On the Differences between L_2Boosting and the Lasso

Michael Vogt*
University of Bonn

We prove that L_2Boosting lacks a theoretical property which is central to the behaviour of ℓ_1-penalized methods such as basis pursuit and the Lasso: Whereas ℓ_1-penalized methods are guaranteed to recover the sparse parameter vector in a high-dimensional linear model under an appropriate restricted nullspace property, L_2Boosting is not guaranteed to do so. Hence, L_2Boosting behaves quite differently from ℓ_1-penalized methods when it comes to parameter recovery/estimation in high-dimensional linear models.

Key words: L_2Boosting; high-dimensional linear models; parameter recovery/estimation; restricted nullspace property; restricted eigenvalue condition.
AMS 2010 subject classifications: 62J05; 62J07; 68Q32.

1 Introduction

The main aim of this paper is to point out an important theoretical difference between L_2Boosting and ℓ_1-penalized methods such as basis pursuit [6] and the Lasso [23]. To do so, we consider the high-dimensional linear model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} \tag{1.1}$$

without noise, where $\boldsymbol{Y} \in \mathbb{R}^n$ is the observation vector, $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ is the design matrix with $p > n$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^p$ denotes the parameter vector. Suppose that $\boldsymbol{\beta}$ is the unique sparsest solution of the equation $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{b}$ and let $S = \{j : \beta_j \neq 0\}$ be the active set with the sparsity index $s = |S|$. Our main result is negative. It shows that L_2Boosting lacks a theoretical property which is central to the behaviour of ℓ_1-penalized methods such as basis pursuit and the Lasso: Whereas ℓ_1-penalized methods are guaranteed to recover the sparse parameter vector $\boldsymbol{\beta}$ in model (1.1) under an appropriate restricted nullspace property, L_2Boosting is not guaranteed to do so.
More formally speaking, we prove the following result: Let $\|\cdot\|_1$ be the usual ℓ_1-norm for vectors and let $S^c = \{1, \ldots, p\} \setminus S$ be the complement of S. Moreover, for any vector $\boldsymbol{b} = (b_1, \ldots, b_p)^\top \in \mathbb{R}^p$ and any index set $\mathcal{T} \subseteq \{1, \ldots, p\}$, define

*Address: Department of Economics and Hausdorff Center for Mathematics, University of Bonn, 53113 Bonn, Germany. Email: michael.vogt@uni-bonn.de.

$\boldsymbol{b}_{\mathcal{T}} = (b_{\mathcal{T},1}, \ldots, b_{\mathcal{T},p})^\top \in \mathbb{R}^p$ by setting $b_{\mathcal{T},j} = b_j \, 1(j \in \mathcal{T})$ for $1 \leq j \leq p$. The design matrix $\boldsymbol{X}$ is said to fulfill the restricted nullspace property $\mathrm{RN}(S, L)$ for the index set S and the constant $L > 0$ if the cone $\mathbb{C}(S, L) = \{\boldsymbol{b} \in \mathbb{R}^p : \|\boldsymbol{b}_{S^c}\|_1 \leq L\|\boldsymbol{b}_S\|_1\}$ and the nullspace $\mathcal{N}(\boldsymbol{X})$ of $\boldsymbol{X}$ only have the zero vector in common, that is, if $\mathbb{C}(S, L) \cap \mathcal{N}(\boldsymbol{X}) = \{\boldsymbol{0}\}$. As is well-known, basis pursuit and the Lasso are guaranteed to recover the sparse vector $\boldsymbol{\beta}$ under the restricted nullspace property $\mathrm{RN}(S, L)$ with $L \geq 1$. We prove that L_2Boosting, in contrast, may fail to recover $\boldsymbol{\beta}$ under $\mathrm{RN}(S, L)$ no matter how large L. In particular, for any $L > 0$, we construct a matrix $\boldsymbol{X}$ with the property $\mathrm{RN}(S, L)$ and a vector $\boldsymbol{\beta}$ which is the unique sparsest solution of $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{b}$ such that the parameter estimate $\boldsymbol{\beta}^{[k]}$ produced by the L_2Boosting algorithm in the k-th iteration step does not converge to $\boldsymbol{\beta}$ as $k \to \infty$. Hence, L_2Boosting fails to recover the sparse parameter vector $\boldsymbol{\beta}$.

