assume that there exists a target coefficient vector $\bar{\mathbf{w}} \in R^p$ such that

$$y_i = \bar{\mathbf{w}}^\top \mathbf{x}_i + \varepsilon_i \qquad (i = 1, \ldots, n), \tag{1}$$

where ε_i are zero-mean independent random noises (but not necessarily identically distributed). Moreover, we assume that the target vector $\bar{\mathbf{w}}$ is sparse. That is, $\bar{k} = \|\bar{\mathbf{w}}\|_0$ is small. Here we use the standard notation

$$\mathrm{supp}(\mathbf{w}) = \{j: \mathbf{w}_j \neq 0\} \qquad \|\mathbf{w}\|_0 = \big|\mathrm{supp}(\mathbf{w})\big|$$

for any vector $\mathbf{w} \in R^p$.

This paper focuses on the feature selection problem, where we are interested in estimating the set of nonzero coefficients $\mathrm{supp}(\bar{\mathbf{w}})$ (also called support set). Let $\mathbf{y}$ denote the vector of $[y_i]$ and X be the $n \times p$ matrix with each row a vector $\mathbf{x}_i$. The standard statistical method is subset selection (L_0 regularization), which computes the following estimator

$$\hat{\mathbf{w}}_{L_0} = \arg\min_{\mathbf{w} \in R^p} \|X\mathbf{w} - \mathbf{y}\|_2^2 \qquad \text{subject to } \|\mathbf{w}\|_0 \leq k, \tag{2}$$

where k is a tuning parameter. This method is arguably a natural method for feature selection because if noise ε_i are i.i.d. Gaussian random variables, then (2) can be regarded as a Bayes

procedure with an appropriately defined sparse prior over $\mathbf{w}$. However, because the optimization problem in (2) is non-convex, the global solution of this problem cannot be efficiently computed. In practice, one can only find an approximate solution of (2). The most popular approximation to L_0 regularization is the L_1 regularization method which is often referred to as Lasso [9]:

$$\hat{\mathbf{w}}_{L_1} = \arg \min_{\mathbf{w} \in R^p} \left[\frac{1}{n} \|X\mathbf{w} - \mathbf{y}\|_2^2 + \lambda \|\mathbf{w}\|_1 \right], \tag{3}$$

where $\lambda > 0$ is an appropriately chosen regularization parameter.

The global optimum of (3) can be easily computed using standard convex programming techniques. It is known that in practice, L_1 regularization often leads to sparse solutions (although often suboptimal). Moreover, its performance has been theoretically analyzed recently. For example, it is known from the compressed sensing literature (e.g., [3]) that under certain conditions referred to as *restricted isometry property* (RIP), the solution of L_1 relaxation (3) approximates the solution of the L_0 regularization problem (2). The prediction and parameter performance of this method has been considered in [1,2,6,10,14,15]. Exact support recovery was considered by various authors such as [7,8,11,18]. It is known that under some more restrictive conditions referred to as *irrepresentable conditions*, L_1 regularization can achieve exact recovery of the support set. However, the L_1 regularization method (3) does not achieve exact recovery of the support set under the RIP type of conditions, which we are interested in here.

Although it is possible to achieve exact recovery using post-processing by thresholding the small coefficients of Lasso solution, this method is suboptimal under RIP in comparison to the L_0 regularization method (2) because it requires the smallest nonzero coefficients to be $\sqrt{\bar{k}}$ times larger than the noise level instead of only requiring the nonzero coefficients to be larger than the noise level with L_0 regularization in (2). This issue, referred to as the *bias* of Lasso for feature selection, was extensively discussed in [13]. Detailed discussion can be found after Theorem 1. It is worth mentioning that under a stronger mutual coherence condition (similar to irrepresentable condition), this post-processing step does not give this bias factor $\sqrt{\bar{k}}$ as shown in [7] (also see [15]). Therefore, the advantage of bias removal for the multi-stage procedure discussed here is only applicable when RIP holds but when the irrepresentable condition and mutual incoherence conditions fail. A thorough discussion of various conditions is beyond the scope of the current paper, and we would like to refer the readers to [10]. Nevertheless, it is worth pointing out that even in the classical $p < n$ setting with the design matrix X being rank p, the irrepresentable condition or the mutual incoherence condition can still be violated while the RIP type sparse-eigenvalue condition used in this paper holds trivially. In fact, this was pointed out in [19] as the main motivation of adaptive Lasso. Adaptive Lasso behaves similarly to the above mentioned post-processing, and thus suffers from the same bias problem.

The bias of Lasso is due to the looseness of convex relaxation for L_0 regularization. Therefore, the remedy is to use a non-convex regularizer that is close to L_0 regularization. One drawback of using nonconvex optimization formulation is that we can only find a local optimal solution and different computational procedure may lead to a different local solution. Therefore, the theoretical analysis has to be integrated with specific computational procedure to show that the local minimum obtained by the procedure has desirable properties (e.g., exact support recovery). Several nonconvex computational procedures have been analyzed in the literature, including an

adaptive forward backward greedy procedure (referred to as FoBa) to approximately solve the regularization method (2) considered in [17], and the MC+ method in [13] to solve a non-convex regularized problem using a path-following procedure. Both methods can achieve unbiased feature selection.

Related to the above mentioned work, a different procedure, referred to as *multi-stage convex relaxation*, was analyzed in [16]. This procedure solves a nonconvex problem using multiple stages of Lasso relaxations, where convex formulations are iteratively refined based on solutions obtained from the previous stages. However, only parameter estimation performance was analyzed in [16]. Unfortunately, the result in [16] does not directly imply that multi-stage convex relaxation achieves unbiased recovery of the support set. The purpose of this paper is to prove such a support recovery result analogous to related result in [13] (which is for a different procedure), and this result complements the parameter estimation result of [16].

2. Multi-stage convex relaxation with capped-L_1 regularization

We are interested in recovering $\bar{\mathbf{w}}$ from noisy observations $\mathbf{y}$ using the following nonconvex regularization formulation:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \left[\frac{1}{n} \| X\mathbf{w} - \mathbf{y} \|_2^2 + \lambda \sum_{j=1}^{p} g\big(|\mathbf{w}_j|\big) \right], \tag{4}$$

where $g(|\mathbf{w}_j|)$ is a regularization function. For simplicity, this paper only considers the specific regularizer

$$g(u) = \min(u, \theta), \tag{5}$$

which is referred to as capped-L_1 regularization in [16]. The parameter θ is a thresholding parameter which says that we use L_1 penalization when a coefficient is sufficiently small, but the penalty does not increase when the coefficient is larger than a threshold θ. Detailed discussions can be found in [16]. Similar to [16], one can analyze general regularization function $g(u)$. However, some of such functions (such as adaptive Lasso) do not completely remove the bias. Therefore we only analyze the simple function (5) in this paper for clarity. While a theoretical justification has been given in [16] for multi-stage convex relaxation, similar procedure has been shown to work well empirically without theoretical justification [4,12]. Moreover, a two-stage version was proposed in [20], which does not remove the bias issue discussed in this paper.

