Journal of Machine Learning Research X (2008) X-X Submitted 5/2008; Published X/XX

On the Consistency of Feature Selection using Greedy Least Squares Regression

Tong Zhang* TZHANG@STAT.RUTGERS.EDU
STATISTICS DEPARTMENT
RUTGERS UNIVERSITY, NJ

Editor: Bin Yu

Abstract

This paper studies the feature selection problem using a greedy least squares regression algorithm. We show that under a certain irrepresentable condition on the design matrix (but independent of the sparse target), the greedy algorithm can select features consistently when the sample size approaches infinity. The condition is identical to a corresponding condition for Lasso.

Moreover, under a sparse eigenvalue condition, the greedy algorithm can reliably identify features as long as each nonzero coefficient is larger than a constant times the noise level. In comparison, Lasso may require the coefficients to be larger than $O(\sqrt{s})$ times the noise level in the worst case, where s is the number of nonzero coefficients.

1. Introduction

We are interested in the statistical feature selection problem for least squares regression. Let $X = [\mathbf{x}_1, \ldots, \mathbf{x}_d] \in \mathbb{R}^{n \times d}$ be an $n \times d$ data matrix with $\mathbf{x}_j \in \mathbb{R}^n$ $(j = 1, \ldots, d)$ as its columns. Assume that the response vector $\mathbf{y} = [y_1, \ldots, y_n] \in \mathbb{R}^n$ is generated from a sparse linear combination of the basis vectors $\{\mathbf{x}_j\}$ plus a zero-mean stochastic noise vector $\mathbf{z} \in \mathbb{R}^n$:

$$\mathbf{y} = X\bar{\beta} + \mathbf{z} = \sum_{j=1}^{d} \bar{\beta}_j \mathbf{x}_j + \mathbf{z}, \tag{1}$$

where most coefficients $\bar{\beta}_j$ equal zero. The goal of feature selection is to identify the set of non-zeros $\{j : \bar{\beta}_j \neq 0\}$, where $\bar{\beta} = [\bar{\beta}_1, \ldots, \bar{\beta}_d]$. The purpose of this paper is study the performance of greedy least squares regression for feature selection.

The following notations are used throughout the paper. Given $\beta \in \mathbb{R}^d$, define

$$\mathrm{supp}(\beta) = \{j : \beta_j \neq 0\}.$$

Given $\mathbf{x} \in \mathbb{R}^n$ and $\bar{F} \subset \{1, \ldots, d\}$, let

$$\hat{\beta}_X(\bar{F}, \mathbf{x}) = \min_{\beta \in \mathbb{R}^d} \frac{1}{n} \|X\beta - \mathbf{x}\|_2^2 \quad \text{subject to} \quad \mathrm{supp}(\beta) \subset \bar{F}.$$

That is, $\hat{\beta}_X(\bar{F}, \mathbf{x})$ is the least squares solution with coefficients restricted to $\bar{F}$.

*. Supported by NSF grant DMS-0706805.

Given $\bar{F} \in \{1, \ldots, d\}$, we let $X_{\bar{F}}$ be the $n \times |\bar{F}|$ matrix that is the restriction of columns of X to $\bar{F}$. That is, $X_{\bar{F}}$'s columns are the basis functions $\mathbf{x}_j$ with $j \in \bar{F}$ arranged in the ascending order. The following quantity, appeared in (Tropp, 2004), is important in our analysis (also see Wainwright, 2006):

$$\mu_X(\bar{F}) = \max_{j \notin \bar{F}} \|(X_{\bar{F}}^T X_{\bar{F}})^{-1} X_{\bar{F}}^T \mathbf{x}_j\|_1.$$

We also define for all $\bar{F} \subset \{1, \ldots, d\}$

$$\rho_X(\bar{F}) = \inf \left\{ \frac{1}{n} \|X\beta\|_2^2 / \|\beta\|_2^2 : \mathrm{supp}(\beta) \subset \bar{F} \right\}.$$

This quantity is the smallest eigenvalue of the restricted design matrix $\frac{1}{n} X_{\bar{F}}^T X_{\bar{F}}$, which has also appeared in previous work such as (Wainwright, 2006; Zhao and Yu, 2006). The requirement that $\rho_X(F)$ is bounded away from zero for small $|F|$ is often referred to as the sparse eigenvalue condition (or the restricted isometry condition).

2. Related work

The feature selection problem of estimating $\mathrm{supp}(\bar{\beta})$ from observation $\mathbf{y}$ defined in (1) has attracted significant attention in recent years. One of the frequently used method for feature selection is Lasso, which solves the following L_1 regularization problem:

$$\hat{\beta} = \arg\min_{\beta} \left[\frac{1}{n} \left\| \sum_{j=1}^{d} \beta_j \mathbf{x}_j - \mathbf{y} \right\|_2^2 + \lambda \|\beta\|_1 \right], \tag{2}$$

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

The effectiveness of feature selection using Lasso was established in (Zhao and Yu, 2006) (also see Meinshausen and Buhlmann, 2006) under irrepresentable conditions that depend on the signs of the true target $\mathrm{sgn}(\bar{\beta})$. Results established in this paper have cruder forms that depend only on the design matrix but not the sparse target $\bar{\beta}$. Such conditions have also been studied by Zhao and Yu (2006) (also see Wainwright, 2006).

In addition to Lasso, greedy algorithms have also been widely used for feature selection. Greedy algorithms for least squares regression are called matching pursuit in the signal processing community (Mallat and Zhang, 1993). The particular algorithm analyzed in this paper (some time referred to as orthogonal matching pursuit or OMP) is presented in Figure 1. The algorithm is often called forward greedy selection in the machine learning literature.

This paper investigates the behavior of greedy least squares algorithm in Figure 1 for feature selection under the stochastic noise model (1). Our result extends that of Tropp (2004), which only considered the situation without stochastic noise. It was shown by Tropp (2004) that $\mu_X(\bar{F}) < 1$ is sufficient for the greedy algorithm to identify the correct feature set $\mathrm{supp}(\bar{\beta})$ when the noise vector $\mathbf{z} = 0$. The main contribution of this paper is to generalize Tropp's analysis to handle non-zero sub-Gaussian stochastic noise vectors. In particular, we will establish conditions on $\min_{j \in \mathrm{supp}(\bar{\beta})} |\bar{\beta}_j|$ and the stopping criterion ϵ in Figure 1, so that

Input: $X = [\mathbf{x}_1, \ldots, \mathbf{x}_d] \in \mathbb{R}^{n \times d}$, $\mathbf{y} \in \mathbb{R}^n$ and $\epsilon > 0$
Output: $F^{(k)}$ and $\beta^{(k)}$
let $\tilde{\mathbf{x}}_j = \mathbf{x}_j / \|\mathbf{x}_j\|_2$ be normalized basis $(j = 1, \ldots, d)$
let $F^{(0)} = \emptyset$ and $\beta^{(0)} = 0$
for $k = 1, 2, \ldots$
  let $i^{(k)} = \arg\max_i |\tilde{\mathbf{x}}_i^T (X\beta^{(k-1)} - \mathbf{y})|$
  if $(|\tilde{\mathbf{x}}_{i^{(k)}}^T (X\beta^{(k-1)} - \mathbf{y})| \leq \epsilon)$ **break**
  let $F^{(k)} = \{i^{(k)}\} \cup F^{(k-1)}$
  let $\beta^{(k)} = \hat{\beta}_X(F^{(k)}, \mathbf{y})$
end