According to this negative result (which, to the best of our knowledge, has not been known so far), L_2Boosting behaves quite differently from ℓ_1-penalized methods when it comes to parameter recovery/estimation in high-dimensional linear models. This comes a bit as a surprise. As L_2Boosting is usually considered to act similar to ℓ_1-penalized methods, one may have expected that it is guaranteed to recover the parameter vector $\boldsymbol{\beta}$ under $\mathrm{RN}(S, L)$ at least for sufficiently large L. There are indeed close connections between L_2Boosting and ℓ_1-penalized methods such as the Lasso. In a very influential paper, Efron et al. [11] established close similarities between the Lasso and forward stagewise linear regression (FS), which is a near relative of L_2Boosting. They proved that the limiting version of FS with step size approaching zero (denoted by FS_0 for short) and the Lasso can be obtained as simple modifications of the least angle regression (LARS) algorithm. In addition, they showed that FS_0 and the Lasso coincide under a positive cone condition, which is related to but more general than orthogonality of the design matrix. Exact equivalence of L_2Boosting and the Lasso in an orthonormal linear model was proven in Bühlmann & Yu [3]. Despite these close connections, L_2Boosting and the Lasso are not the same in general but may produce quite different solution paths. This has been observed in numerous simulation and application studies; cp. for example the recent comparison study by Hepp et al. [17]. The exact theoretical reasons for this difference in behaviour are however not fully understood so far. An important step towards a better understanding was made in Hastie et al. [16]. They showed that FS_0 can be characterized as a restricted version of the Lasso with certain monotonicity constraints, which explains why the FS_0 paths are often much smoother than the Lasso paths. The negative result of this paper is a further step towards a more thorough understanding of the theoretical differences between L_2Boosting and the Lasso.

2 Notation and definitions

2.1 Notation

We briefly summarize the notation used in the paper. As already mentioned in the Introduction, for a general index set $\mathcal{T} \subseteq \{1, \ldots, p\}$ and any vector $\boldsymbol{b} = (b_1, \ldots, b_p)^\top \in \mathbb{R}^p$, we define the vector $\boldsymbol{b}_\mathcal{T} = (b_{\mathcal{T},1}, \ldots, b_{\mathcal{T},p})^\top \in \mathbb{R}^p$ by setting $b_{\mathcal{T},j} = b_j \, 1(j \in \mathcal{T})$ for $1 \leq j \leq p$. We further write $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p)$, where $\boldsymbol{X}_j$ denotes the j-th column of the design matrix $\boldsymbol{X}$. As usual, the symbol $\|v\|_q$ denotes the ℓ_q-norm of a generic vector $v \in \mathbb{R}^N$ for $q \in \mathbb{N} \cup \{\infty\}$. In addition, we use the symbol $\langle v, w \rangle = v^\top w$ to denote the inner product of vectors $v, w \in \mathbb{R}^N$. Finally, the cardinality of a set $\mathcal{T}$ is denoted by $|\mathcal{T}|$.

2.2 L_2Boosting

Boosting methods were originally proposed for classification and go back to Schapire [20] and Freund & Schapire [13]. Since then, a variety of boosting algorithms have been developed for different purposes. The L_2Boosting algorithm has been investigated in the statistics, the signal processing and the approximation literature under different names. In statistics, L_2Boosting methods for regression were developed by Friedman [14]. In signal processing, L_2Boosting is known as matching pursuit and was introduced by Mallat & Zhang [18] and Qian & Chen [19]. In the approximation literature, it goes under the name of pure greedy algorithm; cp. for example Temlyakov's monograph [22]. Prediction consistency of L_2Boosting in high-dimensional linear models was derived by Temlyakov [21] in the noiseless case and by Bühlmann [2] in the noisy case. Results on support recovery for greedy algorithms related to L_2Boosting were established in Tropp [24] and Donoho et al. [9] among others.

The L_2Boosting algorithm proceeds as follows:

Step 0: Initialize the residual and parameter vector by $\boldsymbol{R}^{[0]} = \boldsymbol{Y}$ and $\boldsymbol{\beta}^{[0]} = \boldsymbol{0} \in \mathbb{R}^p$.

