Adaptive Lasso for High Dimensional Regression and Gaussian Graphical Modeling [1]

Shuheng Zhou Sara van de Geer Peter Bühlmann

Seminar für Statistik
ETH Zürich
CH-8092 Zürich, Switzerland

March 13th, 2009

Abstract

We show that the two-stage adaptive Lasso procedure (Zou, 2006) is consistent for high-dimensional model selection in linear and Gaussian graphical models. Our conditions for consistency cover more general situations than those accomplished in previous work: we prove that restricted eigenvalue conditions (Bickel et al., 2008) are also sufficient for sparse structure estimation.

1 Introduction

The problem of inferring the sparsity pattern, i.e. model selection, in high-dimensional problems has recently gained a lot of attention. One important stream of research, which we also adopt here, requires computational feasibility and provable statistical properties of estimation methods or algorithms. Regularization with ℓ_1-type penalization has become extremely popular for model selection in high-dimensional scenarios. The methods are easy to use, due to recent progress in convex optimization (Meier et al., 2008), (Friedman et al., 2008a), and they are asymptotically consistent or oracle optimal when requiring some conditions, e.g. on the design matrix in a linear model or among the variables in a graphical model (Greenshtein and Ritov, 2004; Meinshausen and Bühlmann, 2006; van de Geer, 2008), (Bickel et al., 2008). However, these conditions, referred to as coherence or compatibility conditions, are often very restrictive. The restrictions are due to severe bias problems with ℓ_1-penalization, i.e. shrinking also the estimates which correspond to true signal variables, see also Zou (2006), Meinshausen (2007).

Regularization with the ℓ_q-norm with $q < 1$ would mitigate some of the bias problems but become computationally infeasible as the penalty is non-convex. As an interesting alternative, one can consider multi-step procedures where each of the steps involves a convex optimization only. A prime example is the adaptive Lasso (Zou, 2006) which is a two-step algorithm and whose repeated application corresponds in some "loose" sense to a non-convex penalization scheme (Zou and Li, 2008). We are analyzing in this paper this adaptive Lasso procedure for variable selection in linear models as well as for Gaussian graphical modeling.

[1] Research supported by SNF 20PA21-120050/1.

Both frameworks are related to each other and for both of them, we derive results for model selection under rather weak conditions. In particular, our results imply that the adaptive Lasso can recover the true underlying model in situations where plain ℓ_1-regularization fails (assuming restricted eigenvalue conditions).

1.1 Variable selection in linear models

Consider the linear model

$$Y = X\beta + \epsilon, \tag{1.1}$$

where X is an $n \times p$ design matrix, Y is an $n \times 1$ vector of noisy observations and ϵ being the noise term. The design matrix is treated as either fixed or random. We assume throughout this paper that $p \geq n$ (i.e. high-dimensional) and $\epsilon \sim N(0, \sigma_\epsilon^2 I_n)$.

The sparse object to recover is the unknown parameter $\beta \in \mathbb{R}^p$. We assume that it has a relatively small number s of nonzero coefficients: $S := \text{supp}(\beta) = \{j : \beta_j \neq 0\}$ and $s = |\text{supp}(\beta)|$. Let $\beta_{\min} := \min_{j \in S} |\beta_j|$. Inferring the sparsity pattern, i.e. variable selection, refers to the task of correctly estimating the support set $\text{supp}(\beta)$ based on noisy observations from (1.1). In particular, given some estimator $\widehat{\beta}$, recovery of the relevant variables is understood to be

$$\text{supp}(\widehat{\beta}) = \text{supp}(\beta) \text{ with high probability}. \tag{1.2}$$

Regularized estimation with the ℓ_1-norm penalty, also known as the Lasso (Tibshirani, 1996), refers to the following convex optimization problem:

$$\widehat{\beta} = \arg\min_\beta \frac{1}{2n}\|Y - X\beta\|_2^2 + \lambda_n \|\beta\|_1, \tag{1.3}$$

where the scaling factor $1/(2n)$ is chosen by convenience and $\lambda_n \geq 0$ is a penalization parameter. It is an attractive and computationally tractable method with provable good statistical properties, even if p is much larger than n, for prediction (Greenshtein and Ritov, 2004), for estimation in terms of the ℓ_1- or ℓ_2-loss (van de Geer, 2008; Meinshausen and Yu, 2009; Bickel et al., 2008) and for variable selection (Meinshausen and Bühlmann, 2006; Zhao and Yu, 2006; Wainwright, 2008). For the specific problem of variable selection, it is known that the so-called "neighborhood stability condition" for the design matrix (Meinshausen and Bühlmann, 2006), which has been re-formulated in a nicer form as the "irrepresentable condition" (Zhao and Yu, 2006), is necessary and sufficient for consistent variable selection in the sense of (1.2). Moreover, as this condition is restrictive, its necessity implies that the Lasso only works in a rather restricted range of problems, excluding cases where the design exhibits too strong (empirical) correlations. A key motivation of our work is to continue the exploration of a computationally tractable algorithm for variable selection, while aiming to relax the stringent conditions that are imposed on the design matrix X.

Towards these goals, we analyze the adaptive Lasso procedure, see (2.2) below, for variable selection in the high-dimensional setting. This method was originally proposed by Zou (2006) and he analyzed the case when p is fixed. Further progress of analyzing the adaptive Lasso in the high-dimensional scenario has been achieved by Huang et al. (2008). A more complete understanding of its power, when applied to the high dimensional setting where $p \gg n$ is still lacking. We prove in this paper that variable selection with the adaptive Lasso is possible under rather general incoherence conditions on the design. We do not require more stringent conditions on the design X than Bickel et al. (2008) who give the currently weakest conditions for

convergence of the Lasso in terms of $\|\widehat{\beta} - \beta\|_1$ and $\|\widehat{\beta} - \beta\|_2$. We show that for an initial estimator β_{init} in the two-stage adaptive Lasso procedure with a sufficiently reasonable behavior of $\|\beta_{\text{init}} - \beta\|_\infty$, model selection is possible assuming only a lower bound on the smallest eigenvalue of $X_S^T X_S/n$, where X_S denotes the submatrix of X whose columns are indexed by S, and some restrictions on $\beta_{\min}$ and the sparsity level s. Thus, variable selection is possible under rather general design conditions by the two-stage adaptive Lasso, and it is necessary to move away from plain ℓ_1-regularization, see Meinshausen and Bühlmann (2006), Zhao and Yu (2006).

1.2 Covariance selection in Gaussian graphical models

Covariance selection in a Gaussian graphical model refers to the problem of inferring conditional independencies between a set of jointly Gaussian random variables

$$X_1, \dots X_p \sim N(0, \Sigma) \tag{1.4}$$

(the restriction to mean 0 is without loss of generality). These variables $X_1, \dots, X_p$ correspond to nodes in a graph, labeled by $\{1, \dots, p\}$, and a Gaussian conditional independence graph is then defined as follows:

$$\text{there is an undirected edge between node } i \text{ and } j \;\Leftrightarrow\; \Sigma^{-1}_{ij} \neq 0.$$

The definition of an edge is equivalent to requiring that X_i and X_j are conditionally dependent given all remaining variables $\{X_k;\ k \neq i, j\}$. For details cf. Lauritzen (1996). Estimation of the edge set is thus equivalent to finding the zeroes in the concentration matrix Σ^{-1}.

In the high-dimensional scenario with $p \geq n$, where n denotes the sample size of i.i.d. copies from (1.4), ℓ_1-type regularization has been analyzed.
Meinshausen and Bühlmann (2006) prove that it is possible to consistently infer the edge set by considering many variable selection problems in high-dimensional Gaussian regressions, again requiring a global neighborhood stability or irrepresentable condition which puts some restrictions on the covariance matrix Σ. Later, the GLasso penalization has been proposed (Friedman et al., 2008b; Banerjee et al., 2008) which is a sparse estimator for Σ^{-1} using an ℓ_1-penalty on the non-diagonal elements of Σ^{-1} in the multivariate Gaussian log-likelihood. Ravikumar et al. (2008) recently obtained results for consistent covariance selection ((i.e. inferring the edge set) using the GLasso by imposing mutual incoherence conditions (analogous to the neighborhood stability condition) on the Fisher information matrix (of size $p^2 \times p^2$) of the model, which is an edge-based counterpart of Σ.

We focus here on generalizing conditions for the pursuit via many regressions: we prove in this paper a result for inferring the edge set in a Gaussian graphical model, under a rather general condition on Σ closely related to the restricted eigenvalue assumptions in Bickel et al. (2008) by analyzing the pursuit of many regressions with the adaptive Lasso. We conjecture that the set of conditions which we are imposing are more general than what Ravikumar et al. (2008) require when using the GLasso, although this is a point that needs to be thoroughly studied as we discuss further in Section 7. We also suspect that the GLasso approach is intrinsically more limited, in terms of restrictions for the covariance matrix Σ than the approach from Meinshausen and Bühlmann (2006) via considering many regressions. This has been recognized by Meinshausen (2008) and also studied by Ravikumar et al. (2008) on specific graphical models. On the other hand, for well-behaved problems, GLasso might have an advantage because it exploits the positive definiteness of Σ and Σ^{-1}.

1.3 Related work

Recently, Huang et al. (2008) studied the adaptive Lasso estimators in sparse, high-dimensional linear regression models for a fixed design. Under a rather strong mutual incoherence condition between every pair of relevant and irrelevant covariates and assuming other regularity conditions, they prove that the adaptive Lasso recovers the correct model and has an oracle property. While they have derived the same incoherence condition as one (among others) of ours in (8.4a) in order for the second stage weighted Lasso procedure to achieve model selection consistency, they achieve it by an initial estimator assuming some strong mutual incoherence condition which bounds the pairwise correlations of the columns of the design. This is a much stronger condition than the restricted eigenvalue assumptions that we make, see Bickel et al. (2008).

Meinshausen and Yu (2009) examined the variable selection property of the Lasso followed by a thresholding procedure. Under a relaxed "incoherence design" assumption, Meinshausen and Yu (2009) show that the estimator is still consistent in the ℓ_2-norm sense for fixed designs, and furthermore, it is possible to do hard-thresholding on the ordinary Lasso estimator to achieve variable selection consistency. However the choice of the threshold parameter depends on the the unknown value $\beta_{\min}$ and the sparsity s of β. It is not clear how one can choose such a threshold parameter without knowing $\beta_{\min}$ or s. A more general framework for multi-stage variable selection was studied by Wasserman and Roeder (2008) for various methods and conditions. Their approach controls the probability of false positives (i.e. type I error) but pays a price in terms of false negatives (i.e. type II error) in comparison to the adaptive Lasso (Wasserman and Roeder, 2008).

Finally, our focus is rather different from that of Wainwright (2008, 2007), where the goal was to analyze the least amount of samples that one needs in order to recover a sparse signal via a random or a fixed measurement ensemble that satisfies strong incoherence conditions. It is an open problem to establish a lower bound on the sample size, given p, s and $\beta_{\min}$, to recover the model with the adaptive Lasso, assuming restricted eigenvalue assumptions only.

1.4 Organization of the paper

In Section 2 we define the two-step adaptive Lasso procedure for linear regression and describe our main result: general model selection properties of the second stage weighted procedure for variable selection. Here, the initial estimator β_{init} can be general, and we assume a bound for $\|\beta_{\text{init}} - \beta\|_\infty$. Section 3 presents the restricted eigenvalue conditions we need for deriving bounds for $\|\beta_{\text{init}} - \beta\|_\infty$ with the standard Lasso as initial estimator β_{init}. In Sections 4, 5 and 6, we summarize conditions and results, with the standard Lasso as initial estimator, for linear regression with fixed design, linear regression with random design, and Gaussian graphical modeling, respectively. These results are consequences of our general result in Section 2. Section 8 presents a model selection lemma for the weighted Lasso with general weights. The remainder of the paper contains the proofs.

2 The adaptive Lasso estimator and its general properties

Consider the linear model in (1.1). We distinguish later between fixed and random design.

2.1 The two-stage adaptive Lasso procedure

The adaptive Lasso is the Lasso estimator with a re-weighted penalty function, see (2.2) below. The weights are estimated from an initial estimator β_{init}:

$$w_j := \max\{\frac{1}{|\beta_{j,\text{init}}|}, 1\}. \tag{2.1}$$

We note that the original proposal of Zou (2006) uses $w_j = 1/|\beta_{j,\text{init}}|^\gamma$ for some $\gamma > 0$ with $\gamma = 1$ the most common choice. The adaptive Lasso is now defined by a second-stage weighted Lasso:

$$\widehat{\beta} = \arg\min_\beta \frac{1}{2n}\|Y - X\beta\|_2^2 + \lambda_n \sum_{j=1}^p w_j|\beta_j|. \tag{2.2}$$

2.2 Variable selection with the adaptive Lasso estimator

Correct variable selection with the adaptive Lasso requires some conditions for the design. We first make some assumptions related to the design matrix. For a symmetric matrix A, let $\Lambda_{\min}(A)$ denote the smallest eigenvalue of A.