Since the regularizer (5) is not convex, the resulting optimization problem (4) is a nonconvex regularization problem. However the regularizer in (5) is continuous and piecewise differentiable, and thus its solution is easier to compute than the L_0 regularization method in (2). For example, standard numerical techniques such as sub-gradient descent lead to local minimum solutions. Unfortunately, it is difficult to find the global optimum, and it is also difficult to analyze the quality of the local minimum obtained from the gradient descent method. As a matter of fact, results with non-convex regularization are difficult to reproduce because different numerical optimization procedures can lead to different local minima. Therefore, the quality of the solution heavily depend on the numerical procedure used.

Initialize $\lambda_j^{(0)} = \lambda$ for $j = 1, \ldots, d$.

For $\ell = 1, 2, \ldots$

- Let

$$\hat{\mathbf{w}}^{(\ell)} = \arg\min_{\mathbf{w} \in R^p} \left[\frac{1}{n} \|X\mathbf{w} - \mathbf{y}\|_2^2 + \sum_{j=1}^{p} \lambda_j^{(\ell-1)} |\mathbf{w}_j| \right]. \tag{6}$$

- Let $\lambda_j^{(\ell)} = \lambda I(|\hat{\mathbf{w}}_j^{(\ell)}| \leq \theta)$ $(j = 1, \ldots, d)$.

Figure 1. Multi-stage convex relaxation for sparse regularization.

In the following, we consider a specific numerical procedure referred to as multi-stage convex relaxation in [16]. The algorithm is given in Figure 1. The procedure converges to a local optimal solution of (4) due to a simple concave duality argument, where (4) is rewritten as

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \min_{\{\lambda_j \geq 0\}} \left[\frac{1}{n} \|X\mathbf{w} - \mathbf{y}\|_2^2 + \sum_{j=1}^{p} \lambda_j |\mathbf{w}_j| + \sum_{j=1}^{p} g^*(\lambda_j) \right],$$

with $g^*(\lambda_j) = \max((\lambda - \lambda_j)\theta, 0)$. The procedure of Figure 1 can be regarded as an alternating optimization method to solve this joint optimization problem of $\mathbf{w}$ and $\{\lambda_j\}$, where the first step solves for $\mathbf{w}$ with $\{\lambda_j\}$ fixed, and the second step is the closed form solution of $\{\lambda_j\}$ with $\mathbf{w}$ fixed. A more detailed discussion can be found in [16]. Our goal is to show that this procedure can achieve unbiased feature selection as described in [13].

3. Theoretical analysis

We require some technical conditions for our analysis. First, we assume sub-Gaussian noise as follows.

Assumption 1. *Assume that* $\{\varepsilon_i\}_{i=1,\ldots,n}$ *in* (1) *are independent* (*but not necessarily identically distributed*) *sub-Gaussians*: *there exists* $\sigma \geq 0$ *such that* $\forall i$ *and* $\forall t \in R$,

$$\mathbf{E}_{\varepsilon_i} e^{t\varepsilon_i} \leq e^{\sigma^2 t^2/2}.$$

Both Gaussian and bounded random variables are sub-Gaussian using the above definition. For example, if a random variable $\xi \in [a, b]$, then $\mathbf{E}_\xi e^{t(\xi - \mathbf{E}\xi)} \leq e^{(b-a)^2 t^2/8}$. If a random variable is Gaussian: $\xi \sim N(0, \sigma^2)$, then $\mathbf{E}_\xi e^{t\xi} \leq e^{\sigma^2 t^2/2}$.

We also introduce the concept of sparse eigenvalue, which is standard in the analysis of L_1 regularization.

Definition 1. *Given k, define*

$$\rho_+(k) = \sup\left\{\frac{1}{n}\|X\mathbf{w}\|_2^2/\|\mathbf{w}\|_2^2:\ \|\mathbf{w}\|_0 \le k\right\},$$

$$\rho_-(k) = \inf\left\{\frac{1}{n}\|X\mathbf{w}\|_2^2/\|\mathbf{w}\|_2^2:\ \|\mathbf{w}\|_0 \le k\right\}.$$

The following result for parameter estimation was obtained in [16], under the Assumption 1. If we assume that the target $\bar{\mathbf{w}}$ is sparse, with $\mathbf{E}y_i = \bar{\mathbf{w}}^\top \mathbf{x}_i$, and $\bar{k} = \|\bar{\mathbf{w}}\|_0$, and we choose θ and λ such that

$$\lambda \ge 20\sigma\sqrt{2\rho_+(1)\ln(2p/\eta)/n}$$

and

$$\theta \ge 9\lambda/\rho_-(2\bar{k}+s).$$

Assume that $\rho_+(s)/\rho_-(2\bar{k}+2s) \le 1 + 0.5s/\bar{k}$ for some $s \ge 2\bar{k}$, then with probability larger than $1 - \eta$:

$$
\begin{aligned}
\left\|\hat{\mathbf{w}}^{(\ell)} - \bar{\mathbf{w}}\right\|_2 &\le \frac{17}{\rho_-(2\bar{k}+s)}\left[2\sigma\sqrt{\rho_+(\bar{k})}\left(\sqrt{\frac{7.4\bar{k}}{n}} + \sqrt{\frac{2.7\ln(2/\eta)}{n}}\right) + \lambda\sqrt{k_\theta}\right] \\
&\quad + \frac{0.7^\ell \cdot \sqrt{\bar{k}}\lambda}{\rho_-(2\bar{k}+s)},
\end{aligned}
\tag{7}
$$

where $\hat{\mathbf{w}}^{(\ell)}$ is the solution of (6), and $k_\theta = |\{j \in \bar{F}: |\bar{\mathbf{w}}_j| \le 2\theta\}|$. Here $\bar{F} = \mathrm{supp}(\bar{\mathbf{w}})$.

The condition $\rho_+(s)/\rho_-(2\bar{k}+2s) \le 1 + 0.5s/\bar{k}$ requires the eigenvalue ratio $\rho_+(s)/\rho_-(s)$ to grow sub-linearly in s. Such a condition, referred to as *sparse eigenvalue condition*, is also needed in the standard analysis of L_1 regularization [14,15]. As shown in [14,15], it is slightly weaker than the RIP condition in compressive sensing [3], which requires that for some $\delta_{s'} \in (0, 1)$ and $s' > 2\bar{k}$, the condition

$$1 - \delta_{s'} \le \frac{1}{n}\|X\mathbf{w}\|_2^2/\|\mathbf{w}\|_2^2 \le 1 + \delta_{s'}$$

holds for all $\mathbf{w}$ such that $\|\mathbf{w}\|_0 \le s'$. In other words, RIP defined in [3] requires that $1 - \delta_{s'} \le \rho_-(s') \le \rho_+(s') \le 1 + \delta_{s'}$ for some $s' > 2\bar{k}$. For example, with $s' = 6\bar{k}$, and the restricted isometry constant $\delta_{s'} \le 1/3$, then the sparse eigenvalue condition considered in this paper holds with $s = 2\bar{k}$. Note that in the traditional low-dimensional statistical analysis, one assumes that $\rho_+(s)/\rho_-(2\bar{k}+2s) < \infty$ as $s \to \infty$, which is significantly stronger than the condition we use here. Although in practice it is often difficult to verify the sparse eigenvalue condition for real problems, the parameter estimation result in (7) nevertheless provides important theoretical insights for multi-stage convex relaxation.