Step k: Let $\boldsymbol{R}^{[k-1]} \in \mathbb{R}^n$ and $\boldsymbol{\beta}^{[k-1]} \in \mathbb{R}^p$ be the residual and parameter vector from the previous iteration step. For each $1 \leq j \leq p$, compute the univariate least-squares estimate $\hat{b}_j = \langle \boldsymbol{R}^{[k-1]}, \boldsymbol{X}_j \rangle / \|\boldsymbol{X}_j\|_2^2$ and define the index $j_k \in \{1, \ldots, p\}$ by

$$j_k = \arg\min_{1 \leq j \leq p} \left\| \boldsymbol{R}^{[k-1]} - \hat{b}_j \boldsymbol{X}_j \right\|_2^2. \tag{2.1}$$

Equivalently, j_k can be defined as

$$j_k = \arg\max_{1 \leq j \leq p} \left| \left\langle \boldsymbol{R}^{[k-1]}, \frac{\boldsymbol{X}_j}{\|\boldsymbol{X}_j\|_2} \right\rangle \right|. \tag{2.2}$$

In case of ties, let j_k be the smallest index which fulfills (2.1). Update $\boldsymbol{R}^{[k-1]}$ and $\boldsymbol{\beta}^{[k-1]}$ by

$$\boldsymbol{R}^{[k]} = \boldsymbol{R}^{[k-1]} - \nu \hat{b}_{j_k} \boldsymbol{X}_{j_k} \tag{2.3}$$

and

$$\boldsymbol{\beta}^{[k]} = \boldsymbol{\beta}^{[k-1]} + \nu \hat{b}_{j_k} \boldsymbol{e}_{j_k}, \tag{2.4}$$

where $\boldsymbol{e}_j$ is the j-th standard basis vector of $\mathbb{R}^p$ and $\nu \in (0, 1]$ is a pre-specified step length.

Iterate this procedure until some stopping criterion is satisfied or until some maximal number of iterations is reached.

2.3 Restricted nullspace, restricted eigenvalue and restricted isometry properties

Let $\mathcal{T} \subseteq \{1, \ldots, p\}$ be an arbitrary index set and let L be a positive real constant. Define the cone $\mathbb{C}(\mathcal{T}, L) = \{\boldsymbol{b} \in \mathbb{R}^p : \|\boldsymbol{b}_{\mathcal{T}^c}\|_1 \leq L\|\boldsymbol{b}_{\mathcal{T}}\|_1\}$ and denote the nullspace of $\boldsymbol{X}$ by $\mathcal{N}(\boldsymbol{X})$.

Definition 1. *The design matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ satisfies the restricted nullspace property* $\mathrm{RN}(\mathcal{T}, L)$ *for the index set $\mathcal{T}$ and the constant L if*

$$\mathbb{C}(\mathcal{T}, L) \cap \mathcal{N}(\boldsymbol{X}) = \{\boldsymbol{0}\}.$$

If $\boldsymbol{X}$ satisfies $\mathrm{RN}(\mathcal{T}, L)$ *for any index set $\mathcal{T}$ with $|\mathcal{T}| \leq t$, we say that it fulfills the uniform restricted nullspace property* $\mathrm{RN}_{\mathrm{unif}}(t, L)$ *of order t.*

The $\mathrm{RN}(\mathcal{T}, L)$ property restricts the nullspace $\mathcal{N}(\boldsymbol{X})$ not to intersect with the cone $\mathbb{C}(\mathcal{T}, L)$ except at zero. The cone region $\mathbb{C}(\mathcal{T}, L)$ gets larger with increasing L. Hence, the $\mathrm{RN}(\mathcal{T}, L)$ property becomes more restrictive as L increases. Further discussion of the restricted nullspace property can be found in Donoho & Huo [10], Feuer & Nemirovski [12] and Cohen, Dahmen & DeVore [7] among others.
The $\mathrm{RN}(\mathcal{T}, L)$ property is closely related to restricted eigenvalue properties which were introduced in Bickel et al. [1] and are frequently used in the context of the Lasso. In particular, $\mathrm{RN}(\mathcal{T}, L)$ is equivalent to the following restricted eigenvalue property.