Figure 1: Greedy Least Squares Regression (OMP)

the algorithm finds the correct feature set $\mathrm{supp}(\bar{\beta})$. The selection of stopping criterion ϵ in the greedy algorithm is equivalent to selecting an appropriate regularization condition λ in the Lasso formulation (2), which is necessary both in theory and in practice. The condition on $\min_{j \in \mathrm{supp}(\bar{\beta})} |\bar{\beta}_j|$ also naturally appears in the analysis of Lasso (Zhao and Yu, 2006). In fact, our result shows that the condition of $\min_{j \in \mathrm{supp}(\bar{\beta})} |\bar{\beta}_j|$ required for greedy algorithm is weaker than the corresponding condition for Lasso.

The greedy algorithm analysis employed in this paper is a combination of an observation by Tropp (2004, see Lemma 11) and some technical lemmas for the behavior of greedy least squares regression by Zhang (2008), which are included in Appendix A for completeness. Note that Zhang (2008) only studied a forward-backward procedure, but not the more standard forward greedy algorithm considered here. In this paper, both the employment of the condition $\mu_X(\bar{F}) < 1$ and the proof in Appendix B are new.

As we shall see in this paper, the condition $\mu_X(\bar{F}) \leq 1$ is necessary for the success of the forward greedy procedure. It is worth mentioning that Lasso is consistent in parameter estimation under a weaker sparse eigenvalue condition, even if the condition $\mu_X(\bar{F}) \leq 1$ fails (which means Lasso may not estimate the true feature set correctly): for example, see (Meinshausen and Yu, 2008; Zhang, 2009). Although similar results may be obtained for greedy least squares regression, when the condition $\mu_X(\bar{F}) \leq 1$ fails, it was shown by Zhang (2008) that the performance of greedy algorithm can be improved by incorporating backward steps. In contrast, results in this paper show that if the design matrix satisfies the additional condition $\mu_X(\bar{F}) < 1$, then the standard forward greedy algorithm will be successful without complicated backward steps.

3. Feature Selection using Greedy Least Squares Regression

We would like to establish conditions under which the forward greedy algorithm in Figure 1 never makes any mistake (with large probability), and thus suitable for feature selection. For convenience, we state an assumption before stating the theoretical result.

Assumption 1 *Assume that*

- *The basis functions are normalized such that* $\frac{1}{n}\|\mathbf{x}_j\|_2^2 = 1$ *for all* $j = 1, \ldots, d$.

- *The target function is truly sparse: there exists* $\bar{\beta} \in \mathbb{R}^d$ *with* $\bar{F} = \mathrm{supp}(\bar{\beta})$ *such that* $\mathbf{E}\mathbf{y} = X\bar{\beta}$.
- $\mu_X(\bar{F}) < 1$ *and* $\rho_X(\bar{F}) > 0$.
- $\mathbf{y} = [y_i]_{i=1,\ldots,n}$ *are independent (but not necessarily identically distributed) sub-Gaussians: there exists* $\sigma \geq 0$ *such that* $\forall i \in \{1,\ldots,n\}$ *and* $\forall t \in \mathbb{R}$, $\mathbf{E}_{y_i} e^{t(y_i - \mathbf{E}y_i)} \leq e^{\sigma^2 t^2/2}$.

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

The following theorem gives conditions under which the forward greedy algorithm can identify the correct set of features.

Theorem 1 *Consider the greedy least squares algorithm in Figure 1, where Assumption 1 holds. Given any* $\eta \in (0, 0.5)$*, with probability larger than* $1 - 2\eta$*, if the stopping criterion satisfies*

$$\epsilon > \frac{1}{1 - \mu_X(\bar{F})}\sigma\sqrt{2\ln(2d/\eta)}, \qquad \min_{j \in \bar{F}} |\bar{\beta}_j| \geq 3\epsilon\rho_X(\bar{F})^{-1}/\sqrt{n},$$

then when the procedure stops, we have $F^{(k-1)} = \bar{F}$ *and*

$$\|\beta^{(k-1)} - \bar{\beta}\|_\infty \leq \sigma\sqrt{(2\ln(2|\bar{F}|/\eta))/(n\rho_X(\bar{F}))}.$$

The result is a simple consequence of the following slightly more general theorem (its proof is left to Appendix B).

Theorem 2 *Consider the greedy least squares algorithm in Figure 1, where Assumption 1 holds. Given any* $\eta \in (0, 0.5)$*, with probability larger than* $1 - 2\eta$*, if the stopping criterion satisfies*

$$\epsilon > \frac{1}{1 - \mu_X(\bar{F})}\sigma\sqrt{2\ln(2d/\eta)},$$

then when the procedure stops, the following claims are true:

- $F^{(k-1)} \subset \bar{F}$.
- $|\bar{F} - F^{(k-1)}| \leq 2|\{j \in \bar{F} : |\bar{\beta}_j| < 3\epsilon\rho_X(\bar{F})^{-1}/\sqrt{n}\}|$
- $\|\beta^{(k-1)} - \hat{\beta}_X(\bar{F}, \mathbf{y})\|_2 \leq \epsilon\rho(\bar{F})^{-1}\sqrt{|\bar{F} - F^{(k-1)}|/n}$.
- $\|\hat{\beta}_X(\bar{F}, \mathbf{y}) - \bar{\beta}\|_\infty \leq \sigma\sqrt{(2\ln(2|\bar{F}|/\eta))/(n\rho_X(\bar{F}))}$.

In the following, we discuss some consequences of Theorem 1 and Theorem 2, and compare them with those of Lasso. Let $k(\epsilon)$ be the number of $j \in \bar{F}$ such that $|\bar{\beta}_j| < 3\epsilon\rho_X(\bar{F})^{-1}/\sqrt{n}$. Theorem 2 implies that $|\bar{F} - F^{(k-1)}| \leq 2k(\epsilon)$; that is, $|\bar{F} - F^{(k-1)}|$ is small when $k(\epsilon)$ is small. In such case, the feature set $F^{(k-1)}$ selected by the greedy least squares algorithm is approximately correct. Moreover, we have $\beta^{(k-1)} \approx \bar{\beta}$. In fact, one can show (e.g., see Zhang, 2008) that with probability larger than $1 - \eta$:

$$\|\hat{\beta}_X(\bar{F}, \mathbf{y}) - \bar{\beta}\|_2 \leq \sigma\sqrt{|\bar{F}|/(\rho(\bar{F})n)}[1 + \sqrt{20\ln(1/\eta)}]. \tag{3}$$