In the literature, one is often interested in two types of results, one is parameter estimation bound as in (7), and the other is feature selection consistency: that is, to identify the set of

nonzero coefficients of the truth. Although the parameter estimation bound in (7) is superior to Lasso under certain conditions, the result does not imply that one can correctly select all variables under this condition. Moreover, the specific proof presented in [16] does not directly imply such a result. Therefore, it is important to know whether the multi-stage convex relaxation can achieve unbiased feature selection as studied in [13]. In the following, we present such a result which supplements the parameter estimation bound of (7). While the main high-level argument follows that of [16], there are many differences in the details, and hence a full proof (which is included in Section 5) is still needed. This theorem is the main result of the paper.

It is worth mentioning that although we only consider the simple capped-L_1 regularizer, similar results can be obtained for other regularizers (with virtually the same proof) such that $g'(u) \in [0, \infty)$, $g'(u) > 0$ when u belongs to a neighbor of 0, and $g'(u) = 0$ when $u \geq \theta$, with a threshold $\theta > 0$ appropriately chosen at the order of the noise level – the condition of $g'(u) = 0$ when $u \geq \theta$ ensures the removal of feature selection "bias" of Lasso which we discussed above. As an example, very similar result can be obtained for the MC+ penalty of [13] or SCAD penalty of [5] using the multi-stage convex relaxation procedure here. In fact, in practice there may be additional advantages of using a smooth nonconvex penalty such as MC+ due to the extra smoothness, although such advantage is not revealed in our theoretical analysis.

It should also be mentioned that the multistage convex relaxation method requires two tuning parameters λ and θ (generally we do not regard ℓ as a tuning parameter, as we may take $\ell \to \infty$ or sufficiently large), while Lasso requires only one tuning parameter λ. The extra parameter θ may be used to incorporate the knowledge of the support set size $\bar{k}$, which can be regarded as an explanation of why this method can outperform Lasso in sparse recovery under suitable conditions. However, a disadvantage of the procedure is the extra tuning needed for this method to perform optimally.

Theorem 1. *Let Assumption 1 hold. Assume also that the target $\bar{\mathbf{w}}$ is sparse, with $\mathbf{E} y_i = \bar{\mathbf{w}}^\top \mathbf{x}_i$, and $\bar{k} = \|\bar{\mathbf{w}}\|_0$. Let $\bar{F} = \mathrm{supp}(\bar{\mathbf{w}})$. Choose θ and λ such that*

$$\lambda \geq 7\sigma\sqrt{2\rho_+(1)\ln(2p/\eta)/n}$$

and

$$\theta > 9\lambda/\rho_-(1.5\bar{k} + s).$$

Assume that

$$\min_{j \in \bar{F}} |\bar{\mathbf{w}}_j| > 2\theta$$

and $\rho_+(s)/\rho_-(1.5\bar{k} + 2s) \leq 1 + 2s/(3\bar{k})$ for some $s \geq 1.5\bar{k}$, then with probability larger than $1 - \eta$:

$$\mathrm{supp}\big(\hat{\mathbf{w}}^{(\ell)}\big) = \mathrm{supp}(\bar{\mathbf{w}})$$

when $\ell > L$, where $\hat{\mathbf{w}}^{(\ell)}$ is the solution of (6) and

$$L = \left\lfloor \frac{0.5\ln\bar{k}}{\ln(\rho_-(1.5\bar{k} + s)\theta/(6\lambda))} \right\rfloor + 1.$$

Theorem 1 is the main result of this paper. If

$$\min_{\mathbf{w}_j \in \bar{F}} |\mathbf{w}_j| \geq c\sigma\sqrt{\ln p/n} \tag{8}$$

for a sufficiently large constant c that is independent of $\bar{k}$ (but could depend on the RIP condition), then we can pick both parameters $\lambda = O(\sigma\sqrt{\ln p/n})$ and $\theta = O(\sigma\sqrt{\ln p/n})$ at the noise level, so that Theorem 1 can be applied. In this case, Theorem 1 implies that multi-stage capped-L_1 regularization achieves exact recovery of the support set supp($\bar{\mathbf{w}}$). In comparison, Lasso does not achieve exact sparse recovery under RIP conditions. While running Lasso followed by thresholding small coefficients to zero (or using adaptive Lasso of [19] or the two-stage procedure of [20]) may achieve exact recovery, such a procedure requires the condition that

$$\min_{\mathbf{w}_j \in \bar{F}} |\mathbf{w}_j| \geq c'\sigma\sqrt{\bar{k}\ln p/n}$$

for some constant c' (also depends on the RIP condition). This extra $\sqrt{\bar{k}}$ factor is referred to as the bias of the Lasso procedure in [13]. Moreover, it is known that for exact recovery to hold, the requirement of $\min_{\mathbf{w}_j \in \bar{F}} |\mathbf{w}_j| \geq c\sigma\sqrt{\ln p/n}$ (up to a constant) is necessary for all statistical procedures, in the sense that if $\min_{\mathbf{w}_j \in \bar{F}} |\mathbf{w}_j| \leq c'\sigma\sqrt{\ln p/n}$ for a sufficiently small constant c' (under appropriate RIP conditions), then no statistical procedure can achieve exact recovery with large probability. Therefore, statistical procedures that can achieve exact support recovery under (8) are referred to as (nearly) unbiased feature selection methods in [13]. Theorem 1 shows that multi-stage convex relaxation with capped-L_1 regularization achieves unbiased feature selection.

Results most comparable to what we have obtained here are that of the FoBa procedure in [17] and that of the MC+ procedure in [13]. Both can be regarded as (approximate) optimization methods for nonconvex formulations. The former is a forward backward greedy algorithm, which does not optimize (4), while the latter is a path-following algorithm for solving formulations similar to (4). Although results in [13] are comparable to ours, we should note that unlike our procedure, which is efficient due to the finite number of convex optimization, there is no proof showing that the path-following strategy in [13] is always efficient (in the sense that there may be exponentially many switching points).