Definition 2. *The design matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ satisfies the restricted eigenvalue property* $\mathrm{RE}(\mathcal{T}, L)$ *for the index set $\mathcal{T}$ and the constant L if there exists a constant $\phi > 0$ with*

$$\frac{\|\boldsymbol{X}\boldsymbol{b}\|_2^2}{\|\boldsymbol{b}\|_2^2} \geq \phi \quad \textit{for all non-zero } \boldsymbol{b} \in \mathbb{C}(\mathcal{T}, L).$$

Sufficient conditions for restricted nullspace and eigenvalue properties are often formulated in terms of restricted isometry [4]. The t-restricted isometry constant δ_t of the matrix $\boldsymbol{X}$ is defined as the smallest non-negative number such that

$$(1 - \delta_t) \leq \frac{\|\boldsymbol{X}\boldsymbol{b}\|_2^2}{\|\boldsymbol{b}\|_2^2} \leq (1 + \delta_t)$$

for any non-zero $\mathcal{T}$-sparse vector $\boldsymbol{b}$ whose active set $\mathcal{T}$ has cardinality $|\mathcal{T}| \leq t$. There are several results in the literature which show that the uniform restricted nullspace property $\mathrm{RN}_{\mathrm{unif}}(t, 1)$ holds true if the restricted isometry constants δ_t and δ_{2t} fulfill certain conditions. An example is the following result: If the restricted isometry constant δ_{2t} of the matrix $\boldsymbol{X}$ is such that $\delta_{2t} < 1/3$, then $\boldsymbol{X}$ has the uniform restricted nullspace property $\mathrm{RN}_{\mathrm{unif}}(t, 1)$. Conceptually, restricted isometry conditions are substantially stronger than restricted nullspace/eigenvalue conditions: Restricted isometry conditions require that both a lower and upper bound of the form $(1-\delta) \leq \|\boldsymbol{X}\boldsymbol{b}\|_2^2/\|\boldsymbol{b}\|_2^2 \leq (1+\delta)$ hold for all vectors $\boldsymbol{b}$ that fulfill certain sparsity constraints. Restricted nullspace/eigenvalue conditions, in contrast, only require a lower bound of the form $(1-\delta) \leq \|\boldsymbol{X}\boldsymbol{b}\|_2^2/\|\boldsymbol{b}\|_2^2$ to hold for all vectors $\boldsymbol{b}$ with certain sparsity properties.

3 The main result

Consider the high-dimensional linear model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}$ without noise from (1.1), where $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ is the design matrix and $\boldsymbol{\beta} \in \mathbb{R}^p$ is the sparsest possible parameter vector. As before, we denote the active set by $S = \{j : \beta_j \neq 0\}$ and its cardinality by $s = |S|$. According to the following theorem, L_2Boosting may fail to recover the vector $\boldsymbol{\beta}$ under the uniform restricted nullspace property $\mathrm{RN}_{\mathrm{unif}}(s, L)$ no matter how large L.

Theorem 1. *For any $L > 0$, there exist*

(a) a design matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ for some n and p with $n < p$ which fulfills the uniform restricted nullspace property $\mathrm{RN}_{\mathrm{unif}}(s, L)$ and

(b) a vector $\boldsymbol{\beta} \in \mathbb{R}^p$ with $|S| = s$ which is the unique sparsest solution of the equation $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{b}$

such that

$$\|\boldsymbol{\beta}^{[k]} - \boldsymbol{\beta}\|_1 \not\to 0 \quad \text{as } k \to \infty.$$

Hence, the S-sparse vector $\boldsymbol{\beta}$ is not recovered by L_2Boosting.

The proof of Theorem 1 is given in Section 3.2. There, we construct a design matrix $\boldsymbol{X}$ and a parameter vector $\boldsymbol{\beta}$ for each $L > 0$ which satisfy conditions (a) and (b) and for which the following holds: For any $k \geq 0$, the parameter vector $\boldsymbol{\beta}^{[k]} = \boldsymbol{\beta}_S^{[k]} + \boldsymbol{\beta}_{S^c}^{[k]}$ produced by the boosting algorithm in the k-th iteration step is such that $\boldsymbol{\beta}_S^{[k]} = \mathbf{0}$. Hence, L_2Boosting never selects an index j in the active set S. This implies that $\|\boldsymbol{\beta}^{[k]} - \boldsymbol{\beta}\|_1 \geq \|\boldsymbol{\beta}_S\|_1$ for any k, which in turn yields that $\|\boldsymbol{\beta}^{[k]} - \boldsymbol{\beta}\|_1 \not\to 0$ as $k \to \infty$.