By combining this estimate with Theorem 2, we have

$$\|\beta^{(k-1)} - \bar{\beta}\|_2 \leq \sigma\sqrt{|\bar{F}|/(\rho(\bar{F})n)}[1 + \sqrt{20\ln(1/\eta)}] + \epsilon\rho(\bar{F})^{-1}\sqrt{2k(\epsilon)/n}.$$

That is, $\|\beta^{(k-1)} - \bar{\beta}\|_2 = O(\sqrt{|\bar{F}|/n} + \epsilon\sqrt{k(\epsilon)/n})$. This implies that when $\mu_X(\bar{F}) < 1$, greedy least squares regression leads to a good estimation of the true parameter $\bar{\beta}$. By choosing $\epsilon = O(\sigma\sqrt{\ln(2d/\eta)})$, we obtain $\|\beta^{(k-1)} - \bar{\beta}\|_2 = O(\sqrt{|\bar{F}|/n} + \sqrt{k(\epsilon)\ln d/n})$. The corresponding result for the Lasso estimator $\hat{\beta}$ in (2) is $\|\hat{\beta} - \bar{\beta}\|_2 = O(\lambda\sqrt{|\bar{F}|/n})$, where we require λ to be of the order $\sigma\sqrt{\ln(d/\eta)/n}$ or larger. Therefore, if $k(\epsilon)$ is small, then Lasso is inferior due to the extra $\ln d$ factor. This factor is inherent to the L_1 regularization in Lasso, which introduces a bias that cannot be removed.

In this paper, we are mainly interested in the situation $k(\epsilon) = 0$, which implies that with the stopping criterion ϵ, greedy least squares regression can correctly identify all features with large probability. Note that in order to correctly identify all features ($F^{(k-1)} = \bar{F}$), the requirement $\min_{j\in\bar{F}} |\bar{\beta}_j| \geq 3\epsilon\rho_X(\bar{F})^{-1}/\sqrt{n}$ in Theorem 1 is natural. Observe that we may take $\epsilon = \sigma\sqrt{3\ln(d/\eta)}/(1-\mu_X(\bar{F}))$. This means that under the assumption of Theorem 1, it is possible to identify all features correctly using the greedy least squares algorithm as long as the target coefficients $\bar{\beta}_j$ ($j \in \bar{F}$) are larger than the order $\sigma\sqrt{\ln(d/\eta)/n}$.

In fact, since $\sigma\sqrt{\ln(d/\eta)/n}$ is the noise level, if there exists a target coefficient $\bar{\beta}_j$ that is smaller than $O(\sigma\sqrt{\ln(d/\eta)/n})$ in absolute value, then we cannot distinguish such a small coefficient from zero (or noise) with large probability. Therefore when the condition $\mu_X(\bar{F}) < 1$ holds, it is not possible to do much better than greedy least squares regression except for the constant hidden in $O(\cdot)$ and its dependency on $\rho(\bar{F})$ and $\mu_X(\bar{F})$.

In comparison, for Lasso, the condition required of $\min_{j\in\bar{F}} |\bar{\beta}_j|$ depends not only on $\rho(\bar{F})^{-1}$ and $(1-\mu_X(\bar{F}))^{-1}$, but also on the quantity $\|(X_{\bar{F}}^T X_{\bar{F}})^{-1}\|_{\infty,\infty}$ (see Wainwright, 2006), where

$$\|(X_{\bar{F}}^T X_{\bar{F}})^{-1}\|_{\infty,\infty} = \sup_{\mathbf{u}\in\mathbb{R}^{|\bar{F}|}} \frac{\|(X_{\bar{F}}^T X_{\bar{F}})^{-1}\mathbf{u}\|_\infty}{\|\mathbf{u}\|_\infty}.$$

Consider the matrix $(X_{\bar{F}}^T X_{\bar{F}})^{-1} = I + 0.5B/\sqrt{|\bar{F}|}$, where $B_{i,j} = 1$ when either $i = 1$ or $j = 1$ or $i = j$, and $B_{i,j} = 0$ otherwise. Then it is not hard to verify that $\rho(\bar{F})^{-1} < 2$ and $\|(X_{\bar{F}}^T X_{\bar{F}})^{-1}\|_{\infty,\infty} > 0.5\sqrt{|\bar{F}|}$ (by taking $\mathbf{u} = [1,\ldots,1]$). This means that in the worst case, we can find matrix $X_{\bar{F}}^T X_{\bar{F}}$ such that

$$\|(X_{\bar{F}}^T X_{\bar{F}})^{-1}\|_{\infty,\infty} > 0.25\sqrt{|\bar{F}|}\rho(\bar{F})^{-1}.$$

Therefore, if we only assume that $\rho(\bar{F})$ is bounded away from zero without using the quantity $\|(X_{\bar{F}}^T X_{\bar{F}})^{-1}\|_{\infty,\infty}$, the feature consistency result in (Zhao and Yu, 2006; Wainwright, 2006) for Lasso requires the condition

$$\min_{j\in\mathrm{supp}(\bar{\beta})} |\bar{\beta}_j| \geq c\sigma\sqrt{|\bar{F}|\ln(d/\eta)/n}$$

for some constant c that is proportional to $\rho(\bar{F})^{-1}(1-\mu_X(\bar{F}))^{-1}$. This is a more restrictive condition than that of greedy least squares regression. Unfortunately, the factor $\sqrt{|\bar{F}|}$

cannot be removed for Lasso, unless we make the additional and stronger assumption that $\|(X_{\bar{F}}^T X_{\bar{F}})^{-1}\|_{\infty,\infty} = O(\rho(\bar{F})^{-1})$.

As we discussed after Theorem 2, the bias of L_1-regularization also leads to suboptimal estimation for Lasso. For example, for the greedy algorithm, we can show $\|\hat{\beta}_X(\bar{F}, \mathbf{y}) - \bar{\beta}\|_2 = O(\sigma\sqrt{|\bar{F}|/n})$ and $\|\hat{\beta}_X(\bar{F}, \mathbf{y}) - \bar{\beta}\|_\infty = O(\sigma\sqrt{\ln|\bar{F}|/n})$. Under the conditions of Theorem 1, we have $\beta^{(k-1)} = \hat{\beta}_X(\bar{F}, \mathbf{y})$, and thus $\|\beta^{(k-1)} - \bar{\beta}\|_2 = O(\sigma\sqrt{|\bar{F}|/n})$ and $\|\beta^{(k-1)} - \bar{\beta}\|_\infty = O(\sigma\sqrt{\ln|\bar{F}|/n})$. Under the same conditions, for the Lasso estimator $\hat{\beta}$ of (2), we have $\|\hat{\beta} - \bar{\beta}\|_2 = O(\sigma\sqrt{|\bar{F}|\ln d/n})$ and $\|\hat{\beta} - \bar{\beta}\|_\infty = O(\sigma\sqrt{\ln d/n})$. The $\ln d$ factor (bias) is inherent to Lasso, which can be removed with two-stage procedures (e.g., Zhang, 2009). However, such procedures are less robust and more complicated than the simple greedy algorithm.