4. Simulation study

Numerical examples can be found in [16] that demonstrate the advantage of multi-stage convex relaxation over Lasso. Therefore, we shall not repeat a comprehensive study. Nevertheless, this section presents a simple simulation study to illustrate the theoretical results. The $n \times p$ design matrix X is generated with i.i.d. random Gaussian entries and each column is normalized with 2-norm $\sqrt{n}$. Here $n = 100$ and $p = 250$. We then generate a vector $\bar{\mathbf{w}}$ with $\bar{k} = 30$ nonzero coefficients, and each nonzero coefficient is uniformly generated from the interval $(1, 10)$. The observation is $y = X\bar{\mathbf{w}} + \varepsilon$, where ε is zero-mean i.i.d. Gaussian noise with standard deviation $\sigma = 1$. We study the feature selection performance of multi-stage convex relaxation method

Table 1. Probability of exact support recovery for multi-stage convex relaxation

	λ					
	0.23	0.47	0.94	1.9	3.8	7.5
$\theta = 0.5\lambda$						
$\ell = 1$	0	0	0	0	0	0
$\ell = 2$	0	0	0.02	0	0	0
$\ell = 4$	0	0.05	0.63	0.18	0	0
$\ell = 8$	0	0.12	0.83	0.25	0	0
$\theta = 1\lambda$						
$\ell = 1$	0	0	0	0	0	0
$\ell = 2$	0	0.04	0.15	0.06	0	0
$\ell = 4$	0	0.33	0.86	0.13	0	0
$\ell = 8$	0	0.38	0.93	0.16	0	0
$\theta = 2\lambda$						
$\ell = 1$	0	0	0	0	0	0
$\ell = 2$	0	0.14	0.22	0	0	0
$\ell = 4$	0	0.29	0.6	0.02	0	0
$\ell = 8$	0	0.3	0.62	0.02	0	0
$\theta = 4\lambda$						
$\ell = 1$	0	0	0	0	0	0
$\ell = 2$	0	0.01	0.01	0	0	0
$\ell = 4$	0	0.06	0.06	0	0	0
$\ell = 8$	0	0.06	0.06	0	0	0

in Figure 1 using various configurations of $\lambda = \tau \sigma \sqrt{\ln(p)/n}$ (with $\tau = 1, 2, 4, 8, 16, 32$), and $\theta = \mu\lambda$ for various constants $\mu = 0.5, 1, 2, 4$.

The experiments are repeated for 100 times, and Table 1 reports the probability (percentage in the 100 runs) of exact support recovery for each configuration at various stages ℓ. Note that $\ell = 1$ corresponds to Lasso and $\ell = 2$ is an adaptive Lasso like two stage method [19,20]. The main purpose of this study is to illustrate that it is beneficial to use more than two stages, as predicted by our theory. However, since only $O(\ln(\bar{k}))$ is sufficient, optimal results can be achieved with relatively small number of stages. These conclusions can be clearly seen from Table 1. Specifically the results for $\ell = 2$ are better than those of $\ell = 1$ (standard Lasso), while results of $\ell = 4$ are better than those of $\ell = 2$. Although the performance of $\ell = 8$ is even better, the improve over $\ell = 4$ is small at the optimal configuration of λ and θ. This is consistent with our theory, which implies that a relatively small number of stages is needed to achieve good performance.

5. Proof of Theorem 1

The analysis is an adaptation of [16]. While the main proof structure is similar, there are nevertheless subtle and important differences in the details, and hence a complete proof is still necessary.

The main technical differences are as follows. The proof of [16] tracks the progress from one stage $\ell - 1$ to the next stage ℓ using a bound on 2-norm parameter estimate, while in the current proof we track the progress using the set of variables that differ significantly from the true variables. Moreover, in [16], we compare the current estimated parameter to the true parameter $\bar{\mathbf{w}}$, which is sufficient for parameter estimation. However, in order to establish feature selection result of this paper, it is necessary to compare the current estimated parameter to the least squares solution $\tilde{\mathbf{w}}$ within the true feature set $\bar{F}$ as defined below in (9). These subtle technical differences mean that many details in the proofs presented below differ from that of [16].

5.1. Auxiliary lemmas

We first introduce some definitions. Consider the positive semi-definite matrix $A = n^{-1} X^\top X \in \mathbb{R}^{p \times p}$. Given $s, k \geq 1$ such that $s + k \leq p$. Let I, J be disjoint subsets of $\{1, \ldots, p\}$ with k and s elements, respectively. Let $A_{I,I} \in R^{k \times k}$ be the restriction of A to indices I, $A_{I,J} \in R^{k \times s}$ be the restriction of A to indices I on the left and J on the right. Similarly, we define restriction $\mathbf{w}_I$ of a vector $\mathbf{w} \in R^p$ on I; and for convenience, we allow either $\mathbf{w}_I \in R^k$ or $\mathbf{w}_I \in R^p$ (where components not in I are zeros) depending on the context.

We also need the following quantity in our analysis:

$$\pi(k, s) = \sup_{\mathbf{v} \in R^k, \mathbf{u} \in R^s, I, J} \frac{\mathbf{v}^\top A_{I,J} \mathbf{u} \|\mathbf{v}\|_2}{\mathbf{v}^\top A_{I,I} \mathbf{v} \|\mathbf{u}\|_\infty}.$$

The following two lemmas are taken from [15]. We skip the proof.

Lemma 1. *The following inequality holds*:

$$\pi(k, s) \leq \frac{s^{1/2}}{2} \sqrt{\rho_+(s) / \rho_-(k + s) - 1}.$$

Lemma 2. *Consider $k, s > 0$ and $G \subset \{1, \ldots, d\}$ such that $|G^c| = k$. Given any $\mathbf{w} \in R^p$. Let J be the indices of the s largest components of $\mathbf{w}_G$ (in absolute values), and $I = G^c \cup J$. Then*

$$\max\left(0, \mathbf{w}_I^\top A \mathbf{w}\right) \geq \rho_-(k + s)\left(\|\mathbf{w}_I\|_2 - \pi(k + s, s) \|\mathbf{w}_G\|_1 / s\right) \|\mathbf{w}_I\|_2.$$

Our analysis requires us to keep track of progress with respect to the least squares solution $\tilde{\mathbf{w}}$ with the true feature set $\bar{F}$, which we define below:

$$\tilde{\mathbf{w}} = \arg \min_{\mathbf{w} \in R^p} \|X\mathbf{w} - \mathbf{y}\|_2^2 \qquad \text{subject to } \operatorname{supp}(\mathbf{w}) \subset \bar{F}, \tag{9}$$

where $\bar{F} = \operatorname{supp}(\bar{\mathbf{w}})$.

The following lemmas require varying degrees of modifications from similar lemmas in [16], and thus the proofs are included for completeness.

Lemma 3. *Define $\hat{\varepsilon} = \frac{1}{n} X^\top (X \tilde{\mathbf{w}} - \mathbf{y})$. Under the conditions of Assumption* 1, *with probability larger than $1 - \eta$:*

$$\forall j \in \bar{F}: \qquad |\hat{\varepsilon}_j| = 0, \qquad |\tilde{\mathbf{w}}_j - \bar{\mathbf{w}}_j| \le \sigma \sqrt{2\rho_-(\bar{k})^{-1} \ln(2p/\eta)/n},$$

and

$$\forall j \notin \bar{F}: \qquad |\hat{\varepsilon}_j| \le \sigma \sqrt{2\rho_+(1) \ln(2p/\eta)/n}.$$

Proof. Let $\tilde{P}$ be the projection matrix to the subspace spanned by columns of X in $\bar{F}$, then we know that

$$X\tilde{\mathbf{w}} = \tilde{P}\mathbf{y}$$

and

$$(I - \tilde{P})\mathbf{E}\mathbf{y} = \mathbf{E}\mathbf{y} - X\bar{\mathbf{w}} = 0.$$