3.1 Discussion of Theorem 1

In what follows, we compare the behaviour of L_2Boosting specified in Theorem 1 with that of ℓ_1-penalized methods such as basis pursuit and the Lasso. Similar points apply to the Dantzig selector of Candès & Tao [5]. For brevity, we however restrict attention to basis pursuit and the Lasso. Basis pursuit approximates the sparse parameter vector $\boldsymbol{\beta}$ in model (1.1) by any solution $\boldsymbol{\beta}^{\mathrm{BP}}$ of the minimization problem

$$\underset{\boldsymbol{b} \in \mathbb{R}^p}{\text{minimize}} \|\boldsymbol{b}\|_1 \quad \text{subject to } \boldsymbol{Y} = \boldsymbol{X}\boldsymbol{b}, \tag{3.1}$$

whereas the Lasso (in its Lagrangian form) is defined as any solution $\boldsymbol{\beta}_\lambda^{\mathrm{Lasso}}$ of the problem

$$\underset{\boldsymbol{b} \in \mathbb{R}^p}{\text{minimize}} \left\{ \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}\|_2^2 + \lambda \|\boldsymbol{b}\|_1 \right\} \tag{3.2}$$

with $\lambda > 0$ denoting the penalty constant.

In contrast to L_2Boosting, basis pursuit and the Lasso are guaranteed to recover the S-sparse vector $\boldsymbol{\beta}$ in the high-dimensional linear model (1.1) under an appropriate restricted nullspace property. More precisely, if the design matrix $\boldsymbol{X}$ fulfills $\mathrm{RN}(S, L)$ with $L \geq 1$, then $\boldsymbol{\beta}$ is the unique solution of the minimization problem (3.1), that is, $\boldsymbol{\beta} = \boldsymbol{\beta}^{\mathrm{BP}}$. Moreover, under $\mathrm{RN}(S, L)$ with $L \geq 1$, it holds that $\boldsymbol{\beta} = \lim_{\lambda \to 0} \boldsymbol{\beta}_\lambda^{\mathrm{Lasso}}$. Hence, $\boldsymbol{\beta}$ is recovered as the limit of the Lasso estimator $\boldsymbol{\beta}_\lambda^{\mathrm{Lasso}}$, where the penalty constant λ converges to zero. Letting the penalty constant λ of the Lasso estimator converge to zero corresponds to letting the number of boosting iterations k go to infinity.

The main reason why basis pursuit and the Lasso are ensured to recover the vector $\boldsymbol{\beta}$ under the nullspace contraint $\mathrm{RN}(S, L)$ with $L \geq 1$ is the following: The residuals $\Delta^{\mathrm{BS}} = \boldsymbol{\beta}^{\mathrm{BS}} - \boldsymbol{\beta}$ and $\Delta_\lambda^{\mathrm{Lasso}} = \boldsymbol{\beta}_\lambda^{\mathrm{Lasso}} - \boldsymbol{\beta}$ are guaranteed to lie in the cone $\mathbb{C}(S, 1)$, that is,

$$\|\Delta_{S^c}^{\mathrm{BS}}\|_1 \leq \|\Delta_S^{\mathrm{BS}}\|_1 \quad \text{and} \quad \|\Delta_{\lambda, S^c}^{\mathrm{Lasso}}\|_1 \leq \|\Delta_{\lambda, S}^{\mathrm{Lasso}}\|_1 \quad \text{for any } \lambda > 0.$$

L_2Boosting, in contrast, does not have this property. In particular, we can show the following result on the boosting residuals $\Delta^{[k]} = \boldsymbol{\beta}^{[k]} - \boldsymbol{\beta}$.

Corollary 1. *For any $L > 0$, there exist a design matrix $\boldsymbol{X}$ and a parameter vector $\boldsymbol{\beta}$ with the properties (a) and (b) from Theorem 1 such that*

$$\Delta^{[k]} \notin \mathbb{C}(S, L) \quad \textit{for sufficiently large } k.$$

The proof of Corollary 1 is provided in Section 3.3.

3.2 Proof of Theorem 1

For any given constant $L > 0$, we consider the following design: We let $p = n + 1$ and choose n to be a natural number whose square root is a natural number itself, that is, $n = N^2$ for some $N \in \mathbb{N}$. We pick n sufficiently large, in particular so large that

$$n \geq 5 \quad \text{and} \quad \frac{n + 1 - \sqrt{n}}{\sqrt{n}} > L.$$

The design matrix is given by

$$\boldsymbol{X} = \left(\begin{array}{ccc|c} \gamma & & & \gamma \\ & \ddots & & \vdots \\ & & \gamma & \gamma \\ \hline & & & \\ & & & \\ & & & \end{array}\right) \begin{array}{l} \left.\vphantom{\begin{array}{c}a\\a\\a\end{array}}\right\} s \text{ times} \\ \left.\begin{array}{ccc|c} 1 & & & 1 \\ & \ddots & & \vdots \\ & & 1 & 1 \end{array}\right\} n - s \text{ times} \end{array}$$