For a fixed design matrix X, we define

$$\Lambda_{\min}(s) := \min_{\substack{J_0 \subseteq \{1,\ldots,p\} \\ |J_0| \le s}} \min_{\substack{\gamma \neq 0 \\ \gamma^c_{J_0} = 0}} \frac{\|X\gamma\|_2^2}{n\,\|\gamma_{J_0}\|_2^2}. \tag{2.3}$$

We assume throughout this paper that $\Lambda_{\min}(s) > 0$. As a consequence of this definition we have,

$$\Lambda_{\min}\left(\frac{X_S^T X_S}{n}\right) \ge \Lambda_{\min}(s) > 0. \tag{2.4}$$

Furthermore, we assume for fixed design that the ℓ_2-norm of each column of X is upper bounded by $c_0\sqrt{n}$ for some constant $c_0 > 0$. We then consider the set

$$\mathcal{T} := \left\{\left\|\frac{X^T\epsilon}{n}\right\|_\infty \le c_0\sigma_\epsilon\sqrt{\frac{6\log p}{n}}\right\}. \tag{2.5}$$

The set $\mathcal{T}$ has large probability, as described below in (2.15).

For a random design matrix X we assume:

$$X \text{ has i.i.d. rows } \sim N(0, \Sigma), \tag{2.6}$$

where we assume without loss of generality that the mean is zero and $\Sigma_{jj} = 1, \forall j = 1, \ldots, p$. We then define

$$\Lambda_{\min}(s) := \frac{16}{17} \min_{\substack{J_0 \subseteq \{1,\ldots,p\} \\ |J_0| \le s}} \min_{\substack{\gamma \neq 0 \\ \gamma^c_{J_0} = 0}} \frac{\gamma^T\Sigma\gamma}{\|\gamma_{J_0}\|_2^2}. \tag{2.7}$$

As for fixed design, we assume that $\Lambda_{\min}(s) > 0$ with large probability. The factor $16/17$ allows us to use the same notation $\Lambda_{\min}(s)$ for both fixed and random design. Let Σ_{SS} be the sub-matrix with rows and columns both indexed by the active set S. It then holds that

$$\Lambda_{\min}(\Sigma_{SS}) \quad \geq \quad \frac{17\Lambda_{\min}(s)}{16} > 0. \tag{2.8}$$

Then, a random design X as in (2.6) behaves nicely, with high probability. To be more precise, denote by $\Delta = \frac{X^T X}{n} - \Sigma$, and consider

$$\mathcal{X} := \left\{ \max_{j,k} |\Delta_{jk}| < C_2 \sqrt{\frac{\log p}{n}} \right\}, \tag{2.9}$$

for some constant $C_2 > 4\sqrt{5/3}$. Throughout this paper, we assume for random design that $p < e^{n/4C_2^2}$, i.e. $C_2\sqrt{\frac{\log p}{n}} < 1/2$, such that $\mathcal{X}$ holds with probability at least $1 - \frac{1}{p^2}$ (cf. (2.16) and Lemma 9.3). We note that this implies that on $\mathcal{X}$,

$$\forall j = 1, \ldots, p, \quad \|X_j\|_2^2 \leq \frac{3n}{2}. \tag{2.10}$$

The set $\mathcal{T}$ in (2.5), intersected with $\mathcal{X}$, is also relevant for random design: the constant c_0 equals $\sqrt{3/2}$, following (2.10).

For both, fixed and random design, we consider the quantity

$$r_n(S) := \left\| X_{S^c}^T X_S (X_S^T X_S)^{-1} \right\|_\infty, \tag{2.11}$$

where $\|A\|_\infty = \max_{1 \leq i \leq k} \sum_{j=1}^m |A_{ij}|$ for a $k \times m$ matrix A. The properties of the adaptive Lasso procedure depend on (an upper bound of) $r_n(S)$.

Finally, we denote by

$$\delta := \beta_{\text{init}} - \beta$$

the difference between the initial estimate and the true parameter value.

Theorem 2.1.. *Consider the adaptive Lasso estimator in a linear model as in (1.1) with design X, where $n \leq p$, and for fixed design: the ℓ_2-norm of each column of X is upper bounded by $c_0\sqrt{n}$ for some constant $c_0 > 0$.*

Assume the upper bound $\widetilde{r_n} \geq r_n(S)$ which we require to hold only on $\mathcal{X}$ in case of a random design. Furthermore, assume on $\mathcal{T}$ for a fixed design and on $\mathcal{X} \cap \mathcal{T}$ for a random design, some upper bounds on δ as follows: $1 > \widetilde{\delta}_S \geq \|\delta_S\|_\infty$ and $1 > \widetilde{\delta}_{S^c} \geq \|\delta_{S^c}\|_\infty$. Suppose that on $\mathcal{T}$ for a fixed design and on $\mathcal{X} \cap \mathcal{T}$ for a random design:
for some $1 > \eta > 0$ and some constant $M \geq \frac{4}{\eta}$, λ_n is chosen from the range

$$Mc_0\sigma\widetilde{\delta_{S^c}}\sqrt{\frac{2\log(p-s)}{n}} \geq \lambda_n \geq \frac{4c_0\sigma\widetilde{\delta_{S^c}}}{\eta}\sqrt{\frac{2\log(p-s)}{n}}. \tag{2.12}$$

Furthermore, assume:

$$\widetilde{r_n} \leq \frac{1-\eta}{\widetilde{\delta}_{S^c}}, \tag{2.13}$$

and for $C_1 = \max\left\{\frac{2\widetilde{r}_n}{1-\eta}, \frac{M}{\sqrt{3}}\right\}$

$$\beta_{\min} > \max\left\{2\widetilde{\delta}_S, \frac{2\lambda_n\sqrt{s}}{\Lambda_{\min}(s)}, \frac{4c_0\sigma}{\Lambda_{\min}(s)}\sqrt{\frac{6s\log p}{n}}, C_1\widetilde{\delta}_{S^c}\right\}. \tag{2.14}$$

Then, with probability $1-\mathbb{P}(\mathcal{T}^c)-1/p^2$ *for a fixed design or* $1-\mathbb{P}((\mathcal{X}\cap\mathcal{T})^c)-1/p^2$ *for a random design respectively, the optimal solution* $\widehat{\beta}$ *to (2.2) satisfies* $\mathrm{supp}(\widehat{\beta}) = \mathrm{supp}(\beta)$.

A proof is given in Section 13. We furthermore argue below that the sets $\mathcal{T}$ and $\mathcal{X}$ (and hence also $\mathcal{T}\cap\mathcal{X}$) have large probability.

Remark 2.2.. In general, there are multiple solutions of the adaptive Lasso in (2.2). However, with high probability, the solution of (2.2) is unique. This follows from Wainwright (2008) and we present more details in Section 12.2.

Remark 2.3.. The last term on the right hand side in (2.14) usually dominates all others (under the assumptions we make for the theorem): the order of magnitude is typically $O(\sqrt{s\log(p)/n})$. Furthermore, for a fixed design, we emphasize that $\widetilde{r}_n$, $\widetilde{\delta}_S$ and $\widetilde{\delta}_{S^c}$ are only required to hold on the set $\mathcal{T}$. Similarly, for a random design, we only require some upper bounds to hold on the set $\mathcal{T}\cap\mathcal{X}$.

Remark 2.4.. We note that Theorem 2.1 suggests that we can use any initial estimator that yields a nice bound on $\|\delta\|_\infty = \|\beta_{\text{init}} - \beta\|_\infty$. We consider the Lasso as initial estimator in Sections 4 and 5. The Dantzig selector (Candès and Tao, 2007) could be an alternative having similar properties as the Lasso under the restricted eigenvalue assumptions (Bickel et al., 2008).

Lemma 2.5.. *For a fixed design, we have*

$$\mathbb{P}(\mathcal{T}) \geq 1 - 1/p^2. \tag{2.15}$$

Moreover, for a random design X *as in (2.6) with* $\Sigma_{jj} = 1, \forall j \in \{1,\ldots,p\}$, *and for* $p < e^{n/4C_2^2}$, *where* $C_2 > 4\sqrt{5/3}$, *we have*

$$\mathbb{P}(\mathcal{X}) \geq 1 - 1/p^2. \tag{2.16}$$

Hence, for a random design,

$$\mathbb{P}(\mathcal{X}\cap\mathcal{T}) \geq 1 - 2/p^2.$$

A proof is given in Section 9 (Lemmas 9.1 and 9.3).

3 Restricted eigenvalue conditions

We are analyzing in later sections the properties of the adaptive Lasso when using the standard Lasso as initial estimator:

$$\beta_{\text{init}} := \arg\min_\beta \frac{1}{2n}\|Y - X\beta\|_2^2 + \lambda_{\text{init}}\sum_{j=1}^p |\beta_j|, \tag{3.1}$$

where for some constant B and c_0 to be specified,

$$\lambda_{\text{init}} = Bc_0\sigma_\epsilon\sqrt{\frac{\log p}{n}}. \tag{3.2}$$

As usual, in order to be a sensible procedure, we assume that the different variables (columns in X) are on the same scale. In view of Theorem 2.1, we need to establish bounds for $\delta = \beta_{\text{init}} - \beta$, where β_{init} is defined in (3.1).

To derive such bounds for δ, we build upon recent work by Bickel et al. (2008) under the "restricted eigenvalue" assumptions formalized therein, which are weaker than those in Candès and Tao (2007); Meinshausen and Yu (2009) for deriving ℓ_p bounds on δ, where $p = 1, 2$, for the Dantzig selector and the Lasso respectively. Similar conditions have been used by Koltchinskii (2008) and van de Geer (2007).

3.1 Restricted eigenvalue assumption for fixed design

To introduce the first assumption, we need some more notation. For integers s, m such that $1 \le s \le p/2$ and $m \ge s, s + m \le p$, a vector $\delta \in \mathbb{R}^p$ and a set of indices $J_0 \subseteq \{1, \ldots, p\}$ with $|J_0| \le s$, denoted by J_m the subset of $\{1, \ldots, p\}$ corresponding to the m largest in absolute value coordinates of δ outside of J_0 and defined $J_{0m} \triangleq J_0 \cup J_m$.

Assumption 3.1.. Restricted eigenvalue assumption $RE(s, m, k_0, X)$ (Bickel et al., 2008). *Consider a fixed design. For some integer* $1 \le s \le p/2$, $m \ge s, s + m \le p$, *and a positive number* k_0, *the following condition holds:*

$$\frac{1}{K(s, m, k_0, X)} := \min_{\substack{J_0 \subseteq \{1,\ldots,p\}, \\ |J_0| \le s}} \min_{\substack{\gamma \ne 0, \\ \left\|\gamma_{J_0^c}\right\|_1 \le k_0 \left\|\gamma_{J_0}\right\|_1}} \frac{\|X\gamma\|_2}{\sqrt{n}\,\|\gamma_{J_{0m}}\|_2} > 0. \tag{3.3}$$

We often restrict ourselves to the case with $k_0 = 3$. Apparently, $RE(s, m, k_0, X)$ implies that $RE(s, k_0, X)$ as in Definition 3.1 below holds with $K(s, k_0, X) \le K(s, m, k_0, X)$ for the same X.

Definition 3.1.. Restricted eigenvalue definition $RE(s, k_0, X)$ (Bickel et al., 2008). *Consider a fixed design. For some integer* $1 \le s \le p$ *and a positive number* k_0, *the following condition holds:*

$$\frac{1}{K(s, k_0, X)} := \min_{\substack{J_0 \subseteq \{1,\ldots,p\}, \\ |J_0| \le s}} \min_{\substack{\gamma \ne 0, \\ \left\|\gamma_{J_0^c}\right\|_1 \le k_0 \left\|\gamma_{J_0}\right\|_1}} \frac{\|X\gamma\|_2}{\sqrt{n}\,\|\gamma_{J_0}\|_2} > 0. \tag{3.4}$$

We note that variable selection with the adaptive Lasso is possible under this weaker form of restricted eigenvalues, though with stronger conditions on the sparsity s and $\beta_{\min}$. We omit such results in this paper due to the lack of space.

By an argument in Bickel et al. (2008), it is known that if $RE(s, k_0, X)$ is satisfied with $k_0 \ge 1$, then the square submatrices of size $\le 2s$ of $X^T X/n$ are necessarily positive definite. In fact, it is clear that in (3.4), the set of admissible γ is a superset of that in (2.3). Hence we have the following:

Proposition 3.2.. *Suppose Assumption* $RE(s, k_0, X)$ *holds for* $1 \leq s \leq p/2$ *and some* $k_0 > 0$. *Then* $\Lambda_{\min}(s) \geq \frac{1}{K^2(s,k_0,X)} > 0$ *for* $\Lambda_{\min}(s)$ *as defined in* (2.3).

Note that the quantity $\Lambda_{\min}(s)$ also appears in Theorem 2.1 and hence when applying it, we make use of Proposition 3.2.