4. Feature Selection Consistency

As we have mentioned before, the effectiveness of feature selection using Lasso was established by Zhao and Yu (2006), under more refined irrepresentable conditions that depend on the signs of the true target $\mathrm{sgn}(\bar{\beta})$. In comparison, the condition $\mu_X(\bar{F}) < 1$ in Theorem 2 depends only on the design matrix but not the sparse target $\bar{\beta}$. That is, the condition is with respect to the worst case choice of $\bar{\beta}$ with support $\mathrm{supp}(\bar{\beta}) = \bar{F}$. Due to the complexity of greedy procedure, we cannot establish a simple target dependent condition that ensures feature selection consistency. This means for any specific target, the behavior of forward greedy algorithm and Lasso might be different, and one may be preferred over the other under different scenarios. Experiments in (Zhang, 2008) illustrated this point.

In the following, we introduce the target independent irrepresentable conditions that are equivalent to the irrepresentable conditions of Zhao and Yu (2006) with the worst case choice of $\mathrm{sgn}(\bar{\beta})$ (also see Wainwright, 2006).

Definition 3 *Consider a sequence of problems indexed by n: at each sample size n, let $X^{(n)}$ be an $n \times d^{(n)}$ dimensional data matrix, and we observe $\mathbf{y}^{(n)} \in \mathbb{R}^n$ that is corrupted with noise. Let $\bar{F}^{(n)}$ be the feature set, where $\mathbf{E}\mathbf{y}^{(n)} = X^{(n)}\bar{\beta}^{(n)}$ and $\mathrm{supp}(\bar{\beta}^{(n)}) = \bar{F}^{(n)}$.*

We say that the sequence satisfies the strong target independent irrepresentable *condition if there exists $\delta > 0$ such that $\overline{\lim}_{n\to\infty}\mu_{X^{(n)}}(\bar{F}^{(n)}) \leq 1 - \delta$.*

We say that the sequence satisfies the weak target independent irrepresentable *condition if $\mu_{X^{(n)}}(\bar{F}^{(n)}) \leq 1$ for all sufficiently large n.*

It was shown by Zhao and Yu (2006) that the strong (target independent) irrepresentable condition is sufficient for Lasso to select features consistently for all possible sign combination of $\bar{\beta}^{(n)}$ when $n \to \infty$ (under appropriate assumptions). In addition, the weak (target independent) irrepresentable condition is necessary for Lasso to select features consistently when $n \to \infty$. The target independent irrepresentable conditions are considered by Zhao and Yu (2006) and Wainwright (2006). Similar conditions were also considered by Tropp (2004) without stochastic noise.

Results parallel to that of Lasso can be obtained for Algorithm 1. Specifically, the following two theorems show that the strong target independent irrepresentable condition is sufficient for Algorithm 1 to select features consistently, while the weak target independent irrepresentable condition is necessary.

Theorem 4 *Consider regression problems indexed by the sample size n, and use notations in Definition 3. Let Assumption 1 hold, with noise σ independent of n. Assume that the strong irrepresentable condition holds. For each problem of sample size n, denote by F_n the feature set from Algorithm 1 when it stops with $\epsilon = n^{s/2}$ for some $s \in (0, 1]$. Then for all sufficiently large n, we have*

$$P(F_n \neq \bar{F}^{(n)}) \leq \exp(-n^s/\ln n)$$

if the following conditions hold:

1. $d_n \leq \exp(n^s/\ln n)$
2. $\min_{j\in\bar{F}^{(n)}} |\bar{\beta}_j^{(n)}| \geq 3n^{(s-1)/2}/\rho(\bar{F}^{(n)})$.

Proof When n is sufficiently large, the two conditions of Theorem 1 hold with $\eta = 0.5\exp(-n^s/\ln n)$. Therefore the theorem is a direct consequence. ■

Theorem 5 *Consider regression problems indexed by the sample size n, and use notations in Definition 3. Let Assumption 1 hold, with noise σ independent of n. Assume that the weak irrepresentable condition is violated at sample sizes $n_1 < n_2 < \cdots$. There exist targets $\bar{\beta}^{(n_j)}$ with arbitrarily large $\min_{i\in\bar{F}^{(n_j)}} |\bar{\beta}_i^{(n_j)}|$, such that at each sample size n_j, Algorithm 1 chooses a basis $i^{(1)} \notin \bar{F}$ in the first step with probability larger than 0.5.*

Proof By definition of $\mu_X(\bar{F})$, there exists $\mathbf{v} = (X_{\bar{F}}^T X_{\bar{F}})\mathbf{u} \in \mathbb{R}^{|\bar{F}|}$ such that

$$\begin{aligned}
\mu_X(\bar{F}) =& \max_{j\notin\bar{F}} \|(X_{\bar{F}}^T X_{\bar{F}})^{-1} X_{\bar{F}}^T \mathbf{x}_j\|_1 \\
=& \max_{j\notin\bar{F}} \frac{|\mathbf{v}^T (X_{\bar{F}}^T X_{\bar{F}})^{-1} X_{\bar{F}}^T \mathbf{x}_j|}{\|\mathbf{v}\|_\infty} \\
=& \max_{j\notin\bar{F}} \frac{|\mathbf{u}^T X_{\bar{F}}^T \mathbf{x}_j|}{\|(X_{\bar{F}}^T X_{\bar{F}})\mathbf{u}\|_\infty} \\
=& \frac{\max_{j\notin F} |\mathbf{x}_j^T X_{\bar{F}} \mathbf{u}|}{\max_{i\in\bar{F}} |(\mathbf{x}_i^T X_{\bar{F}})\mathbf{u}|}.
\end{aligned}$$

Therefore if $\mu_X(\bar{F}) > 1$, we can find $\mathbf{u} \in \mathbb{R}^{|\bar{F}|}$ such that $\max_{j\notin F} |\mathbf{x}_j^T X_{\bar{F}} \mathbf{u}| > \max_{i\in\bar{F}} |(\mathbf{x}_i^T X_{\bar{F}})\mathbf{u}|$.

Consider an arbitrary sequence $\delta_n > 0$ $(n = 1, 2, \ldots)$. At any sample size $n = n_j$, since $\mu_{X^{(n)}}(\bar{F}^{(n)}) > 1$, we can find a sufficiently large target vector $\bar{\beta}^{(n)}$ such that

$$\max_{i\in\bar{F}^{(n)}} |\mathbf{x}_i^T X^{(n)} \bar{\beta}^{(n)}| < \max_{j\notin\bar{F}^{(n)}} |\mathbf{x}_j^T X^{(n)} \bar{\beta}^{(n)}| - 2\delta_n. \tag{4}$$

Now we may take $\delta_n = \sigma\sqrt{2n\ln(4d_n)}$; then Lemma 8 implies that with probability larger than 0.5, $\max_{j\in\{1,\ldots,d\}} |\mathbf{x}_j^T(\mathbf{y} - X^{(n)}\bar{\beta}^{(n)})| \leq \delta_n$. Therefore (4) implies that

$$\max_{i\in\bar{F}^{(n)}} [|\mathbf{x}_i^T X^{(n)} \bar{\beta}^{(n)}| + |\mathbf{x}_i^T(\mathbf{y} - X^{(n)}\bar{\beta}^{(n)})|] < \max_{j\notin\bar{F}^{(n)}} [|\mathbf{x}_j^T X^{(n)} \bar{\beta}^{(n)}| - |\mathbf{x}_j^T(\mathbf{y} - X^{(n)}\bar{\beta}^{(n)})|].$$

Therefore

$$\max_{i\in\bar{F}^{(n)}} |\mathbf{x}_i^T\mathbf{y}| < \max_{j\notin\bar{F}^{(n)}} |\mathbf{x}_j^T\mathbf{y}|.$$

This means that we pick $i^{(1)} \notin \bar{F}$ in the first step with probability larger than 0.5. ■

5. Conclusion

We have shown that weak and strong target independent irrepresentable conditions are necessary and sufficient conditions for a greedy least squares regression algorithm to select features consistently. These conditions match the target independent versions of the necessary and sufficient conditions for Lasso by Zhao and Yu (2006).