Therefore for each j

$$n|\hat{\varepsilon}_j| = \left| X_j^\top (X\tilde{\mathbf{w}} - \mathbf{y}) \right| = \left| X_j^\top (I - \tilde{P})(\mathbf{y} - \mathbf{E}\mathbf{y}) \right|.$$

It implies that $\hat{\varepsilon}_j = 0$ if $j \in \bar{F}$. Since for each j: the column X_j satisfies $\|X_j^\top (I - \tilde{P})\|_2^2 \le n\rho_+(1)$, we have from sub-Gaussian tail bound that for all $j \notin \bar{F}$ and $\varepsilon > 0$:

$$P\left[|\hat{\varepsilon}_j| \ge \varepsilon \right] \le 2\exp\left[-n\varepsilon^2 / \left(2\sigma^2 \rho_+(1) \right) \right].$$

Moreover, for each $j \in \bar{F}$, we have

$$|\tilde{\mathbf{w}}_j - \bar{\mathbf{w}}_j| = e_j^\top \left(X_{\bar{F}}^\top X_{\bar{F}} \right)^{-1} X_{\bar{F}}^\top (\mathbf{y} - \mathbf{E}\mathbf{y}).$$

Since $\|e_j^\top (X_{\bar{F}}^\top X_{\bar{F}})^{-1} X_{\bar{F}}^\top\|_2^2 = e_j^\top (X_{\bar{F}}^\top X_{\bar{F}})^{-1} e_j \le n^{-1}\rho_-(\bar{k})^{-1}$, we have for all $\varepsilon > 0$:

$$P\left[|\tilde{\mathbf{w}}_j - \bar{\mathbf{w}}_j| \ge \varepsilon \right] \le 2\exp\left[-n\rho_-(\bar{k})\varepsilon^2 / \left(2n\sigma^2 \right) \right].$$

Taking union bound for $j = 1, \ldots, p$ (each with probability η/p), we obtain the desired inequality. $\qquad\square$

Lemma 4. *Consider $G \subset \{1, \ldots, d\}$ such that $\bar{F} \cap G = \varnothing$. Let $\hat{\mathbf{w}} = \hat{\mathbf{w}}^{(\ell)}$ be the solution of* (6), *and let $\Delta\hat{\mathbf{w}} = \hat{\mathbf{w}} - \tilde{\mathbf{w}}$. Let $\lambda_G = \min_{j \in G} \lambda_j^{(\ell-1)}$ and $\lambda_0 = \max_j \lambda_j^{(\ell-1)}$. If $2\|\hat{\varepsilon}\|_\infty < \lambda_G$, then*

$$\sum_{j \in G} |\hat{\mathbf{w}}_j| \le \frac{2\|\hat{\varepsilon}\|_\infty}{\lambda_G - 2\|\hat{\varepsilon}\|_\infty} \sum_{j \notin \bar{F} \cup G} |\hat{\mathbf{w}}_j| + \frac{\lambda_0}{\lambda_G - 2\|\hat{\varepsilon}\|_\infty} \sum_{j \in \bar{F}} |\Delta\hat{\mathbf{w}}_j| \le \frac{\lambda_0}{\lambda_G - 2\|\hat{\varepsilon}\|_\infty} \|\Delta\hat{\mathbf{w}}_{G^c}\|_1.$$

Proof. For simplicity, let $\lambda_j = \lambda_j^{(\ell-1)}$. The first order equation implies that

$$\frac{1}{n}\sum_{i=1}^{n} 2\big(\mathbf{x}_i^\top \hat{\mathbf{w}} - y_i\big)\mathbf{x}_{i,j} + \lambda_j \, \mathrm{sgn}(\hat{\mathbf{w}}_j) = 0,$$

where $\mathrm{sgn}(\mathbf{w}_j) = 1$ when $\mathbf{w}_j > 0$, $\mathrm{sgn}(\mathbf{w}_j) = -1$ when $\mathbf{w}_j < 0$, and $\mathrm{sgn}(\mathbf{w}_j) \in [-1, 1]$ when $\mathbf{w}_j = 0$. This implies that for all $\mathbf{v} \in \mathbb{R}^p$, we have

$$2\mathbf{v}^\top A \Delta\hat{\mathbf{w}} \leq -2\mathbf{v}^\top \hat{\varepsilon} - \sum_{j=1}^{p} \lambda_j \mathbf{v}_j \, \mathrm{sgn}(\hat{\mathbf{w}}_j). \tag{10}$$

Now, let $\mathbf{v} = \Delta\hat{\mathbf{w}}$ in (10), and notice that $\hat{\varepsilon}_{\bar{F}} = 0$, we obtain

$$0 \leq 2\Delta\hat{\mathbf{w}}^\top A \Delta\hat{\mathbf{w}} \leq 2\big|\Delta\hat{\mathbf{w}}^\top \hat{\varepsilon}\big| - \sum_{j=1}^{p} \lambda_j \Delta\hat{\mathbf{w}}_j \, \mathrm{sgn}(\hat{\mathbf{w}}_j)$$

$$\leq 2\|\Delta\hat{\mathbf{w}}_{\bar{F}^c}\|_1 \|\hat{\varepsilon}\|_\infty - \sum_{j \in \bar{F}} \lambda_j \Delta\hat{\mathbf{w}}_j \, \mathrm{sgn}(\hat{\mathbf{w}}_j) - \sum_{j \notin \bar{F}} \lambda_j \Delta\hat{\mathbf{w}}_j \, \mathrm{sgn}(\hat{\mathbf{w}}_j)$$

$$\leq 2\|\Delta\hat{\mathbf{w}}_{\bar{F}^c}\|_1 \|\hat{\varepsilon}\|_\infty + \sum_{j \in \bar{F}} \lambda_j |\Delta\hat{\mathbf{w}}_j| - \sum_{j \notin \bar{F}} \lambda_j |\hat{\mathbf{w}}_j|$$

$$\leq \sum_{j \in G} \big(2\|\hat{\varepsilon}\|_\infty - \lambda_G\big)|\hat{\mathbf{w}}_j| + \sum_{j \notin G \cup \bar{F}} 2\|\hat{\varepsilon}\|_\infty |\hat{\mathbf{w}}_j| + \sum_{j \in \bar{F}} \lambda_0 |\Delta\hat{\mathbf{w}}_j|.$$

By rearranging the above inequality, we obtain the first desired bound. The second inequality uses $2\|\hat{\varepsilon}\|_\infty \leq \lambda_0$. □

Lemma 5. *Using the notations of Lemma 4, and let J be the indices of the largest s coefficients (in absolute value) of $\hat{\mathbf{w}}_G$. Let $I = G^c \cup J$ and $k = |G^c|$. If $0 \leq \lambda_0/(\lambda_G - 2\|\hat{\varepsilon}\|_\infty) \leq 3$, then*

$$\|\Delta\hat{\mathbf{w}}\|_2 \leq \big(1 + (3k/s)^{0.5}\big)\|\Delta\hat{\mathbf{w}}_I\|_2.$$