3.2 Restricted orthogonality assumption for fixed design

We also present results under a stronger design condition which covers cases where the sparsity s is allowed to be larger than in Corollary 4.4 under Assumption 3.1, see also Corollary 4.5. We define the (s, s')-restricted orthogonality constant (Candès and Tao, 2007) $\theta_{s,s'}$ for $s + s' \leq p$, which is the smallest quantity such that

$$\left| \frac{\langle X_T c, X_{T'} c' \rangle}{n} \right| \leq \theta_{s,s'} \|c\|_2 \left\|c'\right\|_2 \tag{3.5}$$

holds for all disjoint sets $T, T' \subseteq \{1, \ldots, p\}$ of cardinality $|T| \leq s$ and $|T'| < s'$.

Assumption 3.2.. **Restricted orthogonality assumption**. *Consider a fixed design. For some integer* $1 \leq s \leq p/2$, $m \geq s, s + m \leq p$, *and a positive number* k_0, *the condition* $RE(s, s, k_0, X)$ *holds. Furthermore, the following condition holds:*

$$\Lambda_{\min}(s) > 16 k_0 K^2(s, m, k_0, X) \lambda_{\text{init}} s \theta_{1,s}, \tag{3.6}$$

$$s < \frac{n}{96 c_0^2 \sigma^2 K^2(s, s, k_0, X) \log p}, \tag{3.7}$$

where $k_0 \leq 3$.

With such a restriction on the sparsity, we note that (3.6) is a weaker condition than Assumption 3 in Bickel et al. (2008). We assume that (3.6) holds with a constant that is smaller than $2k_0$ as in Assumption 3 of (Bickel et al., 2008), which by itself is a sufficient condition to derive Assumption 3.1.

We refer to Bickel et al. (2008) for more detailed discussions about these assumptions which are weaker than those in Candès and Tao (2007); Meinshausen and Yu (2009) and arguably less restrictive than those in Meinshausen and Bühlmann (2006),
Zhao and Yu (2006) or Wainwright (2008).

4 The adaptive Lasso with fixed design

We first show that the restricted eigenvalue condition ensures to derive upper bounds on the ℓ_∞-norms of $\delta := \beta_{\text{init}} - \beta$,

Lemma 4.1.. *Suppose that condition* $RE(s, 3, X)$ *holds for a fixed design and suppose that*

$$\beta_{\min} \quad \geq \quad 8 K^2(s, 3, X) \lambda_{\text{init}} \sqrt{s}, \tag{4.1}$$

for λ_{init} that satisfies (3.2). Then, the initial estimator (3.1) in model (1.1) guarantees that on the set $\mathcal{T}$ as in (2.5),

$$\|\delta_S\|_\infty \leq 4K^2(s,3,X)\lambda_{\text{init}}\sqrt{s}, \text{ and} \tag{4.2a}$$

$$\|\delta_{S^c}\|_\infty \leq 3K^2(s,3,X)\lambda_{\text{init}}s \tag{4.2b}$$

Suppose that Assumption $RE(s,s,3,X)$ and (4.1) hold. Then on the set $\mathcal{T}$ as in (2.5), (4.2a) holds, while (4.2b) is replaced by

$$\|\delta_{S^c}\|_\infty \leq 16K^2(s,s,3,X)\lambda_{\text{init}}\sqrt{s}. \tag{4.3}$$

A proof is given in Subsection 10.2. Lemma 4.1 leads to the upper bounds $\widetilde{\delta}_S = 4K^2(s,3,X)\lambda_{\text{init}}\sqrt{s}$ and $\widetilde{\delta}_{S^c} = 16K^2(s,s,3,X)\lambda_{\text{init}}\sqrt{s}$. When using these bounds in Theorem 2.1, we see that the range for the regularization parameter in 2.12 depends on the unknown sparsity s. This unpleasant situation can be improved by estimating s using a thresholding procedure as follows.

Lemma 4.2.. Thresholding procedure. *Let the assumptions of Lemma 4.1 hold. Consider the set $\bar{S}$ that includes all $\beta_{j,\text{init}}$ for $j \in \{1,\ldots,p\}$, whose absolute values are larger than $4\lambda_{\text{init}}$. Let $\bar{s} := |\bar{S}|$ be an estimate which is in the same order as the true sparsity s. More specifically, we have, on the set $\mathcal{T}$ in (2.5),*

$$S \subseteq \bar{S} \text{ and } \quad s \leq |\bar{S}| \leq sK^2(s,3,X) \quad \text{for} \quad K \geq 2. \tag{4.4}$$

A proof of Lemma 4.2 is given in Subsection 10.3.

The range for the tuning parameter λ is now specified as follows. For some constant $\frac{4K(s,s,k_0)}{\eta} \leq M \leq \frac{\sqrt{\Lambda_{\min}(s)}}{(1-\eta)c_0\sigma}\sqrt{\frac{n}{2\log p}}$, where $0<\eta<1$, λ_n is chosen such that

$$16MK(s,s,k_0) \geq \frac{\lambda_n}{c_0\sigma\lambda_{\text{init}}\sqrt{\bar{s}}}\sqrt{\frac{n}{2\log(p-s)}} \geq \frac{64K^2(s,s,k_0)}{\eta}, \tag{4.5}$$

where λ_{init} is defined in (3.2) with $B=\sqrt{24}$ and $c_0 \geq 1$ is a small constant to be specified. The following theorem is an immediate result when we substitute $\widetilde{\delta}_{S^c}$ and $\widetilde{\delta}_S$ that appear in Theorem 2.1 with what we derived in Lemma 4.1.

Theorem 4.3.. (Variable selection for fixed design) *Consider the linear model in (1.1) with fixed design X, where $n \leq p$, and each column of X has its ℓ_2-norm upper bounded by $\sqrt{n}$. Suppose condition $RE(s,s,3,X)$ (Assumption 3.1) holds. Suppose on $\mathcal{T}$, for some $1>\eta>0$, λ_n is chosen as in (4.5) with $K(s,s,k_0)=K(s,s,3,X)$ and $c_0=1$. Suppose s satisfies (3.7) and*

$$\widetilde{r_n}\sqrt{s} \leq \frac{1-\eta}{32K^2\lambda_{\text{init}}}, \quad \text{and} \tag{4.6}$$

$$\beta_{\min} > \max\left\{\frac{2\widetilde{r_n}}{1-\eta}, \frac{M}{\sqrt{3}}\right\} 16K^2\lambda_{\text{init}}\sqrt{s} \tag{4.7}$$

where λ_{init} is defined in (3.2) with $B=\sqrt{24}$ and $K=K(s,s,3,X)$. Then, with probability $1-2/p^2$, the adaptive estimator in (2.2) satisfies $\text{supp}(\widehat{\beta}) = \text{supp}(\beta)$.

A proof is given in Section 10.4. A first corollary follows immediately from Theorem 4.3 when we substitute $\widetilde{r}_n = \frac{\sqrt{s}}{\sqrt{\Lambda_{\min}(s)}}$ as shown in Lemma 10.3, formula (10.16) with $c_0 = 1$.

Corollary 4.4.. (Variable selection for fixed design: general bound for $\widetilde{r}_n$) *Consider the linear model in (1.1) with fixed design X, where $n \leq p$, and each column of X has its ℓ_2-norm upper bounded by $\sqrt{n}$. Suppose that condition $RE(s, s, 3, X)$ (Assumption 3.1) holds. Suppose that on $\mathcal{T}$ and for some $1 > \eta > 0$, λ_n is chosen as in (4.5) with $K(s, s, k_0) = K(s, s, 3, X)$ and $c_0 = 1$,*

$$s \leq \frac{\sqrt{\Lambda_{\min}(s)}(1-\eta)}{32K^2\lambda_{\text{init}}} \quad \text{and} \tag{4.8}$$

$$\beta_{\min} > \max\left\{\frac{2\sqrt{s}}{(1-\eta)\sqrt{\Lambda_{\min}(s)}}, \frac{M}{\sqrt{3}}\right\} 16K^2\lambda_{\text{init}}\sqrt{s} \tag{4.9}$$

where λ_{init} is defined in (3.2) with $B = \sqrt{24}$ and $K = K(s, s, 3, X)$. Then, with probability $1 - 2/p^2$, the adaptive estimator in (2.2) satisfies $\text{supp}(\widehat{\beta}) = \text{supp}(\beta)$.

Using the different bound $\widetilde{r}_n = \frac{\theta_{s,1}\sqrt{s}}{\Lambda_{\min}(s)}$ from Lemma 10.3, formula (10.17), our next corollary shows that under Assumption 3.2, we can essentially achieve the sublinear sparsity level of (3.7) while conducting model selection.

Corollary 4.5.. (Variable selection for fixed design: special bound for $\widetilde{r}_n$) *Consider the linear model in (1.1) with fixed design X, where $n \leq p$, and each column of X has ℓ_2-norm upper bounded by $\sqrt{n}$. Suppose that Assumption 3.2 holds for $k_0 = 3$ and $m = s$. Suppose that on $\mathcal{T}$ and for some $1 > \eta > 0$, λ_n is chosen as in (4.5) with $K(s, s, k_0) = K(s, s, 3, X)$ and $c_0 = 1$. Suppose s satisfies (3.7) and*

$$\beta_{\min} > \max\left\{\frac{2\sqrt{s}\theta_{1,s}}{(1-\eta)\Lambda_{\min}(s)}, \frac{M}{\sqrt{3}}\right\} 16K^2\lambda_{\text{init}}\sqrt{s} \tag{4.10}$$

where λ_{init} is defined in (3.2) with $B = \sqrt{24}$ and $K = K(s, s, 3, X)$. Then, with probability $1 - 2/p^2$, the adaptive estimator in (2.2) satisfies $\text{supp}(\widehat{\beta}) = \text{supp}(\beta)$.

It is an open question whether the adaptive Lasso procedure can achieve model selection consistency under such sparsity level under Assumption 3.1 alone.

5 The adaptive Lasso with random design

For a random design X as in (2.6), we make the following assumption on Σ.

Assumption 5.1.. Restricted eigenvalue assumption $RE(s, m, k_0, \Sigma)$ *For some integer $1 \leq s \leq p/2$, $m \geq s, s + m \leq p$, and a positive number k_0, the following condition holds:*

$$\frac{1}{K(s, m, k_0, \Sigma)} := \min_{\substack{J_0 \subseteq \{1,\ldots,p\}, \\ |J_0| \leq s}} \min_{\substack{\gamma \neq 0, \\ \|\gamma_{J_0^c}\|_1 \leq k_0 \|\gamma_{J_0}\|_1}} \frac{\|\Sigma^{1/2}\gamma\|_2}{\|\gamma_{J_{0m}}\|_2} > 0. \tag{5.1}$$

Suppose (2.8) hold and $\Sigma_{jj} = 1, \forall j = 1, \ldots, p$.

It is clear that in (5.1), the set of admissible γ is a superset of that in (2.7). Hence we have:

Proposition 5.1.. *Suppose Assumption $RE(s, s, k_0, \Sigma)$ holds for some $1 \le s \le p$ and some $k_0 > 0$. Then $\frac{17\Lambda_{\min}(s)}{16} \ge \frac{1}{K^2(s,s,k_0,\Sigma)}$ for $\Lambda_{\min}(s)$ as defined in (2.7).*

We now show that with high probability, Assumption $RE(s, m, k_0, X)$ holds for a random realization of X whose row are i.i.d. vectors from $\sim N(0, \Sigma)$, under Assumption 5.1, if $s = o\left(\sqrt{\frac{n}{\log p}}\right)$.

Proposition 5.2.. *Consider a random design X as in (2.6). Assume that Σ satisfies (5.1). Then, on the set $\mathcal{X}$ as defined in (2.9) and with C_2 as in (2.9), X satisfies $RE(s, s, k_0, X)$ as in Assumption 3.1, with*

$$K(s, s, k_0, X) \le \sqrt{2}K(s, s, k_0, \Sigma), \text{ for } s \le \frac{\sqrt{n/\log p}}{32C_2K^2(s, 3, 3, \Sigma)} \tag{5.2}$$

Its proof appears in Subsection 11.1.

We can now state the result for a random design under Assumption 5.1.

Theorem 5.3.. (Variable selection for a random design) *Consider the linear model in (1.1) with random design X as in (2.6) with $n \le p$ and $p < e^{n/4C_2^2}$, where $C_2 > 4\sqrt{5/3}$. Suppose that Assumption 5.1 holds with $m = s$ and $k_0 = 3$. Suppose that on the set $\mathcal{X} \cap \mathcal{T}$ and for some $0 < \eta < 1$, λ_n is chosen as in (4.5) with $K(s, s, k_0) = \sqrt{2}K(s, s, 3, \Sigma)$ and $c_0 = \sqrt{3/2}$; suppose that*

$$s \le \frac{1}{32K^2(s, s, 3, \Sigma)} \min\left\{\frac{1}{C_2}, \frac{\sqrt{\Lambda_{\min}(s)}(1-\eta)}{6\sqrt{6}\sigma}\right\}\sqrt{\frac{n}{\log p}} \tag{5.3}$$

where C_2 is defined in (2.9) In addition $\beta_{\min}$ satisfies (4.9) with $K = \sqrt{2}K(s, s, 3, \Sigma)$. Then, with probability $1 - 3/p^2$, the adaptive Lasso estimator in (2.2) satisfies $\text{supp}(\widehat{\beta}) = \text{supp}(\beta)$.