with $s = \sqrt{n}$ and $\gamma = n$, where empty positions in the matrix correspond to the entry 0.[1] The parameter vector is chosen as

$$\boldsymbol{\beta} = (\underbrace{1, \ldots, 1}_{s \text{ times}}, \underbrace{0, \ldots, 0}_{p - s \text{ times}})^\top \in \mathbb{R}^p.$$

Hence, the active set is $S = \{1, \ldots, s\}$ and the observation vector $\boldsymbol{Y}$ is given by

$$\boldsymbol{Y} = (\underbrace{\gamma, \ldots, \gamma}_{s \text{ times}}, \underbrace{0, \ldots, 0}_{n - s \text{ times}})^\top \in \mathbb{R}^n.$$

We first show that conditions (a) and (b) are fulfilled for our choices of $\boldsymbol{X}$ and $\boldsymbol{\beta}$: The nullspace $\mathcal{N}(\boldsymbol{X})$ of $\boldsymbol{X}$ is one-dimensional and spanned by the vector $\boldsymbol{z} = (-1, \ldots, -1, 1) \in \mathbb{R}^p$. For any subset $\mathcal{T} \subseteq \{1, \ldots, p\}$ with $|\mathcal{T}| \leq s$, it holds that

[1] Note that other choices of s and γ are possible.

$\|\boldsymbol{z}_{\mathcal{T}}\|_1 \leq s = \sqrt{n}$ and $\|\boldsymbol{z}_{\mathcal{T}^c}\|_1 \geq p - s = n + 1 - \sqrt{n}$, which implies that

$$\frac{\|\boldsymbol{z}_{\mathcal{T}}\|_1}{\|\boldsymbol{z}_{\mathcal{T}^c}\|_1} \leq \frac{\sqrt{n}}{n + 1 - \sqrt{n}} \to 0 \quad \text{as } n \to \infty.$$

From this, it immediately follows that the design matrix $\boldsymbol{X}$ satisfies the uniform restricted nullspace property $\mathrm{RN}_{\mathrm{unif}}(s, L)$ for any

$$L < \frac{n + 1 - \sqrt{n}}{\sqrt{n}}.$$

Hence, condition (a) is fulfilled. In order to see that condition (b) is satisfied as well, we make use of the following fact which is a consequence of results due to Donoho & Elad [8] and Gribonval & Nielsen [15]: The vector $\boldsymbol{\beta}$ is the unique sparsest solution of the equation $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{b}$ if $s < \mathrm{spark}(\boldsymbol{X})/2$, where $\mathrm{spark}(\boldsymbol{X})$ is the least number of columns of $\boldsymbol{X}$ that form a linearly dependent set. Since $s = \sqrt{n}$ and $\mathrm{spark}(\boldsymbol{X}) = n$, the inequality $s < \mathrm{spark}(\boldsymbol{X})/2$ holds for any $n \geq 5$. Consequently, the parameter vector $\boldsymbol{\beta}$ is guaranteed to satisfy condition (b) for any $n \geq 5$.
We now prove that $\|\boldsymbol{\beta}^{[k]} - \boldsymbol{\beta}\|_1 \not\to 0$ as $k \to \infty$. To do so, we verify the following lemma.

Lemma 1. *Suppose that the vector $\boldsymbol{\beta}^{[k]}$ obtained in the k-th iteration step of the boosting algorithm has the form*

$$\boldsymbol{\beta}^{[k]} = (\underbrace{0, \ldots, 0}_{s \text{ times}}, -c_{s+1}, \ldots, -c_n, c_p)^\top \text{ with } c_j \in [0, 1] \text{ for } s + 1 \leq j \leq p. \tag{3.3}$$

Then, in the $(k+1)$-th iteration step, the boosting algorithm produces a vector of the form

$$\boldsymbol{\beta}^{[k+1]} = (\underbrace{0, \ldots, 0}_{s \text{ times}}, -\tilde{c}_{s+1}, \ldots, -\tilde{c}_n, \tilde{c}_p)^\top \text{ with } \tilde{c}_j \in [0, 1] \text{ for } s + 1 \leq j \leq p. \tag{3.4}$$