Moreover, if the eigenvalue $\rho(\bar{F})$ is bounded away from zero, then the greedy algorithm can reliably identify features as long as each nonzero coefficient is larger than a constant times the noise level. In comparison, under the same condition, Lasso may require the coefficients to be larger than $O(\sqrt{s})$ times the noise level, where s is the number of nonzero coefficients. This implies that under some conditions, greedy least squares regression may potentially select features more effectively than Lasso in the presence of stochastic noise.

Although the target independent versions of the irrepresentable conditions for greedy least squares regression match those of Lasso, our result does not show which algorithm is better for any specific target. In fact, the target specific behaviors of the two algorithms are different, and one may be preferred over the other under different scenarios.

Acknowledgments

This paper is the outcome of some private discussion with Joel Tropp, which drew the author's attention to the similarities and differences between orthogonal matching pursuit and Lasso for estimating sparse signals.

References

S. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.

Nicolai Meinshausen and Peter Buhlmann. High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics*, 34:1436–1462, 2006.

Nicolai Meinshausen and Bin Yu. Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics*, 2008. to appear.

B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM J. Comput.*, 24(2): 227–234, 1995. ISSN 0097-5397.

Joel A. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Trans. Info. Theory*, 50(10):2231–2242, 2004.

Martin Wainwright. Sharp thresholds for high-dimensional and noisy recovery of sparsity. Technical report, Department of Statistics, UC. Berkeley, 2006.

Tong Zhang. Some sharp performance bounds for least squares regression with L_1 regularization. *The Annals of Statistics*, 2009. to appear.

Tong Zhang. Forward-backward greedy algorithm for learning sparse representations. Technical report, Rutgers Statistics Department, 2008. A short version is to appear in NIPS 08.

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

Appendix A. Auxiliary Lemmas

The following technical lemmas from (Zhang, 2008) are needed to analyze the behavior of greedy least squares regression under stochastic noise. For completeness, we include them here with proofs.

The following lemma relates the squared error reduction in one greedy step to correlation coefficients.

Lemma 6 *For all $\mathbf{x}, \mathbf{x}', \mathbf{y} \in \mathbb{R}^n$, we have*

$$\inf_{\alpha\in\mathbb{R}} \|\mathbf{x} + \alpha\mathbf{x}' - \mathbf{y}\|_2^2 = \|\mathbf{x} - \mathbf{y}\|_2^2 - ((\mathbf{x} - \mathbf{y})^T\mathbf{x}')^2/\|\mathbf{x}'\|_2^2.$$

Proof The equality follows from simple algebra with the optimal α achieved at $-(\mathbf{x} - \mathbf{y})^T\mathbf{x}'/\|\mathbf{x}'\|_2^2$. ■

The following lemma provides a bound on the squared error reduction of one forward greedy step. Some ingredients of the proof has appeared in (Natarajan, 1995).

Lemma 7 *Let Assumption 1 hold. Consider $F \subset \bar{F} \subset \{1, \ldots, d\}$. Let $\beta' = \hat{\beta}_X(\bar{F}, \mathbf{y})$, $\beta = \hat{\beta}_X(F, \mathbf{y})$, $\mathbf{x}' = X\beta'$, and $\mathbf{x} = X\beta$. Then*

$$\inf_{\alpha\in\mathbb{R}, j\in\bar{F}-F} \|\mathbf{x} + \alpha\mathbf{x}_j - \mathbf{y}\|_2^2 \le \|\mathbf{x} - \mathbf{y}\|_2^2 - \frac{\rho(\bar{F})}{|\bar{F} - F|}\|\mathbf{x} - \mathbf{x}'\|_2^2.$$

Proof For all $j \in F$, we have $\|\mathbf{x}+\alpha\mathbf{x}_j-\mathbf{y}\|_2^2$ achieves the minimum at $\alpha = 0$. This implies that $(\mathbf{x} - \mathbf{y})^T\mathbf{x}_j = 0$ for $j \in F$. Therefore we have

$$\begin{aligned}
&(\mathbf{x} - \mathbf{y})^T \sum_{j\in\bar{F}-F} (\beta'_j - \beta_j)\mathbf{x}_j \\
=&(\mathbf{x} - \mathbf{y})^T \sum_{j\in\bar{F}\cup F} (\beta'_j - \beta_j)\mathbf{x}_j = (\mathbf{x} - \mathbf{y})^T(\mathbf{x}' - \mathbf{x}) \\
=&- \|\mathbf{x} - \mathbf{x}'\|_2^2 + (\mathbf{x}' - \mathbf{y})^T(\mathbf{x}' - \mathbf{x}) = -\|\mathbf{x} - \mathbf{x}'\|_2^2.
\end{aligned}$$

The last quality follows from the definition of $\beta' = \hat{\beta}_X(\bar{F}, \mathbf{y})$ and $F \subset \bar{F}$, which implies that $(\mathbf{x}'-\mathbf{y})^T(\mathbf{x}'-\mathbf{x}) = 0$. Now, let $s' = |\bar{F}-F|$, then the above inequality leads to the following

derivation $\forall \eta > 0$:

$$\begin{aligned}
& s' \inf_{j \in \bar{F}-F} \|\mathbf{x} + \eta(\beta'_j - \beta_j)\mathbf{x}_j - \mathbf{y}\|_2^2 \\
\leq & \sum_{j \in \bar{F}-F} \|\mathbf{x} + \eta(\beta'_j - \beta_j)\mathbf{x}_j - \mathbf{y}\|_2^2 \\
= & s'\|\mathbf{x}-\mathbf{y}\|_2^2 + \eta^2 \sum_{j \in \bar{F}-F} (\beta'_j - \beta_j)^2 \|\mathbf{x}_j\|_2^2 + 2\eta(\mathbf{x}-\mathbf{y})^T \sum_{j \in \bar{F}-F} (\beta'_j - \beta_j)\mathbf{x}_j \\
= & s'\|\mathbf{x}-\mathbf{y}\|_2^2 + n\eta^2 \sum_{j \in \bar{F}-F} (\beta'_j - \beta_j)^2 - 2\eta\|\mathbf{x}'-\mathbf{x}\|_2^2.
\end{aligned}$$