A proof is given in Section 11.3.

6 The adaptive Lasso in Gaussian graphical modeling

Consider the problem of covariance selection described in Section 1.2.

6.1 The many regressions pursuit procedure

The procedure for covariance selection in a Gaussian graphical model based on a pursuit of many regressions has been proposed and studied in Meinshausen and Bühlmann (2006).

Consider $X_1, \ldots, X_p \sim \mathcal{N}(0, \Sigma)$ as in (1.4). We can regress X_i versus the other variables $\{X_k;\ k \ne i\}$:

$$X_i = \sum_{j \ne i} \beta_j^i X_i + V_i \tag{6.1}$$

where V_i is a normally distributed random variable with mean zero. Then, denoting by $Q = \Sigma^{-1}$, it is well known that

$$\beta_j^i = -\frac{Q_{ij}}{Q_{ii}}. \tag{6.2}$$

In particular, this implies that

$$\begin{aligned} &\text{there is an undirected edge between } i \text{ and } j \\ \Leftrightarrow \quad &\Sigma^{-1}_{ij} \neq 0 \;\Leftrightarrow\; \beta_j^i \neq 0 \text{ and/or } \beta_i^j \neq 0, \end{aligned}$$

where the last statement holds due to the symmetry of Σ^{-1}.

The estimation of the edge set can then be done by one of the following rules:

$$\begin{aligned} &\text{there is an edge between } i \text{ and } j \;\Leftrightarrow\; \widehat{\beta}_j^i \neq 0 \text{ and } \widehat{\beta}_i^j \neq 0, \\ &\text{there is an edge between } i \text{ and } j \;\Leftrightarrow\; \widehat{\beta}_j^i \neq 0 \text{ or } \widehat{\beta}_i^j \neq 0. \end{aligned}$$

Our obvious proposal is to use the adaptive Lasso estimates $\widehat{\beta}^i_{j;n}$ in the corresponding regressions as described in (6.1). The discrepancy between the "and" or "or" rule above vanishes with high probability.

The theoretical analysis follows by our result for random design linear models (Theorem 5.3) and controlling the error over p different regressions. Let $\beta_{\min} = \min_{i,j} |\beta_j^i|$ and s be the largest node degree. Our conditions on sparsity and $\beta_{\min}$ for linear models need to hold for all p regressions simultaneously and they are as follows.

Assumption 6.1.. *β_j^i from (6.1) satisfy the conditions on $\beta_{\min}$ as in (4.9) $\forall i, j \in \{1, \ldots, p\}$ under Assumption 5.1.*

Equivalently, by assuming $\Sigma^{-1}_{jj} = 1$ for all $j = 1, \ldots, p$ (see Assumption 5.1) and due to (6.2), the non-zero elements of $|\Sigma^{-1}_{ij}|$ are required to be upper-bounded by the value of $\beta_{\min}$.

Assumption 6.2.. *The covariance matrix Σ satisfies the restricted eigenvalue condition in Assumption 5.1. In addition, (2.8) is required to hold on every subset $S \subset \{1, \ldots, p\}$ such that $|S| \leq s$.*

Assumption 6.3.. *The size of the neighborhood, for all nodes, is bounded by an integer $1 < s < p/2$ that satisfies (5.3) under Assumption 5.1.*

The following result can then be immediately derived using the union bound for the p regression in the many regressions pursuit.

Theorem 6.1.. **(Covariance selection in Gaussian Graphical Models)** *Consider the Gaussian graphical model with n i.i.d. samples from (1.4), where $n \leq p < e^{n/4C_2^2}$, where $C_2 > 4\sqrt{5/3}$. Suppose that Assumptions 6.1 - 6.3 hold. Then,*

$$\mathbb{P}\left(\mathrm{supp}(\widehat{\Sigma}_n^{-1}) = \mathrm{supp}(\Sigma^{-1})\right) \geq 1 - 3/p.$$

7 Discussion

We have presented results for high-dimensional model selection in regression and Gaussian graphical modeling. We make some assumptions on (fixed or random) designs in terms of restricted eigenvalues. Such assumptions are among the weakest for deriving oracle inequalities in terms of $\|\widehat{\beta} - \beta\|_q$ $(q = 1, 2)$ (Bickel et al., 2008). We show here that under such restricted eigenvalue assumptions, the two-stage adaptive Lasso is able to correctly infer the relevant variables in regression or the edge set in a Gaussian graphical model. The ordinary Lasso can easily fail since the neighborhood stability condition, or the equivalent irrepresentable condition, are necessary and sufficient (Meinshausen and Bühlmann, 2006; Zhao and Yu, 2006). It is easy to construct examples where the neighborhood stability condition fails but the restricted eigenvalue condition holds for the situation where $n > p$, see for example Zou (2006).

In the high-dimensional context, the relation between the neighborhood stability condition and the restricted eigenvalue assumption is not clear. However, the latter is a condition on an average behavior (as an eigenvalue condition) while the former requires a relation for a maximum: thus, we conjecture that the restricted eigenvalue assumption is in general less restrictive than the neighborhood stability condition. In particular, although it appears non-trivial to derive a general relation between these two conditions, one can certainly derive relations between them under additional assumptions; A thorough exposition of such relations is an interesting direction for future work, given the frequent appearance of both types of conditions in the literature, for example in Meinshausen and Bühlmann (2006); Zhao and Yu (2006); Wainwright (2008); Candès and Tao (2007); Meinshausen and Yu (2009); Bickel et al. (2008). For high-dimensional Gaussian graphical modeling, using the reasoning above, the restricted eigenvalue assumptions we make appears in general less restrictive (and easier to check) than the assumptions in Meinshausen and Bühlmann (2006) and in Ravikumar et al. (2008) who analyze the GLasso algorithm Banerjee et al. (2008); Friedman et al. (2008b).

8 Analysis of the weighted Lasso

In the sequel, for clarity, we denote by β^* the true parameter in the linear model (1.1). Inspired by the adaptive Lasso estimator defined in (2.2), we consider here the weighted Lasso with weights $0 < w_j$ $(j = 1, \ldots, p)$ which solves the following optimization problem:

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{2n} \|Y - X\beta\|_2^2 + \lambda_n \sum_{j=1}^{p} w_j |\beta_j|, \tag{8.1}$$

The only distinction between the adaptive and weighted Lasso is that we assume that the weights are estimated in the former and pre-specified in the latter approach. However, our theory below though does not depend whether the weights are random or not. For convenience we denote by

$$w_{\max}(S) = \max_{i \in S} w_i, \qquad w_{\min}(S^c) = \min_{j \in S^c} w_j. \tag{8.2}$$

A slightly stronger notion than inferring the support of β^* is the recovery of the sign-pattern:

$$\text{sgn}(\widehat{\beta}_n) = \text{sgn}(\beta^*).$$

Furthermore, there are generally multiple solutions of the adaptive Lasso estimator in (2.2) and in the weighted Lasso in (8.1). However, with high probability, the solution is unique, see also Remark 2.2 and Section 12.2.

As before, we denote by $\|A\|_\infty = \max_{1\le i\le k}\sum_{j=1}^m |A_{ij}|$ for a $k\times m$ matrix A. First, let us state the following conditions that are imposed on the design matrix for the ordinary Lasso by Zhao and Yu (2006) and Wainwright (2008):

$$\left\|X_{S^c}^T X_S (X_S^T X_S)^{-1}\right\|_\infty \le 1-\eta, \quad \text{for some } \eta\in(0,1], \text{ and} \tag{8.3a}$$

$$\Lambda_{\min}\left(\tfrac{1}{n}X_S^T X_S\right) \ge \Lambda_{\min}(s) > 0, \tag{8.3b}$$

where $\Lambda_{\min}(A)$ is the smallest eigenvalue of A. Note that the second condition coincides with ours in (2.4). Meinshausen and Bühlmann (2006) formulated such conditions for a random design.

We impose the following incoherence conditions on the weighted Lasso.

Definition 8.1.. ($(\vec{w},S)$-incoherence condition) *Let X be an $n\times p$ matrix and let $S\subset\{1,\dots,p\}$ be nonempty. Let $\vec{w}=(w_1,w_2,\dots,w_p)^T$ be a weight vector, where $w_j>0\forall j$. Let $\vec{b}=(\mathrm{sgn}(\beta_i^*)w_i)_{i\in S}$. We say that X is $(\vec{w},S)$-incoherent if for some $\eta\in(0,1)$,*

$$\forall j\in S^c, \qquad \left|X_j^T X_S (X_S^T X_S)^{-1}\vec{b}\right| \le w_j(1-\eta), \tag{8.4a}$$

$$\Lambda_{\min}\left(\tfrac{1}{n}X_S^T X_S\right) \ge \Lambda_{\min}(s) > 0, \tag{8.4b}$$

where a sufficient condition for (8.4a) is

$$\forall j\in S^c, \qquad \left\|X_{S^c}^T X_S (X_S^T X_S)^{-1}\right\|_\infty \le \frac{w_{\min}(S^c)}{w_{\max}(S)}(1-\eta). \tag{8.5}$$

We now state a general lemma about recovering the signs for the weighted Lasso estimator as defined in (2.2).

Lemma 8.2.. (Sign recovery Lemma) *Consider the linear model in (1.1) where the design matrix X satisfies (8.4a) and (8.4b). Let $c_0=\max_{j\in S^c}\|X_j\|_2/\sqrt{n}$. Suppose that $w_j>0,\forall j=1,\dots,p$ and λ_n is chosen such that*

$$\lambda_n w_{\min}(S^c) \ge \frac{4c_0\sigma}{\eta}\sqrt{\frac{2\log(p-s)}{n}},$$

where $w_{\min}(S^c), w_{\max}(S)$ are as defined in (8.2). Furthermore, assume

$$\beta_{\min} > \max\left\{\frac{4c_0\sigma}{\Lambda_{\min}(s)}\sqrt{\frac{6s\log p}{n}}, \frac{2\lambda_n w_{\max}(S)\sqrt{s}}{\Lambda_{\min}(s)}\right\} \tag{8.6}$$

Then for $\widehat{\beta}$ in (8.1):

$$\mathbb{P}\left(\mathrm{sgn}(\widehat{\beta})=\mathrm{sgn}(\beta^*)\right)\ge 1-2/p^2,$$

Moreover, with $\mathcal{T}$ defined in (2.5), we have $\mathbb{P}\left((\mathrm{sgn}(\widehat{\beta})\ne\mathrm{sgn}(\beta^*))\cap\mathcal{T}\right)\le 2/p^2$.

A proof is given in Section 12.3. Note that in case $w_{\min}(S^c)=w_{\max}(S)=1$, conditions (8.5) and (8.6) reduce to (8.3a) and the the statement of Lemma 8.2 is exactly the same as Theorem 1 in Wainwright (2008).

9 Proof of Lemma 2.5

Lemma 9.1. *For fixed design X with $\max_j \|X_j\|_2 \leq c_o\sqrt{n}$ we have for $\mathcal{T}$ as defined in (2.5),*

$$\mathbb{P}\left(\mathcal{T}^c\right) \leq 1/p^2. \tag{9.1}$$

Proof. Define the random variables

$$Y_j = \frac{1}{n}\sum_{i=1}^{n} \epsilon_i X_{i,j}.$$

Note that $\max_{1\leq j\leq p}|Y_j| = \|X^T\epsilon/n\|_\infty$. We have $\mathbb{E}\left(Y_j\right) = 0$ and $\mathsf{Var}(Y_j) = \frac{\|X_j\|_2^2\sigma_\epsilon^2}{n^2} \leq \frac{c_0\sigma_\epsilon^2}{n}$. Obviously, Y_j has its tail probability dominated by that of $Z \sim N(0, \frac{c_0^2\sigma_\epsilon^2}{n})$:

$$\mathbb{P}\left(|Y_j| \geq t\right) \leq \mathbb{P}\left(|Z| \geq t\right) \leq \frac{c_0\sigma_\epsilon}{\sqrt{n}t}\exp\left(\frac{-nt^2}{2c_0^2\sigma_\epsilon^2}\right).$$

We can now apply the union bound to obtain:

$$\begin{aligned}\mathbb{P}\left(\max_{1\leq j\leq p}|Y_j| \geq t\right) &\leq p\frac{c_0\sigma_\epsilon}{\sqrt{n}t}\exp\left(\frac{-nt^2}{2c_0^2\sigma_\epsilon^2}\right)\\ &= \exp\left(-\left(\frac{nt^2}{2c_0^2\sigma_\epsilon^2} + \log\frac{t\sqrt{n}}{c_0\sigma_\epsilon} - \log p\right)\right).\end{aligned}$$

By choosing $t = c_o\sigma_\epsilon\sqrt{6\log(p)/n}$, the right-hand side is bounded by $1/p^2$. □

We now show that $\mathbb{P}\left(\mathcal{X}\right) \geq 1 - 1/p^2$.