Proof. Using $\lambda_0/(\lambda_G - 2\|\hat{\varepsilon}\|_\infty) \leq 3$, we obtain from Lemma 4

$$\|\hat{\mathbf{w}}_G\|_1 \leq 3\|\Delta\hat{\mathbf{w}} - \hat{\mathbf{w}}_G\|_1.$$

Therefore,

$$\|\Delta\hat{\mathbf{w}} - \Delta\hat{\mathbf{w}}_I\|_\infty \leq \|\Delta\hat{\mathbf{w}}_J\|_1/s$$
$$= s^{-1}[\|\Delta\hat{\mathbf{w}}_G\|_1 - \|\Delta\hat{\mathbf{w}} - \Delta\hat{\mathbf{w}}_I\|_1]$$
$$\leq s^{-1}[3\|\Delta\hat{\mathbf{w}} - \hat{\mathbf{w}}_G\|_1 - \|\Delta\hat{\mathbf{w}} - \Delta\hat{\mathbf{w}}_I\|_1],$$

which implies that

$$
\begin{aligned}
\|\Delta\hat{\mathbf{w}} - \Delta\hat{\mathbf{w}}_I\|_2 &\le (\|\Delta\hat{\mathbf{w}} - \Delta\hat{\mathbf{w}}_I\|_1 \|\Delta\hat{\mathbf{w}} - \Delta\hat{\mathbf{w}}_I\|_\infty)^{1/2} \\
&\le \left[\|\Delta\hat{\mathbf{w}} - \Delta\hat{\mathbf{w}}_I\|_1 (3\|\Delta\hat{\mathbf{w}} - \hat{\mathbf{w}}_G\|_1 - \|\Delta\hat{\mathbf{w}} - \Delta\hat{\mathbf{w}}_I\|_1)\right]^{1/2} s^{-1/2} \\
&\le \left[(3\|\Delta\hat{\mathbf{w}} - \hat{\mathbf{w}}_G\|_1/2)^2\right]^{1/2} s^{-1/2} \\
&\le (3/2)s^{-1/2}\|\Delta\hat{\mathbf{w}} - \hat{\mathbf{w}}_G\|_1 \\
&\le (3/2)s^{-1/2}\bar{k}^{1/2}\|\Delta\hat{\mathbf{w}} - \hat{\mathbf{w}}_G\|_2 \le (3k/s)^{1/2}\|\Delta\hat{\mathbf{w}}_I\|_2.
\end{aligned}
$$

The third inequality uses the simple algebraic inequality $a(3b - a) \le (3b/2)^2$. By rearranging this inequality, we obtain the desired bound. Note that in the above derivation, we have used the fact that $\bar{F} \cap G = \varnothing$, which implies that $\Delta\hat{\mathbf{w}}_G = \hat{\mathbf{w}}_G$, and thus $\Delta\hat{\mathbf{w}} - \hat{\mathbf{w}}_G = \Delta\hat{\mathbf{w}}_{G^c}$. $\qquad\square$

Lemma 6. *Let the conditions of Lemma* 4 *and Lemma* 5 *hold, and let* $k = |G^c|$. *If* $t = 1 - \pi(k + s, s)k^{1/2}s^{-1} \in (0, 4/3)$, *and* $0 \le \lambda_0/(\lambda_G - 2\|\hat{\varepsilon}\|_\infty) \le (4 - t)/(4 - 3t)$, *then*

$$
\|\Delta\hat{\mathbf{w}}\|_2 \le \left(1 + (3k/s)^{0.5}\right)\|\Delta\hat{\mathbf{w}}_I\|_2
$$

$$
\le \frac{1 + (3k/s)^{0.5}}{t\rho_-(k + s)}\left[2\|\hat{\varepsilon}_{G^c}\|_2 + \left(\sum_{j\in\bar{F}}(\lambda_j^{(\ell-1)})^2\right)^{1/2}\right].
$$

Proof. Let J be the indices of the largest s coefficients (in absolute value) of $\hat{\mathbf{w}}_G$, and $I = G^c \cup J$. The conditions of the lemma imply that

$$
\begin{aligned}
\max\left(0, \Delta\hat{\mathbf{w}}_I^\top A\Delta\hat{\mathbf{w}}\right) &\ge \rho_-(k + s)\left[\|\Delta\hat{\mathbf{w}}_I\|_2 - \pi(k + s, s)\|\hat{\mathbf{w}}_G\|_1/s\right]\|\Delta\hat{\mathbf{w}}_I\|_2 \\
&\ge \rho_-(k + s)\left[1 - (1 - t)(4 - t)(4 - 3t)^{-1}\right]\|\Delta\hat{\mathbf{w}}_I\|_2^2 \\
&\ge 0.5t\rho_-(k + s)\|\Delta\hat{\mathbf{w}}_I\|_2^2.
\end{aligned}
$$

In the above derivation, the first inequality is due to Lemma 2; the second inequality is due to the conditions of this lemma plus Lemma 4, which implies that

$$
\|\hat{\mathbf{w}}_G\|_1 \le \frac{\lambda_0}{\lambda_G - 2\|\hat{\varepsilon}\|_\infty}\|\hat{\mathbf{w}}_{G^c}\|_1 \le \frac{\lambda_0}{\lambda_G - 2\|\hat{\varepsilon}\|_\infty}\sqrt{k}\|\hat{\mathbf{w}}_I\|_2;
$$

and the last inequality follows from $1 - (1 - t)(4 - t)(4 - 3t)^{-1} \ge 0.5t$, which holds for $t \in (0, 4/3)$.

If $\Delta\hat{\mathbf{w}}_I^\top A\Delta\hat{\mathbf{w}} \le 0$, then the above inequality, together with Lemma 5, imply the lemma. Therefore in the following, we can assume that

$$
\Delta\hat{\mathbf{w}}_I^\top A\Delta\hat{\mathbf{w}} \ge 0.5t\rho_-(k + s)\|\Delta\hat{\mathbf{w}}_I\|_2^2.
$$

Moreover, let $\lambda_j = \lambda_j^{(\ell-1)}$. We obtain from (10) with $\mathbf{v} = \Delta\hat{\mathbf{w}}_I$ the following:

$$2\Delta\hat{\mathbf{w}}_I^\top A \Delta\hat{\mathbf{w}} \leq -2\Delta\hat{\mathbf{w}}_I^\top \hat{\varepsilon} - \sum_{j\in I} \lambda_j \Delta\hat{\mathbf{w}}_j \operatorname{sgn}(\hat{\mathbf{w}}_j)$$

$$= -2\Delta\hat{\mathbf{w}}_I^\top \hat{\varepsilon}_{G^c} - 2\Delta\hat{\mathbf{w}}_I^\top \hat{\varepsilon}_G - \sum_{j\in\bar{F}} \lambda_j \Delta\hat{\mathbf{w}}_j \operatorname{sgn}(\hat{\mathbf{w}}_j)$$

$$- \sum_{j\in G} \lambda_j |\Delta\hat{\mathbf{w}}_j| - \sum_{j\in\bar{F}^c\cap G^c} \lambda_j |\Delta\hat{\mathbf{w}}_j|$$

$$\leq 2\|\Delta\hat{\mathbf{w}}_I\|_2 \|\hat{\varepsilon}_{G^c}\|_2 + 2\|\hat{\varepsilon}_G\|_\infty \sum_{j\in G} |\Delta\hat{\mathbf{w}}_j| + \sum_{j\in\bar{F}} \lambda_j |\Delta\hat{\mathbf{w}}_j| - \sum_{j\in G} \lambda_j |\Delta\hat{\mathbf{w}}_j|$$

$$\leq 2\|\Delta\hat{\mathbf{w}}_I\|_2 \|\hat{\varepsilon}_{G^c}\|_2 + \left(\sum_{j\in\bar{F}} \lambda_j^2\right)^{1/2} \|\Delta\hat{\mathbf{w}}_I\|_2.$$