Note that in the last equation, we have used $\|\mathbf{x}_j\|_2^2 = n$ in Assumption 1. By optimizing over η, we obtain

$$\begin{aligned}
& s' \inf_{j \in \bar{F}-F} \|\mathbf{x} + \eta(\beta'_j - \beta_j)\mathbf{x}_j - \mathbf{y}\|_2^2 \\
\leq & s'\|\mathbf{x}-\mathbf{y}\|_2^2 - \frac{\|\mathbf{x}'-\mathbf{x}\|_2^4}{n\sum_{j \in \bar{F}}(\beta'_j - \beta_j)^2} \leq s'\|\mathbf{x}-\mathbf{y}\|_2^2 - \rho(\bar{F})\|\mathbf{x}'-\mathbf{x}\|_2^2.
\end{aligned}$$

This leads to the lemma. ■

The following lemma is a standard empirical processes bound for sub-Gaussian random variables. The bound is used to derive probability estimates in our analysis.

Lemma 8 *Consider n independent random variables $\xi_1, \ldots, \xi_n$ such that $\mathbf{E}e^{t(\xi_i - \mathbf{E}\xi_i)} \leq e^{\sigma_i^2 t^2/2}$ for all t and i. Consider $g_{i,j}$ for $i = 1, \ldots, n$ and $j = 1, \ldots, m$, we have for all $\eta \in (0,1)$, with probability larger than $1-\eta$:*

$$\sup_j |\sum_{i=1}^n g_{i,j}(\xi_i - \mathbf{E}\xi_i)| \leq a\sqrt{2\ln(2m/\eta)},$$

where $a^2 = \sup_j \sum_{i=1}^n g_{i,j}^2 \sigma_i^2$.

Proof For a fixed j, we let $s_j = \sum_{i=1}^n g_{i,j}(\xi_i - \mathbf{E}\xi_i)$; then by assumption, $\mathbf{E}(e^{ts_j} + e^{-ts_j}) \leq 2e^{a^2t^2/2}$, which implies that for all $\epsilon > 0$: $P(|s_j| \geq \epsilon)e^{t\epsilon} \leq 2e^{a^2t^2/2}$. Now let $t = \epsilon/a^2$, we obtain:

$$P\left(\left|\sum_{i=1}^n g_{i,j}(\xi_i - \mathbf{E}\xi_i)\right| \geq \epsilon\right) \leq 2e^{-\epsilon^2/2a^2}.$$

This implies that

$$P\left[\sup_j \left|\sum_{i=1}^n g_{i,j}(\xi_i - \mathbf{E}\xi_i)\right| \geq \epsilon\right] \leq m \sup_j P\left[\left|\sum_{i=1}^n g_{i,j}(\xi_i - \mathbf{E}\xi_i)\right| \geq \epsilon\right] \leq 2me^{-\epsilon^2/(2a^2)}.$$

This implies the lemma. ■

The following lemma gives a bound on the infinity norm of the difference between the estimated parameter $\hat{\beta}_X(\bar{F}, \mathbf{y})$ and the true parameter $\bar{\beta}$ when the set of features $\bar{F}$ are known in advance.

Lemma 9 *Let Assumption 1 hold. Consider any fixed $\bar{F} \subset \{1,\ldots,d\}$. For all $\eta \in (0,1)$, with probability larger than $1-\eta$, we have*

$$\|\hat{\beta}_X(\bar{F},\mathbf{y}) - \hat{\beta}_X(\bar{F},\mathbf{E}\mathbf{y})\|_\infty \leq \sigma\sqrt{(2\ln(2|\bar{F}|/\eta))/(n\rho(\bar{F}))}.$$

Proof For simplicity, let $G = X_{\bar{F}}$ and $\bar{k} = |\bar{F}|$. Let $\hat{\beta}' \in \mathbb{R}^{\bar{k}}$ and $\bar{\beta}' \in \mathbb{R}^{\bar{k}}$ be the restrictions of $\hat{\beta}_X(\bar{F},\mathbf{y}) \in \mathbb{R}^d$ and $\hat{\beta}_X(\bar{F},\mathbf{E}\mathbf{y}) \in \mathbb{R}^d$ to $\bar{F}$ respectively. Algebraically, we have $\hat{\beta}' = (G^TG)^{-1}G^T\mathbf{y}$ and $\bar{\beta}' = (G^TG)^{-1}G^T\mathbf{E}\mathbf{y}$. It follows that

$$\hat{\beta}' - \bar{\beta}' = (G^TG)^{-1}G^T(\mathbf{y} - \mathbf{E}\mathbf{y}).$$

Therefore

$$|\hat{\beta}'_j - \bar{\beta}'_j| = |\mathbf{e}_j^T(G^TG)^{-1}G^T(\mathbf{y} - \mathbf{E}\mathbf{y})|.$$

Lemma 8 implies that with probability larger than $1-\eta$, for all j:

$$|\mathbf{e}_j^T(G^TG)^{-1}G^T(\mathbf{y} - \mathbf{E}\mathbf{y})| \leq \sigma\|\mathbf{e}_j^T(G^TG)^{-1}G^T\|_2\sqrt{2\ln(2\bar{k}/\eta)}.$$

Since by definition, $\rho(\bar{F})n$ is no larger than the smallest eigenvalue of G^TG, the desired inequality follows from the estimate

$$\|\mathbf{e}_j^T(G^TG)^{-1}G^T\|_2^2 = \mathbf{e}_j^T(G^TG)^{-1}\mathbf{e}_j \leq 1/(n\rho(\bar{F})).$$

■

The following lemma estimates the correlation coefficient of $\mathbf{x}_j$ with $j \notin \bar{F}$.

Lemma 10 *Let Assumption 1 hold. Assume also that $\mathbf{E}\mathbf{y} = X\bar{\beta}$ with $\mathrm{supp}(\bar{\beta}) \subset \bar{F} \subset \{1,\ldots,d\}$. For all $\eta \in (0,1)$, with probability larger than $1-\eta$, we have*

$$\max_{j\notin\bar{F}} |(X\hat{\beta}_X(\bar{F},\mathbf{y}) - \mathbf{y})^T\mathbf{x}_j| \leq \sigma\sqrt{2n\ln(2d/\eta)}.$$

Proof Let P be the projection operator to the subspace spanned by $\{\mathbf{x}_j : j \in \bar{F}\}$ on $\mathbb{R}^n$. Lemma 8 implies that with probability larger than $1-\eta$:

$$\sup_{j=1,\ldots,d} |(\mathbf{y} - \mathbf{E}\mathbf{y})^T(I-P)\mathbf{x}_j| \leq \sigma\sqrt{2n\ln(2d/\eta)},$$

where we have used $\|\mathbf{x}_j\|_2 = \sqrt{n}$. Let $\hat{\mathbf{x}} = X\hat{\beta}_X(\bar{F},\mathbf{y})$. Since $(\hat{\mathbf{x}} - \mathbf{y})^T\mathbf{x}_j = 0$ for all $j \in \bar{F}$, we have $\hat{\mathbf{x}} = P\mathbf{y}$. Moreover, since $(I-P)\mathbf{E}\mathbf{y} = (I-P)X\bar{\beta} = 0$, we have

$$(\mathbf{y} - \mathbf{E}\mathbf{y})^T(I-P)\mathbf{x}_j = ((I-P)\mathbf{y} - (I-P)\mathbf{E}\mathbf{y})^T\mathbf{x}_j = (\mathbf{y} - \hat{\mathbf{x}})^T\mathbf{x}_j.$$