We denote $\Sigma_{ii} := \sigma_i^2$ throughout the rest of this proof. We first state the following large inequality bound for the nondiagonal entries of Σ, adapted from Lemma 38 (Zhou et al., 2008) by plugging in $\sigma_i^2 = 1, \forall i = 1,\ldots,p$ and using the fact that $|\Sigma_{jk}| = |\rho_{jk}\sigma_j\sigma_k| \leq 1, \forall j \neq k$, where ρ_{jk} is the correlation coefficient between variables X_j and X_k.

Lemma 9.2. (Zhou et al., 2008) *Let $\Psi_{jk} = (1+\Sigma_{jk}^2)/2$. For $0 \leq \tau \leq \Psi_{jk}$,*

$$\mathbb{P}\left(|\Delta_{jk}| > \tau\right) \leq \exp\left\{-\frac{3n\tau^2}{10(1+\Sigma_{jk}^2)}\right\} \leq \exp\left\{-\frac{3n\tau^2}{20}\right\}. \tag{9.2}$$

We now also state a large deviation bound for the χ_n^2 distribution Johnstone (2001):

$$\mathbb{P}\left(\frac{\chi_n^2}{n} - 1 > \tau\right) \leq \exp\left(\frac{-3n\tau^2}{16}\right), \text{ for } 0 \leq \tau \leq \frac{1}{2}. \tag{9.3}$$

Hence by the union bound, we have $j = 1,\ldots,p$, for $\tau < 1/2$,

$$\mathbb{P}\left(\max_{j=1,\ldots,p}\frac{\|X_j\|_2^2}{n} - 1 > \tau\right) \leq p\exp\left(\frac{-3n\tau^2}{16}\right). \tag{9.4}$$

Lemma 9.3. *For a random design X as in (2.6) with $\Sigma_{jj} = 1, \forall j \in \{1, \ldots, p\}$, and for $p < e^{n/4C_2^2}$, where $C_2 > 4\sqrt{5/3}$, we have*

$$\mathbb{P}(\mathcal{X}) \geq 1 - 1/p^2.$$

Proof. Now it is clear that we have $p(p-1)/2$ unique non-diagonal entries $\sigma_{jk}, \forall j \neq k$ and p diagonal entries. By the union bound and by taking $\tau = C_2\sqrt{\frac{\log p}{n}}$ in (9.4) and (9.2), we have

$$\begin{aligned}
\mathbb{P}(\mathcal{X}^c) &= \mathbb{P}\left(\max_{jk} |\Delta_{jk}| \geq C_2\sqrt{\frac{\log p}{n}}\right) \\
&\leq p\exp\left(-\frac{3C_2^2 \log p}{16}\right) + \frac{p^2 - p}{2}\exp\left(-\frac{3C_2^2 \log p}{20}\right) \\
&\leq p^2 \exp\left(-\frac{3C_2^2 \log p}{20}\right) = p^{-\frac{3C_2^2}{20}+2} < \frac{1}{p^2}
\end{aligned}$$

for $C_2 > 4\sqrt{5/3}$. Finally, $p < e^{n/4C_2^2}$ guarantees that $C_2\sqrt{\frac{\log p}{n}} < 1/2$. □

10 Proofs for Section 4

Throughout this section, we have $\lambda_{\text{init}} = Bc_0\sigma_\epsilon\sqrt{\frac{\log p}{n}}$ with $B = \sqrt{24}$.

10.1 The Lasso as initial estimator

Lemma 4.1 crucially uses the bound on the ℓ_1-loss of the initial Lasso estimator.

Our proof follows that of Bickel et al. (2008). Let β_{init} be as in (3.1) and $\delta = \beta_{\text{init}} - \beta^*$. The set $\mathcal{T}$ is defined in (2.5). We first show Lemma 10.1; we then apply condition $RE(s, k_0, X)$ on δ with $k_0 = 3$ under $\mathcal{T}$ to derive various norm bounds.

Lemma 10.1. *For fixed design, on $\mathcal{T}$, $\|\delta_{S^c}\|_1 \leq 3\|\delta_S\|_1$.*

Proof. Since β_{init} is a Lasso solution, we have

$$\begin{aligned}
\lambda_{\text{init}}\|\beta^*\|_1 - \lambda_{\text{init}}\|\beta_{\text{init}}\|_1 &\geq \frac{1}{2n}\|Y - X\beta_{\text{init}}\|_2^2 - \frac{1}{2n}\|Y - X\beta^*\|_2^2 \\
&\geq \frac{1}{2n}\|X\delta\|_2^2 - \frac{\delta^T X^T \epsilon}{n}
\end{aligned}$$

Hence on the set $\mathcal{T}$ as in (2.5), we have

$$\begin{aligned}
\|X\delta\|_n^2 &\leq 2\lambda_{\text{init}}\|\beta^*\|_1 - 2\lambda_{\text{init}}\|\beta_{\text{init}}\|_1 + 2\left\|\frac{X^T\epsilon}{n}\right\|_\infty \|\delta\|_1 \\
&\leq \lambda_{\text{init}}\left(2\|\beta^*\|_1 - 2\|\beta_{\text{init}}\|_1 + \|\delta\|_1\right), \qquad (10.1)
\end{aligned}$$

where by the triangle inequality, and $\beta^*_{S^c} = 0$, we have

$$\begin{aligned} 0 &\leq 2\|\beta^*\|_1 - 2\|\beta_{\text{init}}\|_1 + \|\delta\|_1 \\ &= 2\|\beta^*_S\|_1 - 2\|\beta_{S,\text{init}}\|_1 - 2\|\delta_{S^c}\|_1 + \|\delta_S\|_1 + \|\delta_{S^c}\|_1 \\ &\leq 3\|\delta_S\|_1 - \|\delta_{S^c}\|_1 . \end{aligned} \tag{10.2}$$

Thus Lemma 10.1 holds. □

Proposition 10.2.. (**ℓ_p-loss for the initial estimator**, (Bickel et al., 2008)) *Consider the linear model in (1.1) with fixed design satisfying* $\max_j \|X_j\|_2 \leq c_0\sqrt{n}$. *Suppose that* $RE(s,3,X)$ *holds. Let* $\delta = \beta_{\text{init}} - \beta^*$ *with* β_{init} *defined in* (3.1) *with*

$$\lambda_{\text{init}} = Bc_0\sigma_\epsilon\sqrt{\frac{\log p}{n}}.$$

Then, on the set $\mathcal{T}$ *in* (2.5),

$$\|\delta_S\|_2 \leq 4K^2(s,3,X)\lambda_{\text{init}}\sqrt{s}. \tag{10.3}$$

$$\|\delta\|_1 \leq 4K^2(s,3,X)\lambda_{\text{init}}s; \tag{10.4}$$

Moreover, under the stronger assumption $RE(s,s,3,X)$, *and on the set* $\mathcal{T}$ *as in* (2.5),

$$\|\delta\|_2 \leq 16K^2(s,s,3,X)\lambda_{\text{init}}\sqrt{s}. \tag{10.5}$$

Proof. On the set $\mathcal{T}$, by (10.1) and (10.2),

$$\begin{aligned} \|X\delta\|_n^2 + \lambda_{\text{init}}\|\delta\|_1 &\leq \lambda_{\min}\left(3\|\delta_S\|_1 - \|\delta_{S^c}\|_1 + \|\delta_S\|_1 + \|\delta_{S^c}\|_1\right) \\ &= 4\lambda_{\text{init}}\|\delta_S\|_1 \leq 4\lambda_{\text{init}}\sqrt{s}\,\|\delta_S\|_2 \qquad (10.6) \\ &\leq 4\lambda_{\text{init}}\sqrt{s}K(s,3,X)\|X\delta\|_n \qquad (10.7) \\ &\leq 4K^2(s,3,X)\lambda_{\text{init}}^2 s + \|X\delta\|_n^2 , \end{aligned}$$

where (10.7) is due to condition $RE(s,3,X)$ and Lemma 10.1. Hence (10.4) holds. Now by $RE(s,3,X)$ and (10.6), we have

$$\|\delta_S\|_2^2 \leq K^2(s,3,X)\|X\delta\|_n^2 \leq K^2(s,3,X)4\lambda_{\text{init}}\sqrt{s}\,\|\delta_S\|_2 . \tag{10.8}$$

Hence (10.3) holds. Finally, on the set $\mathcal{T}$, given Lemma 10.1, by $RE(s,s,3,X)$ and (10.6), we have

$$\begin{aligned} \|\delta_{SS'}\|_2^2 &\leq K^2(s,s,3,X)\|X\delta\|_n^2 \\ &\leq K^2(s,s,3,X)4\lambda_{\text{init}}\sqrt{s}\,\|\delta_S\|_2 \\ &\leq K^2(s,s,3,X)4\lambda_{\text{init}}\sqrt{s}\,\|\delta_{SS'}\|_2 . \end{aligned}$$

Hence from the following inequality (10.9) (e.g., cf. (B.28) in Bickel et al. (2008))

$$\|\delta\|_2 \leq (1+k_0)\|\delta_{SS'}\|_2 , \tag{10.9}$$

we obtain (10.5). □

10.2 Proof of Lemma 4.1

By Proposition 10.2, and (B.26) in Bickel et al. (2008),

$$\begin{aligned} \|\delta_S\|_2 &\leq 4K(s,3,X)^2\lambda_{\text{init}}\sqrt{s}, \\ \|\delta\|_1 &\leq 4K(s,3,X)^2\lambda_{\text{init}}s, \quad \text{where} \\ \|\delta_{S^c}\|_1 &\leq 3\|\delta\|_1, \end{aligned}$$

due to a property of the Lasso estimator (see, for example Bickel et al. (2008)). This allows us to conclude that on the set $\mathcal{T}$ as in (2.5),

$$\|\delta_S\|_\infty \leq \|\delta_S\|_2 \leq 4K(s,3,X)^2\lambda_{\text{init}}\sqrt{s}, \tag{10.10}$$

$$\|\delta_{S^c}\|_1 \leq \frac{3}{4}\|\delta\|_1 \leq 3K(s,3,X)^2\lambda_{\text{init}}s. \tag{10.11}$$

Thus we have by (4.1), (10.10) and (10.11),

$$\forall\, i \in S, \quad |\beta_{i,\text{init}}| \geq \beta_{\min} - \|\delta_S\|_\infty \geq 4K(s,3,X)^2\lambda_{\text{init}}\sqrt{s}, \tag{10.12}$$

$$\forall j \in S^c, \quad |\beta_{j,\text{init}}| \leq \|\delta_{S^c}\|_\infty \leq \|\delta_{S^c}\|_1 \leq 3K(s,3,X)^2\lambda_{\text{init}}s. \tag{10.13}$$

□

10.3 Proof of Lemma 4.2

If we threshold β_{init} at the value of $4\lambda_{\text{init}}$, by (10.12), we have $\bar{S} \supseteq S$. Moreover, by (10.11), we include at most $3K(s,3,X)^2 s/4$ more entries from S^c in $\bar{S}$; thus for $K(s,3,X) \geq 2$,

$$s \leq |\bar{S}| \leq s + \frac{3sK(s,3,X)^2}{4} \leq sK(s,3,X)^2.$$

In addition, we have $\forall j \in S^c$, by (10.5),

$$\begin{aligned} |\beta_{j,\text{init}}| &\leq \|\delta_{S^c}\|_\infty \leq \|\delta_{S^c}\|_2 \\ &\leq 16K^2(s,s,3,X)\lambda_{\text{init}}\sqrt{s}, \end{aligned}$$

under Assumption $RE(s,s,3,X)$ and condition $\mathcal{T}$. □

10.4 Proof of Theorem 4.3

It is clear that once we finish checking conditions on λ_n in (2.12), on s as in (2.13) and on $\beta_{\min}$ as in (2.14) hold, we can invoke Theorem 2.1 to finish the proof. Formula (4.1) is satisfied assuming (4.9). Hence by choosing

$$\widetilde{\delta}_S := 4K^2(s,3,X)\lambda_{\text{init}}\sqrt{s}, \tag{10.14}$$

$$\widetilde{\delta}_{S^c} := 16K^2(s,s,3,X)\lambda_{\text{init}}\sqrt{s}, \tag{10.15}$$

we have $\widetilde{\delta}_S \geq \|\delta_S\|_\infty$ and $\widetilde{\delta}_{S^c} \geq \|\delta_{S^c}\|_\infty$ by (4.2a) and (4.3). Now by (4.4),

$$\begin{aligned}
\lambda_n &\geq \frac{64\sigma K^2(s,s,3,X)\lambda_{\text{init}}\sqrt{|\bar{S}|}}{\eta}\sqrt{\frac{2\log(p-s)}{n}} \\
&\geq \frac{4\sigma}{\eta}\sqrt{\frac{2\log(p-s)}{n}}16K^2(s,s,3,X)\lambda_{\text{init}}\sqrt{s} \\
&= \frac{4\sigma\widetilde{\delta}_{S^c}}{\eta}\sqrt{\frac{2\log(p-s)}{n}}
\end{aligned}$$

and

$$\begin{aligned}
\lambda_n &\leq \frac{16MK^2(s,3,3,X)\sigma\lambda_{\text{init}}\sqrt{|\bar{S}|}}{K(s,3,X)}\sqrt{\frac{2\log(p-s)}{n}} \\
&\leq M\sigma 16K^2(s,3,3,X)\lambda_{\text{init}}\sqrt{s}\sqrt{\frac{2\log(p-s)}{n}} \\
&= M\sigma\widetilde{\delta}_{S^c}\sqrt{\frac{2\log(p-s)}{n}},
\end{aligned}$$

and thus (2.12) holds with $c_0 = 1$. Furthermore, for the sparsity s, (4.6) guarantees that (2.13) holds by (10.15). Finally, regarding $\beta_{\min}$, (2.14) holds given (4.7), as $\frac{16MK^2\lambda_{\text{init}}\sqrt{s}}{\sqrt{3}}$ clearly dominates the first and the third term in (2.14) by the definition of (10.14) and (10.15), and the fact that $\frac{1}{\Lambda_{\min}(s)} \leq K^2(s,k_0,X)$ by Proposition 3.2; and it also dominates the second term given (3.7) and the upper bound on λ_n. □