Note that the equality uses the fact that $G \subset \bar{F}^c$, and $\Delta\hat{\mathbf{w}}_j \operatorname{sgn}(\hat{\mathbf{w}}_j) = |\hat{\mathbf{w}}_j|$ for $j \in \bar{F}^c$. The last inequality uses the fact that $\forall j \in G$: $\lambda_j \geq \lambda_G \geq 2\|\hat{\varepsilon}_G\|_\infty$. Now by combining the above two estimates, we obtain

$$\|\Delta\hat{\mathbf{w}}_I\|_2 \leq \frac{1}{t\rho_-(k+s)} \left[2\|\hat{\varepsilon}_{G^c}\|_2 + \left(\sum_{j\in\bar{F}} \lambda_j^2\right)^{1/2}\right].$$

The desired bound follows from Lemma 5. $\qquad\square$

Lemma 7. *Given $\mathbf{w} \in R^p$, and define $\lambda_j = \lambda I(|\mathbf{w}_j| \leq \theta)$. We have*

$$\left(\sum_{j\in\bar{F}} \lambda_j^2\right)^{1/2} \leq \lambda\sqrt{\sum_{j\in\bar{F}} I(|\bar{\mathbf{w}}_j| \leq 2\theta)} + \lambda|\{j \in \bar{F}: |\bar{\mathbf{w}}_j - \mathbf{w}_j| \geq \theta\}|^{1/2}.$$

Proof. If $|\bar{\mathbf{w}}_j - \mathbf{w}_j| \geq \theta$, then

$$I(|\mathbf{w}_j| \leq \theta) \leq 1 \leq I(|\bar{\mathbf{w}}_j - \mathbf{w}_j| \geq \theta);$$

otherwise, $I(|\mathbf{w}_j| \leq \theta) \leq I(|\bar{\mathbf{w}}_j| \leq 2\theta)$. It follows that the following inequality always holds:

$$I(|\mathbf{w}_j| \leq \theta) \leq I(|\bar{\mathbf{w}}_j| \leq 2\theta) + I(|\bar{\mathbf{w}}_j - \mathbf{w}_j| \geq \theta).$$

The desired bound is a direct consequence of the above result and the 2-norm triangle inequality

$$\left(\sum_j (x_j + \Delta x_j)^2\right)^{1/2} \leq \left(\sum_j x_j^2\right)^{1/2} + \left(\sum_j \Delta x_j^2\right)^{1/2},$$

with $x_j = \lambda I(|\bar{\mathbf{w}}_j| \leq 2\theta)$ and $\Delta x_j = \lambda I(|\bar{\mathbf{w}}_j - \mathbf{w}_j| \geq \theta)$. $\qquad\square$

Lemma 8. *Define* $F^{(\ell)} = \{j \colon |\hat{\mathbf{w}}_j^{(\ell)} - \bar{\mathbf{w}}_j| \ge \theta\}$. *Under the conditions of Theorem* 1, *we have for all* $s \ge 2\bar{k}$:

$$\|\hat{\mathbf{w}}^{(\ell)} - \tilde{\mathbf{w}}\|_2 \le \frac{5.7\lambda}{\rho_-(1.5\bar{k} + s)}\sqrt{|F^{(\ell-1)}|},$$

and

$$\sqrt{|F^{(\ell)}|} \le \frac{6\lambda\theta^{-1}}{\rho_-(1.5\bar{k} + s)}\sqrt{|F^{(\ell-1)}|}.$$

Proof. For all $t \in [0.5, 4/3]$, by using Lemma 3, we know that the condition of the theorem implies that

$$\frac{\lambda}{\lambda - 2\|\hat{\varepsilon}\|_\infty} \le 7/5 \le \frac{4-t}{4-3t}.$$

Moreover, Lemma 1 implies that the condition

$$0.5 \le t = 1 - \pi(1.5\bar{k} + s, s)(1.5\bar{k})^{0.5}/s$$

is also satisfied. This means that the conditions of Lemma 6 (with $\lambda_0 = \lambda_G = \lambda$) are satisfied.

Now, we assume that at some $\ell \ge 1$,

$$|G_\ell^c| \le 1.5\bar{k}, \qquad \text{where } G_\ell = \left\{j \notin \bar{F} \colon \lambda_j^{(\ell-1)} = \lambda\right\}, \tag{11}$$

then it is easy to verify that $G_\ell^c \setminus \bar{F} \subset F^{(\ell-1)}$.

Moreover, with the definition of $G = G_\ell$ in Lemma 6 and Lemma 7, we can set $\lambda_0 = \lambda_G = \lambda$ and obtain (note also that $\hat{\varepsilon}_{\bar{F}} = 0$)

$$\|\hat{\mathbf{w}}^{(\ell)} - \tilde{\mathbf{w}}\|_2 \le \frac{1 + \sqrt{3}}{t\rho_-(1.5\bar{k} + s)}\left[2\|\hat{\varepsilon}_{G_\ell^c \setminus \bar{F}}\|_2 + \left(\sum_{j \in \bar{F}}(\lambda_j^{(\ell-1)})^2\right)^{1/2}\right]$$

$$\le \frac{1 + \sqrt{3}}{t\rho_-(1.5\bar{k} + s)}\left[2\sqrt{|F^{(\ell-1)} \setminus \bar{F}|}\|\hat{\varepsilon}\|_\infty + \sqrt{|F^{(\ell-1)} \cap \bar{F}|}\lambda\right]$$

$$\le \frac{1 + \sqrt{3}}{t\rho_-(1.5\bar{k} + s)}\left[(2/7)\sqrt{|F^{(\ell-1)} \setminus \bar{F}|} + \sqrt{|F^{(\ell-1)} \cap \bar{F}|}\right]\lambda$$

$$\le \frac{1 + \sqrt{3}}{0.5\rho_-(1.5\bar{k} + s)}\left[\sqrt{1.082|F^{(\ell-1)}|}\right]\lambda$$

$$\le \frac{5.7\lambda}{\rho_-(1.5\bar{k} + s)}\sqrt{|F^{(\ell-1)}|},$$

where the first inequality is due to Lemma 6. The second inequality uses the facts that $G_\ell^c \setminus \bar{F} \subset F^{(\ell-1)} \setminus \bar{F}$, and Lemma 7 with $I(|\bar{\mathbf{w}}_j| \le 2\theta) = 0$ (for all $j \in \bar{F}$). The third inequality uses $2\|\hat{\varepsilon}\|_\infty \le (2/7)\lambda$, and the fourth inequality uses $(2/7)a + b \le \sqrt{1.082(a^2 + b^2)}$.