Combine this equation with the previous inequality, we obtain the desired bound. ■

Appendix B. Proof of Theorem 2

Before the formal proof, we give a brief outline of the main argument. First, Lemma 10 implies that with large probability, $\max_{j\notin\bar{F}}|(X\hat{\beta}_X(\bar{F},\mathbf{y})-\mathbf{y})^T\mathbf{x}_j|$ is small. Using this fact, and the assumption that $\mu_X(\bar{F})<1$, it follows from Lemma 11 below that when $\mathrm{supp}(\beta^{(k-1)})\subset\bar{F}$, either $\max_i|(X\beta^{(k-1)}-\mathbf{y})^T\mathbf{x}_i|$ is sufficiently small (which implies that the greedy procedure stops and $\beta^{(k-1)}\approx\bar{\beta}$), or $\max_{j\notin\bar{F}}|(X\beta^{(k-1)}-\mathbf{y})^T\mathbf{x}_j|<\max_{i\in\bar{F}}|(X\beta^{(k-1)}-\mathbf{y})^T\mathbf{x}_i|$ (which implies that the greedy procedure chooses a direction $i^{(k)}\in\bar{F}$ in the next iteration). The claims of the theorem then follow by induction.

We start the formal proof by introducing the following critical lemma, which generalizes the essential idea of Tropp (2004). Note that the result there only considered the case $X\hat{\beta}_X(\bar{F},\mathbf{y})=\mathbf{y}$.

Lemma 11 *Consider $\beta\in\mathbb{R}^d$ such that $\mathrm{supp}(\beta)\subset\bar{F}$. We have*

$$\max_{j\notin\bar{F}}|(X\beta-\mathbf{y})^T\mathbf{x}_j|\leq\max_{j\notin\bar{F}}|(X\hat{\beta}_X(\bar{F},\mathbf{y})-\mathbf{y})^T\mathbf{x}_j|+\mu_X(\bar{F})\max_{i\in\bar{F}}|(X\beta-\mathbf{y})^T\mathbf{x}_i|.$$

Proof Let $\beta'=\hat{\beta}_X(\bar{F},\mathbf{y})$. Note that $(X\beta'-\mathbf{y})^T\mathbf{x}_i=0$ when $i\in\bar{F}$. Therefore

$$\max_{i\in\bar{F}}|(X\beta-\mathbf{y})^T\mathbf{x}_i|=\max_{i\in\bar{F}}|\mathbf{x}_i^TX(\beta-\beta')|=\|X_{\bar{F}}^TX_{\bar{F}}(\beta-\beta')\|_\infty.$$

Let $\mathbf{v}=X_{\bar{F}}^TX_{\bar{F}}(\beta-\beta')$, then the definition of $\mu_X(\bar{F})$ implies that

$$\mu_X(\bar{F})\geq\max_{j\notin\bar{F}}\frac{|\mathbf{x}_j^TX_{\bar{F}}(X_{\bar{F}}^TX_{\bar{F}})^{-1}\mathbf{v}|}{\|\mathbf{v}\|_\infty}=\frac{\max_{j\notin\bar{F}}|\mathbf{x}_j^TX_{\bar{F}}(\beta-\beta')|}{\|X_{\bar{F}}^TX_{\bar{F}}(\beta-\beta')\|_\infty}.$$

We obtain from the above

$$\max_{j\notin\bar{F}}|\mathbf{x}_j^TX(\beta-\beta')|\leq\mu_X(\bar{F})\|X_{\bar{F}}^TX_{\bar{F}}(\beta-\beta')\|_\infty=\mu_X(\bar{F})\max_{i\in\bar{F}}|(X\beta-\mathbf{y})^T\mathbf{x}_i|.$$

Now, the lemma follows from the simple inequality

$$\max_{j\notin\bar{F}}|(X\beta-\mathbf{y})^T\mathbf{x}_j|\leq\max_{j\notin\bar{F}}|\mathbf{x}_j^TX(\beta-\beta')|+\max_{j\notin\bar{F}}|\mathbf{x}_j^T(X\beta'-\mathbf{y})|$$

■

Now we are ready to prove the theorem. From Lemma 9, we obtain with probability larger than $1-\eta$,

$$\|\hat{\beta}_X(\bar{F},\mathbf{y})-\bar{\beta}\|_\infty\leq\sigma\sqrt{(2\ln(2|\bar{F}|/\eta))/(n\rho(\bar{F}))}.$$

From Lemma 10, we obtain with probability larger than $1-\eta$,

$$\max_{j\notin\bar{F}}|(X(\hat{\beta}_X(\bar{F},\mathbf{y}))-\mathbf{y})^T\tilde{\mathbf{x}}_j|\leq\sigma\sqrt{2\ln(2d/\eta)}<(1-\mu_X(\bar{F}))\epsilon. \tag{5}$$

With probability larger than $1-2\eta$, both claims hold.

We now proceed by induction on k to show that $F^{(k-1)} \subset \bar{F}$ before the procedure stops. Assume the claim is true after $k-1$ steps for $k \geq 1$. By induction hypothesis, we have $F^{(k-1)} \subset \bar{F}$ at the beginning of step k. Therefore Lemma 7 implies

$$\min_{\alpha, i \in \bar{F}} \|X\beta^{(k-1)} + \alpha \mathbf{x}_i - \mathbf{y}\|_2^2 \leq \|X\beta^{(k-1)} - \mathbf{y}\|_2^2 - \frac{\rho(\bar{F})}{|\bar{F} - F^{(k-1)}|} \|X(\beta^{(k-1)} - \hat{\beta}_X(\bar{F}, \mathbf{y}))\|_2^2. \quad (6)$$

Using the above inequality, we obtain from Lemma 6 the following bound on the correlation coefficients:

$$\begin{aligned}
\max_{i \in \bar{F}} |(X\beta^{(k-1)} - \mathbf{y})^T \tilde{\mathbf{x}}_i| &= \left(\|X\beta^{(k-1)} - \mathbf{y}\|_2^2 - \min_{\alpha, i \in \bar{F}} \|X\beta^{(k-1)} + \alpha \mathbf{x}_i - \mathbf{y}\|_2^2 \right)^{1/2} \\
&\geq \frac{\sqrt{\rho(\bar{F})}}{\sqrt{|\bar{F} - F^{(k-1)}|}} \|X(\beta^{(k-1)} - \hat{\beta}_X(\bar{F}, \mathbf{y}))\|_2 \\
&\geq \frac{\sqrt{n}\rho(\bar{F})}{\sqrt{|\bar{F} - F^{(k-1)}|}} \|\beta^{(k-1)} - \hat{\beta}_X(\bar{F}, \mathbf{y})\|_2.
\end{aligned}$$

We only need to consider the following two scenarios:

- $\|\beta^{(k-1)} - \hat{\beta}_X(\bar{F}, \mathbf{y})\|_2 > \epsilon \rho(\bar{F})^{-1} \sqrt{|\bar{F} - F^{(k-1)}|/n}$.