10.5 Bounds for r_n

Lemma 10.3.. *Consider a fixed design X with $\max_j \|X_j\|_2 \leq c_0\sqrt{n}$ and assume that (2.4) holds. Then for all subsets S with $|S| \leq s$,*

$$r_n := \left\|X_{S^c}^T X_S (X_S^T X_S)^{-1}\right\|_\infty \leq \frac{c_0\sqrt{s}}{\sqrt{\Lambda_{\min}(s)}}. \tag{10.16}$$

$$r_n \leq \frac{\theta_{1,s}\sqrt{s}}{\Lambda_{\min}(s)}, \tag{10.17}$$

where $\theta_{1,s}$ is given in 3.5

Proof. As a shorthand, we let $P_S = X_S(X_S^T X_S)^{-1}X_S^T$ denote the projection matrix and define

$$\forall j \in S^c, \;\; r_j = (X_S^T X_S)^{-1}X_S^T X_j.$$

Bounding $\|r_j\|_1 \;\forall j$ yields a bound on r_n. First we have for all $j \in S^c$,

$$\begin{aligned}
\|X_S r_j\|_2 = \left\|X_S(X_S^T X_S)^{-1}X_S^T X_j\right\|_2 &= \|P_S X_j\|_2 \\
&\leq \|X_j\|_2 \leq c_0\sqrt{n}.
\end{aligned} \tag{10.18}$$

On the other hand, by the restricted eigenvalue assumption, we have

$$\|X_S r_j\|_2^2 = r_j^T X_S^T X_S r \geq n\Lambda_{\min}\left(\frac{X_S^T X_S}{n}\right)\|r_j\|_2^2.$$

Thus we have that $\|r_j\|_2 \leq \frac{c_0}{\sqrt{\Lambda_{\min}(s)}}, \forall j \in S^c$, and hence

$$r_n = \max_{j\in S^c}\|r_j\|_1 \leq \max_{j\in S^c}\sqrt{s}\,\|r_j\|_2 = \sqrt{s}\max_{j\in S^c}\|r_j\|_2 \leq \frac{c_0\sqrt{s}}{\sqrt{\Lambda_{\min}(s)}}.$$

Next we note that using (3.5), we can bound r_n as follows, which has essentially been shown in Candès and Tao (2007). For $P_S X_j = X_S r_j$, with

$$\|r_j\|_2 \quad \leq \quad \frac{\|X_S r_j\|_2}{\sqrt{n\Lambda_{\min}(s)}} = \frac{\|P_S X_j\|_2}{\sqrt{n\Lambda_{\min}(s)}}$$

we have

$$\begin{aligned}\frac{\|P_S X_j\|_2^2}{n} &= \frac{\langle\, P_S X_j, X_j\,\rangle}{n} = \frac{\langle\, X_S r_j, X_j\,\rangle}{n} \\ &\leq \theta_{1,s}\|r_j\|_2 \leq \theta_{1,s}\frac{\|X_S r_j\|_2}{\sqrt{n\Lambda_{\min(s)}}} = \theta_{1,s}\frac{\|P_S X_j\|_2}{\sqrt{n\Lambda_{\min}(s)}}\end{aligned}$$

Hence,

$$\|P_S X_j\|_2 \quad \leq \quad \frac{\sqrt{n}\theta_{1,s}}{\sqrt{\Lambda_{\min}(s)}} \text{ and } r_n \leq \frac{\sqrt{s}\theta_{1,s}}{\Lambda_{\min}(s)}.$$

□

11 Proofs for Section 5

11.1 Proof of Proposition 5.2

We first bound $\|X\gamma\|_n^2 - \gamma^T\Sigma\gamma$.

$$\begin{aligned}&\left|\|X\gamma\|_n^2 - \gamma^T\Sigma\gamma\right| = \left|\gamma^T\widehat{\Sigma}\gamma - \gamma^T\Sigma\gamma\right| = \left|\sum_{j=1}^p\sum_{k=1}^p \gamma_j\gamma_k(\widehat{\Sigma}_{jk} - \Sigma_{jk})\right| \\ \leq\ &\left|\sum_{j\in S}\sum_{k\in S}\gamma_j\gamma_k(\widehat{\Sigma}_{jk} - \Sigma_{jk})\right| + \left|\sum_{j\in S^c}\sum_{k\in S^c}\gamma_j\gamma_k(\widehat{\Sigma}_{jk} - \Sigma_{jk})\right| \\ +\ &2\left|\sum_{j\in S}\sum_{k\in S^c}\gamma_j\gamma_k(\widehat{\Sigma}_{jk} - \Sigma_{jk})\right| \\ \leq\ &\max_{j,k}|\Delta_{jk}|\left(\|\gamma_S\|_1^2 + 2\|\gamma_S\|_1\|\gamma_{S^c}\|_1 + \|\gamma_{S^c}\|_1^2\right),\end{aligned}$$

where $\Delta = \widehat{\Sigma} - \Sigma$. Now given that $\|\gamma_{S^c}\|_1 \leq k_0 \|\gamma_S\|_1$ and $\|\gamma_S\|_1^2 \leq s \|\gamma_S\|_2^2$, we have

$$\begin{aligned} & \left| \|X\gamma\|_n^2 - \gamma^T \Sigma \gamma \right| \leq \max_{j,k} |\Delta_{jk}| \, \|\gamma_S\|_1^2 \, (1 + 2k_0 + k_0^2) \\ \leq & \max_{j,k} |\Delta_{jk}| \, \|\gamma_S\|_1^2 \, (1 + k_0)^2 \leq s(1 + k_0)^2 \max_{j,k} |\Delta_{jk}| \, \|\gamma_S\|_2^2 . \end{aligned}$$

Let $\gamma_{SS'} = \gamma_S \cup \gamma_{S'}$, where $\gamma_{S'}$ denote the subset of $\{1, \ldots, p\}$ corresponding to the s largest coordinates of γ in their absolute values in γ_{S^c}. We have on $\mathcal{X}$, using Assumption 5.1,

$$\begin{aligned} & \|X\gamma\|_n^2 \geq \gamma^T \Sigma \gamma - s(1 + k_0)^2 \max_{j,k} |\Delta_{jk}| \, \|\gamma_S\|_2^2 \\ \geq & \frac{\|\gamma_{SS'}\|_2^2}{K(s, s, k_0, \Sigma)^2} - s(1 + k_0)^2 \max_{j,k} |\Delta_{jk}| \, \|\gamma_S\|_2^2 \geq \frac{\|\gamma_{SS'}\|_2^2}{2K(s, s, k_0, \Sigma)^2}, \end{aligned}$$

and hence (5.2) holds. □

11.2 Eigenvalue bounds

We now show that (2.4) is satisfied with high probability for a random design X, given its population correspondent as in (2.8).

Lemma 11.1. *Let X be a random design as in (2.6). Let $s \leq \frac{\Lambda_{\min}(s)}{16 C_2} \sqrt{\frac{n}{\log p}}$ for C_2 as defined in (2.9). We have on the set $\mathcal{X}$,*

$$\Lambda_{\min} \left(\frac{X_S^T X_S}{n} \right) \geq \Lambda_{\min}(s), \tag{11.1}$$

for all subsets $S \subset \{1, \ldots, p\}$ with $|S| \leq s$ where (2.8) hold.

Proof. On the set $\mathcal{X}$, for all subsets S with $|S| \leq s$,

$$\begin{aligned} \left| \Lambda_{\min} \left(\frac{X_S^T X_S}{n} \right) - \Lambda_{\min}(\Sigma_{SS}) \right| & \leq \left\| \left(\frac{X_S^T X_S}{n} \right) - \Sigma_{SS} \right\|_2 & (11.2) \\ & \leq \left\| \left(\frac{X_S^T X_S}{n} \right) - \Sigma_{SS} \right\|_\infty & (11.3) \\ & \leq s C_2 \sqrt{\frac{\log p}{n}} \leq \frac{\Lambda_{\min}(s)}{16}, & (11.4) \end{aligned}$$

where $\|.\|_2$ denotes here the operator norm of a matrix. (11.2) is a standard result in matrix perturbation theory, (11.3) is due to the fact that $\widehat{\Sigma}$ and Σ are symmetric, and (11.4) is due to (2.9) and the bound on s. Hence for all subsets S with $|S| \leq s$ that satisfy $\Lambda_{\min}(\Sigma_{SS}) \geq \frac{17}{16} \Lambda_{\min}(s)$ (as defined in (2.7)), (11.1) holds. □

11.3 Proof of Theorem 5.3

As corollary of Lemmas 10.3 and 11.1, we have

Corollary 11.2.. *Consider a random design X. Then on the set $\mathcal{X}$ defined in (2.9), (10.16) holds with $c_0 = \sqrt{3/2}$, for all subsets S with $|S| \leq s$.*

It is clear that (4.1) is always satisfied given (4.9), where $K = \sqrt{2}K(s,3,3,\Sigma)$, as $K(s,s,k_0,X) \leq \sqrt{2}K(s,s,k_0,\Sigma)$ by Proposition 5.2. We now show that the conditions on λ_n, s and $\beta_{\min}$ as required by Theorem 2.1 are satisfied on $\mathcal{X} \cap \mathcal{T}$. First we take

$$\widetilde{\delta}_S := 8K^2(s,s,3,\Sigma)\lambda_{\text{init}}\sqrt{s} \tag{11.5}$$

$$\widetilde{\delta}_{S^c} := 32K^2(s,s,3,\Sigma)\lambda_{\text{init}}\sqrt{s}, \tag{11.6}$$

$$\widetilde{r_n} := \frac{\sqrt{3s}}{\sqrt{2\Lambda_{\min}(s)}}, \tag{11.7}$$

where (11.7) holds by Corollary 11.2, for which

$$s \leq \frac{1}{32C_2K^2(s,s,3,\Sigma)} \leq \frac{\Lambda_{\min}(s)}{16C_2}\sqrt{\frac{n}{\log p}},$$

by Proposition 5.1. It is clear that

$$\widetilde{\delta}_S \geq 4K^2(s,s,3,X)\lambda_{\text{init}}\sqrt{s} \geq \|\delta_S\|_\infty, \text{ and} \tag{11.8}$$

$$\widetilde{\delta}_{S^c} \geq 16K^2(s,s,3,X)\lambda_{\text{init}}\sqrt{s} \geq \|\delta_{S^c}\|_\infty \tag{11.9}$$

given (4.2a) and (4.2b), and Proposition 5.2. Regarding the condition on λ_n, by Proposition 5.2, (4.4) and (11.6), we have

$$\begin{aligned}
\lambda_n &\geq \frac{128c_0\sigma K^2(s,s,3,\Sigma)\lambda_{\text{init}}\sqrt{|\bar{S}|}}{\eta}\sqrt{\frac{2\log(p-s)}{n}} \\
&\geq \frac{4c_0\sigma(32K^2(s,s,3,\Sigma)\lambda_{\text{init}}\sqrt{s})}{\eta}\sqrt{\frac{2\log(p-s)}{n}} \\
&= \frac{4c_0\sigma\widetilde{\delta}_{S^c}}{\eta}\sqrt{\frac{2\log(p-s)}{n}}
\end{aligned}$$

and

$$\begin{aligned}
\lambda_n &\leq \frac{16M\sqrt{2}K(s,s,3,\Sigma)K(s,s,3,X)c_0\sigma\lambda_{\text{init}}\sqrt{|\bar{S}|}}{K(s,s,3,X)}\sqrt{\frac{2\log(p-s)}{n}} \\
&\leq Mc_0\sigma 32K^2(s,3,3,\Sigma)\lambda_{\text{init}}\sqrt{s}\sqrt{\frac{2\log(p-s)}{n}} \\
&= Mc_0\sigma\widetilde{\delta}_{S^c}\sqrt{\frac{2\log(p-s)}{n}},
\end{aligned}$$

where we used the fact that $K(s,3,X) \leq K(s,s,3,X)$. Hence (2.12) is satisfied. In addition, for $K = \sqrt{2}K(s,s,3,\Sigma)$, the sparsity condition (2.13) holds by Corollary 11.2. Condition (4.7) implies that the condition (2.14) for $\beta_{\min}$ holds, given (11.5) and (11.6) and Proposition 5.1. We can then invoke Theorem 2.1 to finish the proof with $c_0 = \sqrt{3/2}$. □

12 Proof of the sign recovery Lemma

12.1 Preliminaries

We first state necessary and sufficient conditions for the event $\text{sgn}(\widehat{\beta}) = \text{sgn}(\beta^*)$. Note that this is essentially equivalent to Lemmas 2 and 3 in Wainwright (2008). First, for $\widehat{\Sigma} = X^T X/n$, let $\widehat{\Sigma}_{RT} = \frac{1}{n} X_R^T X_T$ be the submatrix of $\widehat{\Sigma}$ with rows and columns indexed by R and T respectively.