Since Lemma 3 implies that

$$\|\tilde{\mathbf{w}} - \bar{\mathbf{w}}\|_{\infty} \leq (1/7)\lambda/\sqrt{\rho_+(1)\rho_-(\bar{k})},$$

we know that $j \in F^{(\ell)}$ implies that

$$\left|\tilde{\mathbf{w}}_j - \hat{\mathbf{w}}_j^{(\ell)}\right| \geq \theta - (1/7)\lambda/\sqrt{\rho_+(1)\rho_-(\bar{k})} \geq (41/42)\theta.$$

Therefore,

$$\sqrt{|F^{(\ell)}|} \leq (41\theta/42)^{-1} \left\|\tilde{\mathbf{w}} - \hat{\mathbf{w}}^{(\ell)}\right\|_2$$

$$\leq \frac{5.7\lambda(41\theta/42)^{-1}}{\rho_-(1.5\bar{k} + s)}\sqrt{|F^{(\ell-1)}|}$$

$$\leq \frac{6\lambda\theta^{-1}}{\rho_-(1.5\bar{k} + s)}\sqrt{|F^{(\ell-1)}|}.$$

That is, under the assumption of (11), the lemma holds at ℓ.

Therefore, next we only need to prove by induction on ℓ that (11) holds for all $\ell = 1, 2, \ldots$. When $\ell = 1$, we have $G_1^c = \bar{F}$, which implies that (11) holds.

Now assume that (11) holds at ℓ for some $\ell \geq 1$. Then by the induction hypothesis we know that the lemma holds at ℓ. This means that

$$\sqrt{|G_{\ell+1}^c \setminus \bar{F}|} \leq \sqrt{|F^{(\ell)}|}$$

$$\leq \frac{6\lambda\theta^{-1}}{\rho_-(1.5\bar{k} + s)}\sqrt{|F^{(\ell-1)}|}$$

$$\leq \sqrt{0.5|F^{(\ell-1)}|}$$

$$\leq \cdots \leq 0.5^{\ell/2}|F^{(0)}|.$$

The first inequality is due to the fact $G_{\ell+1}^c \setminus \bar{F} \subset F^{(\ell)}$. The second inequality uses the assumption of θ in the theorem. The last inequality uses induction. Now note that $F^{(0)} = \bar{F}$, we thus have $|G_{\ell+1}^c \setminus \bar{F}| \leq 0.5\bar{k}$. This completes the induction step. $\qquad\square$

5.2. Proof of Theorem 1

Define

$$\beta = \frac{6\lambda\theta^{-1}}{\rho_-(1.5\bar{k} + s)}.$$

We have $\beta < 1$ by the assumption of the theorem. Using induction, we have from Lemma 8 that

$$
\begin{aligned}
\sqrt{\left|F^{(L)}\right|} &\leq \beta \sqrt{\left|F^{(L-1)}\right|} \\
&\leq \cdots \\
&\leq \beta^L \sqrt{\left|F^{(0)}\right|} \\
&\leq \beta^L \sqrt{\bar{k}} < 1.
\end{aligned}
$$

This means that when $\ell > L$, $|F^{(\ell-1)}| = 0$. Therefore by applying Lemma 8 again, we obtain

$$
\left\| \hat{\mathbf{w}}^{(\ell)} - \tilde{\mathbf{w}} \right\|_2 = 0.
$$

Since Lemma 3 implies that

$$
\|\tilde{\mathbf{w}} - \bar{\mathbf{w}}\|_\infty \leq (1/7)\lambda/\sqrt{\rho_+(1)\rho_-(\bar{k})} < \theta,
$$

we have

$$
\operatorname{supp}(\tilde{\mathbf{w}}) = \operatorname{supp}(\bar{\mathbf{w}}).
$$

This implies that $\operatorname{supp}(\hat{\mathbf{w}}^{(\ell)}) = \operatorname{supp}(\bar{\mathbf{w}})$.

6. Discussion

This paper investigated the performance of multi-stage convex relaxation for feature selection, where it is shown that under RIP, the procedure can achieve unbiased feature selection. This result complements that of [16] which studies the parameter estimation performance of multi-stage convex relaxation. It also complements similar results obtained in [17] and [13] for different computational procedures. One advantage of our result over that in [13] is that the multi-stage convex relaxation method is provably efficient because the correct feature set can be obtained after no more than $O(\log \bar{k})$ number of iterations. In comparison, a computational efficiency statement for the path-following method of [13] remains open.

Acknowledgements

References

[1] Bickel, P.J., Ritov, Y. and Tsybakov, A.B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469

[2] Bunea, F., Tsybakov, A. and Wegkamp, M. (2007). Sparsity oracle inequalities for the Lasso. *Electron. J. Stat.* **1** 169–194. MR2312149

[3] Candes, E.J. and Tao, T. (2005). Decoding by linear programming. *IEEE Trans. Inform. Theory* **51** 4203–4215. MR2243152

[4] Candès, E.J., Wakin, M.B. and Boyd, S.P. (2008). Enhancing sparsity by reweighted l_1 minimization. *J. Fourier Anal. Appl.* **14** 877–905. MR2461611

[5] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

[6] Koltchinskii, V. (2009). Sparsity in penalized empirical risk minimization. *Ann. Inst. Henri Poincaré Probab. Stat.* **45** 7–57.

[7] Lounici, K. (2008). Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electron. J. Stat.* **2** 90–102. MR2386087

[8] Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363

[9] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

[10] van de Geer, S.A. and Bühlmann, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat.* **3** 1360–1392. MR2576316

[11] Wainwright, M.J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Inform. Theory* **55** 2183–2202. MR2729873

[12] Wipf, D.P. and Nagarajan, S. (2010). Iterative reweighted ℓ_1 and ℓ_2 methods for finding sparse solutions. *Journal of Selected Topics in Signal Processing* **4** 317–329.

[13] Zhang, C.H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942. MR2604701

[14] Zhang, C.H. and Huang, J. (2008). The sparsity and bias of the LASSO selection in high-dimensional linear regression. *Ann. Statist.* **36** 1567–1594. MR2435448

[15] Zhang, T. (2009). Some sharp performance bounds for least squares regression with L_1 regularization. *Ann. Statist.* **37** 2109–2144. MR2543687

[16] Zhang, T. (2010). Analysis of multi-stage convex relaxation for sparse regularization. *J. Mach. Learn. Res.* **11** 1081–1107. MR2629825

[17] Zhang, T. (2011). Adaptive forward–backward greedy algorithm for learning sparse representations. *IEEE Trans. Inform. Theory* **57** 4689–4708. MR2840485

[18] Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563. MR2274449

[19] Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469

[20] Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.* **36** 1509–1533. MR2435443