  In this case, we have

  $$\max_{i \in \bar{F}} |(X\beta^{(k-1)} - \mathbf{y})^T \tilde{\mathbf{x}}_i| > \epsilon > \max_{j \notin \bar{F}} |(X\hat{\beta}_X(\bar{F}, \mathbf{y}) - \mathbf{y})^T \tilde{\mathbf{x}}_j| / (1 - \mu_X(\bar{F})).$$

  The second inequality is due to (5). Now by combining this estimate with Lemma 11, we obtain

  $$\max_{j \notin \bar{F}} |(X\beta^{(k-1)} - \mathbf{y})^T \tilde{\mathbf{x}}_j| < \max_{i \in \bar{F}} |(X\beta^{(k-1)} - \mathbf{y})^T \tilde{\mathbf{x}}_i|.$$

  This implies that $i^{(k)} \in \bar{F}$ and the procedure does not stop.

- $\|\beta^{(k-1)} - \hat{\beta}_X(\bar{F}, \mathbf{y})\|_2 \leq \epsilon \rho(\bar{F})^{-1} \sqrt{|\bar{F} - F^{(k-1)}|/n}$.

  In this case, we have the following scenarios:

  1. $i^{(k)} \notin \bar{F}$: By Lemma 11, we must have:

     $$\begin{aligned}
     \max_{i \in \bar{F}} |(X\beta - \mathbf{y})^T \tilde{\mathbf{x}}_i| &\leq \max_{j \notin \bar{F}} |(X\beta - \mathbf{y})^T \tilde{\mathbf{x}}_j| \\
     &\leq \max_{j \notin \bar{F}} |(X\hat{\beta}_X(\bar{F}, \mathbf{y}) - \mathbf{y})^T \tilde{\mathbf{x}}_j| + \mu_X(\bar{F}) \max_{i \in \bar{F}} |(X\beta - \mathbf{y})^T \tilde{\mathbf{x}}_i| \\
     &\leq \max_{j \notin \bar{F}} |(X\hat{\beta}_X(\bar{F}, \mathbf{y}) - \mathbf{y})^T \tilde{\mathbf{x}}_j| + \mu_X(\bar{F}) \max_{j \notin \bar{F}} |(X\beta - \mathbf{y})^T \tilde{\mathbf{x}}_j|.
     \end{aligned}$$

     Therefore

     $$\max_{j \notin \bar{F}} |(X\beta - \mathbf{y})^T \tilde{\mathbf{x}}_j| \leq \frac{1}{1 - \mu_X(\bar{F})} \max_{j \notin \bar{F}} |(X\hat{\beta}_X(\bar{F}, \mathbf{y}) - \mathbf{y})^T \tilde{\mathbf{x}}_j| < \epsilon.$$

     The last inequality is due to (5). This implies that the procedure stops.

2. $i^{(k)} \in \bar{F}$ and the procedure does not stop.
3. $i^{(k)} \in \bar{F}$ and the procedure stops.

The above situations imply that if the procedure does not stop, then $i^{(k)} \in \bar{F}$. If the procedure stops, then

$$\|\beta^{(k-1)} - \hat{\beta}_X(\bar{F}, \mathbf{y})\|_2 \leq \epsilon \rho(\bar{F})^{-1} \sqrt{|\bar{F} - F^{(k-1)}|/n}.$$

Therefore by induction, when the procedure stops, the following three claims hold:

- $F^{(k-1)} \subset \bar{F}$.
- $\|\beta^{(k-1)} - \hat{\beta}_X(\bar{F}, \mathbf{y})\|_2 \leq \epsilon \rho_X(\bar{F})^{-1} \sqrt{|\bar{F} - F^{(k-1)}|/n}$.
- $\|\hat{\beta}_X(\bar{F}, \mathbf{y}) - \bar{\beta}\|_\infty \leq \sigma \sqrt{(2\ln(2|\bar{F}|/\eta))/(n\rho_X(\bar{F}))} < \epsilon/\sqrt{n\rho_X(\bar{F})}$.

Now, we can let $\gamma = \sqrt{8}\epsilon\rho_X(\bar{F})^{-1}/\sqrt{n}$, then the above claims imply that

$$\begin{aligned}
&\gamma\sqrt{|\{j \in \bar{F} - F^{(k-1)} : |\bar{\beta}_j| \geq \gamma\}|} \\
\leq &\left[\sum_{j \in \bar{F} - F^{(k-1)}} |\bar{\beta}_j|^2\right]^{1/2} \\
\leq &\left[\sum_{j \in \bar{F} - F^{(k-1)}} |\bar{\beta}_j - \hat{\beta}_X(\bar{F}, \mathbf{y})|^2\right]^{1/2} + \left[\sum_{j \in \bar{F} - F^{(k-1)}} |\hat{\beta}_X(\bar{F}, \mathbf{y})|^2\right]^{1/2} \\
\leq &\sqrt{|\bar{F} - F^{(k-1)}|}\|\bar{\beta} - \hat{\beta}_X(\bar{F}, \mathbf{y})\|_\infty + \|\beta^{(k-1)} - \hat{\beta}_X(\bar{F}, \mathbf{y})\|_2 \\
< &\sqrt{|\bar{F} - F^{(k-1)}|/(n\rho_X(\bar{F}))}\epsilon + \epsilon\rho_X(\bar{F})^{-1}\sqrt{|\bar{F} - F^{(k-1)}|/n} \\
\leq &2\epsilon\rho_X(\bar{F})^{-1}\sqrt{|\bar{F} - F^{(k-1)}|/n} = \gamma\sqrt{|\bar{F} - F^{(k-1)}|/2}.
\end{aligned}$$

Therefore

$$2|\{j \in \bar{F} - F^{(k-1)} : |\bar{\beta}_j| \geq \gamma\}| \leq |\bar{F} - F^{(k-1)}|,$$

which implies that

$$\begin{aligned}
|\bar{F} - F^{(k-1)}| = &2|\bar{F} - F^{(k-1)}| - |\bar{F} - F^{(k-1)}| \\
\leq &2|\bar{F} - F^{(k-1)}| - 2|\{j \in \bar{F} - F^{(k-1)} : |\bar{\beta}_j| \geq \gamma\}| \\
= &2|\{j \in \bar{F} - F^{(k-1)} : |\bar{\beta}_j| < \gamma\}| \\
\leq &2|\{j \in \bar{F} : |\bar{\beta}_j| < \gamma\}|.
\end{aligned}$$

This proves the theorem.