Lemma 12.1.. *Let $\vec{b} := (\text{sgn}(\beta_j^*) w_j)_{j \in S}$. Let $\vec{w} = (w_1, w_2, \ldots, w_p)$, where $w_j > 0, \forall j$, be a positive weight vector. Assume that the matrix $X_S^T X_S$ is invertible. Then for any given $\lambda_n > 0$ and noise vector $\epsilon \in \mathbb{R}^n$, there exists a solution $\widehat{\beta}$ for the weighted Lasso such that*

$$\text{sgn}(\widehat{\beta}) = \text{sgn}(\beta^*),$$

if and only if the following two conditions hold:

$$\left| \widehat{\Sigma}_{S^c S} (\widehat{\Sigma}_{S,S})^{-1} \left[\frac{X_S^T \epsilon}{n} - \lambda_n \vec{b} \right] - \frac{X_{S^c}^T \epsilon}{n} \right| \leq \lambda_n \vec{w}_{S^c}, \tag{12.1a}$$

$$\text{sgn}\left(\beta_S^* + (\widehat{\Sigma}_{SS})^{-1} \left[\frac{X_S^T \epsilon}{n} - \lambda_n \vec{b} \right] \right) = \text{sgn}(\beta_S^*). \tag{12.1b}$$

Finally, if (12.1a) holds with strict inequality, then the solution of the weighted Lasso is unique.

Proof. Recall that we observe $Y = X\beta^* + \epsilon$ and $\vec{b} := (\text{sgn}(\beta_i^*) w_i)_{i \in S}$. Let $w = (w_1, w_2, \ldots, w_p)$ be the weight vector.

First observe that the KKT conditions imply that $\widehat{\beta} \in \mathbb{R}^p$ is a solution, if and only if there exists a subgradient

$$\vec{g} \in \partial \sum_{j=1}^{p} w_j |\widehat{\beta}_j| = \{ z \in \mathbb{R}^p |\ z_i = \text{sgn}(\widehat{\beta}) w_i \text{ for } \widehat{\beta}_i \neq 0, \text{ and } |z_j| \leq w_j \text{ otherwise} \}$$

such that

$$\frac{1}{n} X^T X \widehat{\beta} - \frac{1}{n} X^T Y + \lambda_n \vec{g} = 0, \tag{12.2}$$

which is equivalent to the following linear system by substituting $Y = X\beta^* + \epsilon$ and re-arranging:

$$\widehat{\Sigma}(\widehat{\beta} - \beta^*) - \frac{1}{n} X^T \epsilon + \lambda_n \vec{g} = 0. \tag{12.3}$$

Hence, given X, β^*, ϵ and $\lambda_n > 0$ the event $\text{sgn}(\widehat{\beta}) = \text{sgn}(\beta_S^*)$ holds if and only if

1. there exist a point $\widehat{\beta} \in \mathbb{R}^p$ and a subgradient $\vec{g} \in \partial \sum_{j=1}^p w_j |\widehat{\beta}_j|$ such that (12.3) holds, and
2. $\text{sgn}(\widehat{\beta}_S) = \text{sgn}(\beta_S^*)$ and $\widehat{\beta}_{S^c} = \beta_{S^c}^* = 0$, which implies that $\vec{g}_S = \vec{b}$ and $|\vec{g}_j| \leq w_j \forall j \in S^c$ by definition of $\vec{g}$.

Plugging $\widehat{\beta}_{S^c} = \beta_{S^c}^* = 0$ and $\vec{g}_S = \vec{b}$ in (12.3) shows that $\text{sgn}(\widehat{\beta}) = \text{sgn}(\beta^*)$ if and only if

1. there exists a point $\widehat{\beta} \in \mathbb{R}^p$ and a subgradient $\vec{g} \in \partial \sum_{j=1}^p w_j |\widehat{\beta}_j|$ such that

$$\widehat{\Sigma}_{S^cS}(\widehat{\beta}_S - \beta^*_S) - \frac{X^T_{S^c}\epsilon}{n} = -\lambda_n \vec{g}_{S^c}, \tag{12.4a}$$

$$\widehat{\Sigma}_{SS}(\widehat{\beta}_S - \beta^*_S) - \frac{X^T_S \epsilon}{n} = -\lambda_n \vec{g}_S = -\lambda_n \vec{b}, \tag{12.4b}$$

2. and $\text{sgn}(\widehat{\beta}_S) = \text{sgn}(\beta^*_S)$ and $\widehat{\beta}_{S^c} = \beta^*_{S^c} = 0$.

Using invertibility of $X^T_S X_S$, we can solve for $\widehat{\beta}_S$ and $\vec{g}_{S^c}$ using (12.4a) and (12.4b) to obtain

$$\begin{aligned} -\lambda_n \vec{g}_{S^c} &= \widehat{\Sigma}_{S^cS}(\widehat{\Sigma}_{SS})^{-1}\left[\frac{X^T_S\epsilon}{n} - \lambda_n \vec{b}\right] - \frac{X^T_{S^c}\epsilon}{n}, \\ \widehat{\beta}_S &= \beta^*_S + (\widehat{\Sigma}_{SS})^{-1}\left[\frac{1}{n}X^T_S\epsilon - \lambda_n \vec{b}\right]. \end{aligned}$$

Thus, given invertibility of $X^T_S X_S$, $\text{sgn}(\widehat{\beta}) = \text{sgn}(\beta^*)$ holds if and only if

1. there exists simultaneously a point $\widehat{\beta} \in \mathbb{R}^p$ and a subgradient $\vec{g} \in \partial \sum_{j=1}^p w_j |\widehat{\beta}_j|$ such that

$$-\lambda_n \vec{g}_{S^c} = \widehat{\Sigma}_{S^cS}(\widehat{\Sigma}_{SS})^{-1}\left[\frac{X^T_S\epsilon}{n} - \lambda_n \vec{b}\right] - \frac{X^T_{S^c}\epsilon}{n}, \tag{12.5a}$$

$$\widehat{\beta}_S = \beta^*_S + (\widehat{\Sigma}_{SS})^{-1}\left[\frac{X^T_S\epsilon}{n} - \lambda_n \vec{b}\right], \tag{12.5b}$$

2. and $\text{sgn}(\widehat{\beta}_S) = \text{sgn}(\beta^*_S)$ and $\widehat{\beta}_{S^c} = \beta^*_{S^c} = 0$.

Thus, for $\text{sgn}(\widehat{\beta}) = \text{sgn}(\beta^*)$ to hold, there exists simultaneously a point $\widehat{\beta} \in \mathbb{R}^p$ and a subgradient $\vec{g} \in \partial \sum_{j=1}^p w_j |\widehat{\beta}_j|$ such that

$$\begin{aligned} \left|\widehat{\Sigma}_{S^cS}(\widehat{\Sigma}_{SS})^{-1}\left[\frac{X^T_S\epsilon}{n} - \lambda_n \vec{b}\right] - \frac{X^T_{S^c}\epsilon}{n}\right| &= |-\lambda_n \vec{g}_{S^c}| \leq \lambda_n \vec{w}_{S^c}, \\ \text{sgn}(\widehat{\beta}_S) = \text{sgn}\left(\beta^*_S + (\widehat{\Sigma}_{SS})^{-1}\left[\frac{1}{n}X^T_S\epsilon - \lambda_n \vec{b}\right]\right) &= \text{sgn}(\beta^*_S), \end{aligned}$$

given that $|\vec{g}_{S^c}| \leq \vec{w}_{S^c}$ by definition of $\vec{g}$. Thus (12.1a) and (12.1b) hold for the given X, β^*, ϵ and $\lambda_n > 0$. Thus we have shown the lemma in one direction.

For the reverse direction, given X, β^*, ϵ, and suppose that (12.1a) and (12.1b) hold for some $\lambda_n > 0$, we first construct a point $\widehat{\beta} \in \mathbb{R}^p$ by letting $\widehat{\beta}_{S^c} = \beta^*_{S^c} = 0$ and

$$\widehat{\beta}_S = \beta^*_S + (\widehat{\Sigma}_{SS})^{-1}\left[\frac{1}{n}X^T_S\epsilon - \lambda_n \vec{b}\right]$$

which guarantees that

$$\text{sgn}(\widehat{\beta}_S) = \text{sgn}\left(\beta^*_S + (\widehat{\Sigma}_{SS})^{-1}\left[\frac{1}{n}X^T_S\epsilon - \lambda_n \vec{b}\right]\right) = \text{sgn}(\beta^*_S)$$

by (12.1b). We simultaneously construct $\vec{g}$ by letting $\vec{g}_S = \vec{b}$ and

$$\vec{g}_{S^c} = -\frac{1}{\lambda_n}\left(\widehat{\Sigma}_{S^cS}(\widehat{\Sigma}_{SS})^{-1}\left[\frac{1}{n}X_S^T\epsilon - \lambda_n\vec{b}\right] - \frac{1}{n}X_{S^c}^T\epsilon\right), \tag{12.6}$$

which guarantees that $|\vec{g}_j| \leq w_j, \forall j \in S^c$ due to (12.1b); hence $\vec{g} \in \partial\sum_{j=1}^p w_j|\widehat{\beta}_j|$. Thus, we have found a point $\widehat{\beta} \in \mathbb{R}^p$ and a subgradient $\vec{g} \in \partial\sum_{j=1}^p w_j|\widehat{\beta}_j|$ such that $\text{sgn}(\widehat{\beta}) = \text{sgn}(\beta^*)$ and the set of equations (12.5a) and (12.5b) is satisfied. Hence, by invertibility of $X_S^TX_S$, $\text{sgn}(\widehat{\beta}) = \text{sgn}(\beta^*)$ for the given $X, \beta^*, \epsilon, \lambda_n$. □

12.2 Uniqueness of solution

Finally, the uniqueness proof follows a similar argument in the revised draft of Wainwright (2008). We omit the details. In fact, it is illustrative to rewrite the adaptive (or weighted) Lasso program as follows: Let $W = \text{diag}(w_1, \ldots, w_p)$, for $w_j > 0$, and let the solution to (2.2) be

$$\widehat{\beta} = W^{-1}\widehat{\beta}_0, \quad \text{where}$$

$$\widehat{\beta}_0 \;:=\; \arg\min_{\beta_0} \frac{1}{2n}\|Y - XW^{-1}\beta_0\|_2^2 + \lambda_n\|\beta_0\|_1. \tag{12.7}$$

Now we can just take XW^{-1} as the design matrix and $\beta_0 := W\beta$ as the sparse vector that we recover through $\widehat{\beta}_0$, by solving the standard Lasso problem as in (12.7). It is clear that uniqueness of $\widehat{\beta}_0$ to (12.7) is equivalent to uniqueness of $\widehat{\beta}$ as W is a positive-definite matrix.

12.3 Proof of Lemma 8.2

Let $e_i \in \mathbb{R}^s$ be the vector with 1 in i^{th} position and zero elsewhere; hence $\|e_i\|_2 = 1$.

We first define a set of random variables that are relevant for (12.1a) and (12.1b):

$$\begin{aligned}
\forall j \in S^c,\;\; V_j &:= X_j^TX_S(X_S^TX_S)^{-1}\lambda_n\vec{b} + X_j^T\left\{I_{n\times n} - X_S(X_S^TX_S)^{-1}X_S^T\right\}\frac{\epsilon}{n},\\
\forall i \in S,\;\; U_i &:= e_i^T\left(\tfrac{1}{n}X_S^TX_S\right)^{-1}\left[\tfrac{1}{n}X_S^T\epsilon - \lambda_n\vec{b}\right].
\end{aligned}$$

Condition (12.1a) holds if and only if the event

$$\mathcal{E}(V) := \{\forall j \in S^c,\; |V_j| \leq \lambda_n w_j\}$$

is true. For Condition (12.1b), the event

$$\mathcal{E}(U) := \left\{\max_{i\in S}|U_i| \leq \beta_{\min}\right\},$$

is sufficient to guarantee that Condition (12.1b) holds.

We first prove that $\mathbb{P}\left(\mathcal{E}(V)\right)$ and $\mathbb{P}\left(\mathcal{E}(U)\right)$ both are large.

Analysis of $\mathcal{E}(V)$. Note that

$$\mu_j = \mathbb{E}\left(V_j\right) = \lambda_n X_j^T X_S (X_S^T X_S)^{-1} \vec{b}, \;\; j \in S^c.$$

By (8.4a), we have $\forall j \in S^c$,

$$|\mu_j| \leq \lambda_n w_j (1-\eta). \tag{12.8}$$

Denote by $P = X_S(X_S^T X_S)^{-1} X_S^T = P^2$ the projection matrix. Let

$$\widetilde{V}_j = X_j^T \left\{ \left[I_{n\times n} - X_S (X_S^T X_S)^{-1} X_S^T \right] \frac{\epsilon}{n} \right\}, \;\; j \in S^c \tag{12.9}$$

which is a zero-mean Gaussian random variable with variance

$$\mathsf{Var}(\widetilde{V}_j) = \frac{\sigma^2}{n^2} X_j^T \left\{ \left[(I_{n\times n} - P) \right] \left[(I_{n\times n} - P) \right]^T \right\} X_j \leq \frac{\sigma^2}{n^2} \left\| X_j \right\|_2^2 = \frac{\sigma^2 c_0^2}{n}$$

since $\left\| I - P \right\|_2 \leq 1$. Using the tail bound for a Gaussian random variable

$$\begin{aligned} \mathbb{P}\left(\left| \widetilde{V}_j \right| \geq t \right) &\leq \frac{\sqrt{\mathsf{Var}(\widetilde{V}_j)}}{t} \exp\left(\frac{-t^2}{2\mathsf{Var}(\widetilde{V}_j)} \right) \\ &\leq \frac{\sigma c_0}{\sqrt{n} t} \exp\left(\frac{-nt^2}{2\sigma^2 c_0^2} \right), \end{aligned} \tag{12.10}$$

with $t = \frac{\eta \lambda_n w_{\min}(S^c)}{2} \geq 2c_0 \sigma \sqrt{\frac{2\log(p-s)}{n}}$ and the union bound, we have

$$\begin{aligned} \mathbb{P}\left(\max_{j \in S^c} \left| \widetilde{V}_j \right| \geq \frac{\eta \lambda_n w_{\min}(S^c)}{2} \right) &\leq \frac{(p-s)\exp\left(-4\log(p-s)\right)}{2\sqrt{2\log(p-s)}} \\ &\leq \frac{1}{2(p-s)^3 \sqrt{2\log(p-s)}}. \end{aligned}$$

Thus, with probability at least $1 - \frac{1}{2(p-s)^3}$,

$$\begin{aligned} \forall j \in S^c, \quad |V_j| &\leq |\mu_j| + |\widetilde{V}_j| \leq \lambda_n w_j (1-\eta) + \frac{\eta \lambda_n w_{\min}(S^c)}{2} \\ &\leq \lambda_n w_j (1 - \eta/2), \end{aligned}$$

and $\mathcal{E}(V)$ holds; in fact, it holds with straight inequality for $\eta > 0$.

Analysis of $\mathcal{E}(U)$. By the triangle inequality, and on the set $\mathcal{T}$,

$$\begin{aligned} \max_{i \in S} |U_i| &\leq \left\| \left(X_S^T X_S / n \right)^{-1} \right\|_\infty \left\| X_S^T \epsilon / n \right\|_\infty + \left\| \left(X_S^T X_S / n \right)^{-1} \right\|_\infty \lambda_n w_{\max} \\ &\leq \frac{\sqrt{s}}{\Lambda_{\min}(s)} \left(c_0 \sigma \sqrt{24 \log p / n} + \lambda_n w_{\max}(S) \right) < \beta_{\min}, \end{aligned}$$

where

$$\left\| \left(X_S^T X_S / n \right)^{-1} \right\|_\infty \leq \sqrt{s} \left\| \left(X_S^T X_S / n \right)^{-1} \right\|_2 = \frac{\sqrt{s}}{\Lambda_{\min}\left(X_S^T X_S / n \right)} \leq \frac{\sqrt{s}}{\Lambda_{\min}(s)},$$

by standard matrix norm comparison results and the restricted eigenvalue assumption. Hence, $\mathcal{E}(U)$ holds on the set $\mathcal{T}$. Denote by $\mathcal{F} = \mathcal{E}(U)^c \cup \mathcal{E}(V)^c$. Then we have

$$\begin{aligned}\mathbb{P}(\mathcal{F}) &= \mathbb{P}(\mathcal{F} \cap \mathcal{T}^c) + \mathbb{P}(\mathcal{F} \cap \mathcal{T}) \\ &\leq \mathbb{P}(\mathcal{T}^c) + \mathbb{P}(\mathcal{E}(V)^c \cap \mathcal{T}) \\ &\leq \mathbb{P}(\mathcal{T}^c) + \mathbb{P}(\mathcal{E}(V)^c) \leq 2/p^2\end{aligned}$$

by Lemma 9.1 and the analysis of $\mathcal{E}(U)$ and as $\mathcal{E}(V)$ above.

13 Proof of Theorem 2.1

We note that for a fixed design X, once we finish checking that the incoherence conditions and conditions on λ_n and $\beta_{\min}$ as in (8.6) are satisfied, we can then invoke Lemma 8.2 to finish the theorem. For a random design, our proof follows the case of a fixed design after we exclude the bad event $\mathcal{X}^c$ for $\mathcal{X}$ as defined in (2.9). We now show that on $\mathcal{X} \cap \mathcal{T}$, where for a fixed design $\mathcal{X}^c = \emptyset$, all conditions in Lemma 8.2 for $c_0^2 = 3/2$ are indeed satisfied.

First by Lemma 11.1, we have $\Lambda_{\min}(X_S^T X_S/n) \geq \Lambda_{\min}(s)$ and hence (8.3b) hold under $\mathcal{X} \cap \mathcal{T}$, given (2.8). Now we have by $\beta_{\min} \geq 2\widetilde{\delta}_S \geq 2\|\delta_S\|_\infty$,

$$\begin{aligned}\forall\, j \in S, \quad |\beta_{j,\text{init}}| &\geq \beta_{\min} - \|\delta_S\|_\infty \geq \frac{\beta_{\min}}{2} \quad \text{and hence} \\ w_{\max} &\leq \max\left\{\frac{2}{\beta_{\min}}, 1\right\}.\end{aligned}$$

It also holds by $1 > \widetilde{\delta}_{S^c} \geq \|\delta_{S^c}\|_\infty$

$$\forall j \in S^c, \quad |\beta_{j,\text{init}}| \leq \|\delta_{S^c}\|_\infty \leq \widetilde{\delta}_{S^c} < 1 \quad \text{and} \quad w_{\min} \geq \frac{1}{\|\delta_{S^c}\|_\infty}.$$

Hence the choice of λ_n in (2.12) guarantees that

$$\lambda_n w_{\min} \geq \frac{\lambda_n}{\|\delta_{S^c}\|_\infty} \geq \frac{\lambda_n}{\widetilde{\delta}_{S^c}} \geq \frac{4c_0\sigma}{\eta}\sqrt{\frac{2\log(p-s)}{n}}.$$

We now show that the incoherence condition as in (8.5) holds given $\widetilde{r_n} \geq r_n$.

1. Suppose $\beta_{\min} \leq 2$ satisfies (2.14), we have $w_{\max} = 2/\beta_{\min}$ and hence

$$\frac{w_{\min}(1-\eta)}{w_{\max}} \geq \frac{\beta_{\min}(1-\eta)}{2\|\delta_{S^c}\|_\infty} \geq \frac{\beta_{\min}(1-\eta)}{2\widetilde{\delta}_{S^c}} \geq \widetilde{r_n} \geq r_n. \tag{13.1}$$

2. Suppose $\beta_{\min} > 2$: then $w_{\max}(S) = 1$ and by assumption,

$$\frac{w_{\min}(1-\eta)}{w_{\max}} \geq \frac{1-\eta}{\|\delta_{S^c}\|_\infty} \geq \frac{1-\eta}{\widetilde{\delta}_{S^c}} \geq \widetilde{r_n} \geq r_n.$$

It is clear that (8.6) is satisfied given (2.14), if

$$\beta_{\min} \geq \max\left\{\frac{4\lambda_n\sqrt{s}}{\beta_{\min}\Lambda_{\min}(s)}, \frac{2\lambda_n\sqrt{s}}{\Lambda_{\min}(s)}\right\}. \tag{13.2}$$

We only need to be concerned with the first term: given the last two terms in the $\beta_{\min}$ bound, we have

$$\begin{aligned}
\beta_{\min}^2 &\geq \frac{4Mc_0\sigma\widetilde{\delta}_{S^c}\sqrt{s}}{\Lambda_{\min}(s)}\sqrt{\frac{2\log p}{n}} \quad \text{hence} \\
\beta_{\min} &\geq \frac{4\sqrt{s}}{\beta_{\min}\Lambda_{\min}(s)}Mc_0\sigma\widetilde{\delta}_{S^c}\sqrt{\frac{2\log(p-s)}{n}} \geq \frac{4\lambda_n\sqrt{s}}{\beta_{\min}\Lambda_{\min}(s)}.
\end{aligned}$$

Finally, we have for both fixed and random designs, let $\mathcal{F}$ be a shorthand for the event $\text{sgn}(\widehat{\beta}) \neq \text{sgn}(\beta^*)$. We have

$$\mathbb{P}(\mathcal{F}) \leq \mathbb{P}((\mathcal{T}\cap\mathcal{X})^c) + \mathbb{P}(\mathcal{F}\cap\mathcal{T}\cap\mathcal{X}) \leq \mathbb{P}((\mathcal{T}\cap\mathcal{X})^c) + 1/p^2,$$

where $\mathcal{X}^c = \emptyset$ for a fixed design, and the last term has been bounded using Lemma 8.2 for a fixed design or conditioned on a random design on the set $\mathcal{X}$ with $c_0 = \sqrt{3/2}$.

References

BANERJEE, O., GHAOUI, L. E. and D'ASPREMONT, A. (2008). Model selection through sparse maximum likelihood estimation. *Journal of Machine Learning Research* **9** 485–516.

BICKEL, P. J., , RITOV, Y. and TSYBAKOV, A. (2008). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics, to appear* .

CANDÈS, E. and TAO, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics* **35** 2313–2351.

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2008a). Regularized paths for generalized linear models via coordinate descent. Tech. rep., Department of Statistics, Stanford University.

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2008b). Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics* **9** 432–441.

GREENSHTEIN, E. and RITOV, Y. (2004). Persistency in high dimensional linear predictor-selection and the virtue of over-parametrization. *Bernoulli* **10** 971–988.

HUANG, J., MA, S. and ZHANG, C.-H. (2008). Adaptive Lasso for sparse high-dimensional regression models. *Statist. Sinica* **18** 1603–1618.

JOHNSTONE, I. (2001). Chi-square oracle inequalities. *In State of the Art in Probability and Statistics, Festchrift for Willem R. van Zwet, M. de Gunst and C. Klaassen and A. van der Waart editors, IMS Lecture Notes - Monographs* **36** 399–418.

KOLTCHINSKII, V. (2008). Dantzig selector and sparsity oracle inequalities. *Bernoulli* To appear.

LAURITZEN, S. L. (1996). *Graphical Models*. Oxford University Press.

MEIER, L., VAN DE GEER, S. and BÜHLMANN, P. (2008). The group Lasso for logistic regression. *Journal of the Royal Statistical Society Series B* **70** 53–71.

MEINSHAUSEN, N. (2007). Relaxed Lasso. *Computational Statistics and Data Analysis* **52** 374–393.

MEINSHAUSEN, N. (2008). A note on the Lasso for graphical gaussian model selection. *Statistics and Probability Letters* **78** 880–884.

MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High dimensional graphs and variable selection with the Lasso. *Annals of Statistics* **34** 1436–1462.

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

RAVIKUMAR, P., WAINWRIGHT, M., RASKUTTI, G. and YU, B. (2008). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. In *Advances in Neural Information Processing Systems*. MIT Press. Longer version in arXiv:0811.3628v1.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288.

VAN DE GEER, S. A. (2007). The deterministic Lasso. *The JSM Proceedings* .

VAN DE GEER, S. A. (2008). High-dimensional generalized linear models and the Lasso. *The Annals of Statistics* **36** 614–645.

WAINWRIGHT, M. (2007). Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. Tech. Rep. 725, Department of Statistics, UC Berkeley.

WAINWRIGHT, M. (2008). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming. *IEEE Trans. Inform. Theory* To appear, also posted as Technical Report 709, 2006, Department of Statistics, UC Berkeley.

WASSERMAN, L. and ROEDER, K. (2008). High dimensional variable selection. *The Annals of Statistics* To appear.

ZHAO, P. and YU, B. (2006). On model selection consistency of Lasso. *Journal of Machine Learning Research* **7** 2541–2567.

ZHOU, S., LAFFERTY, J. and WASSERMAN, L. (2008). Time varying undirected graphs. In *Proceedings of the 21st Annual Conference on Computational Learning Theory (COLT'08)*.

ZOU, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association* **101** 1418–1429.

ZOU, H. and LI, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models (with discussion). *Annals of Statistics* **36** 1509–1566.