Since the vector $\boldsymbol{\beta}^{[0]} = \boldsymbol{0} \in \mathbb{R}^p$ obviously has the form (3.3), Lemma 1 and a simple induction argument yield that for any $k \geq 0$, the vector $\boldsymbol{\beta}^{[k]}$ produced by the boosting algorithm is of the form $\boldsymbol{\beta}^{[k]} = (0, \ldots, 0, -c_{s+1}^{[k]}, \ldots, -c_n^{[k]}, c_p^{[k]})^\top$ with $c_j^{[k]} \in [0, 1]$ for $s + 1 \leq j \leq p$. From this, it immediately follows that $\|\boldsymbol{\beta}^{[k]} - \boldsymbol{\beta}\|_1 \geq \|\boldsymbol{\beta}_S\|_1 = s$ for any $k \geq 0$, which in turn implies that $\|\boldsymbol{\beta}^{[k]} - \boldsymbol{\beta}\|_1 \not\to 0$ as $k \to \infty$. To complete the proof of Theorem 1, it remains to verify Lemma 1.

Proof of Lemma 1. Since

$$\|\boldsymbol{X}_j\|_2 = \begin{cases} \gamma & \text{for } 1 \le j \le s \\ 1 & \text{for } s+1 \le j \le n \\ \sqrt{(\gamma^2-1)s+n} & \text{for } j = p \end{cases}$$

and

$$\begin{aligned} \boldsymbol{R}^{[k]} &= \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}^{[k]} = \boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{\beta}^{[k]}) \\ &= (\underbrace{\gamma(1-c_p), \ldots, \gamma(1-c_p)}_{s \text{ times}}, c_{s+1} - c_p, \ldots, c_n - c_p)^\top, \end{aligned}$$

we get that

$$\rho_j^{[k]} := \frac{\boldsymbol{X}_j^\top \boldsymbol{R}^{[k]}}{\|\boldsymbol{X}_j\|_2} = \begin{cases} \gamma(1-c_p) & \text{for } 1 \le j \le s \\ c_j - c_p & \text{for } s+1 \le j \le n \\ \dfrac{s\gamma^2(1-c_p) + \sum_{j=s+1}^n (c_j - c_p)}{\sqrt{(\gamma^2-1)s+n}} & \text{for } j = p. \end{cases}$$

The boosting algorithm picks the index $j_{k+1} = \arg\max_{1 \le j \le p} |\rho_j^{[k]}|$ in the $(k+1)$-th iteration step. (In case of ties, j_{k+1} is the smallest index j with $|\rho_j^{[k]}| \ge |\rho_i^{[k]}|$ for all i.) If $\max_{1 \le j \le s} |\rho_j^{[k]}| \ge \max_{s+1 \le j \le n} |\rho_j^{[k]}|$, that is, if $\gamma(1-c_p) \ge \max_{s+1 \le j \le n} |c_j - c_p|$, then

$$\begin{aligned} \rho_p^{[k]} &\ge \frac{s\gamma^2(1-c_p) - (n-s)\max_{s+1 \le j \le n} |c_j - c_p|}{\sqrt{(\gamma^2-1)s+n}} \\ &\ge \gamma(1-c_p) \underbrace{\frac{s\gamma - n + s}{\sqrt{(\gamma^2-1)s+n}}}_{>1 \text{ for all } n \ge 4}, \end{aligned}$$

which implies that $|\rho_p^{[k]}| > \max_{1 \le j \le s} |\rho_j^{[k]}|$ for any $n \ge 4$. Hence, only two cases are possible:

(A) $j_{k+1} = p$, or put differently, $|\rho_p^{[k]}| > |\rho_j^{[k]}|$ for all $j \ne p$.

(B) $j_{k+1} \in \{s+1, \ldots, n\}$, or put differently, there exists $j^* \in \{s+1, \ldots, n\}$ such that $|\rho_{j^*}^{[k]}| > |\rho_j^{[k]}|$ for all $1 \le j \le s$ and $|\rho_{j^*}^{[k]}| \ge |\rho_j^{[k]}|$ for all $s+1 \le j \le p$.

In case (A), $\boldsymbol{\beta}^{[k+1]} = (0, \ldots, 0, -\tilde{c}_{s+1}, \ldots, -\tilde{c}_n, \tilde{c}_p)^\top$, where $\tilde{c}_j = c_j$ for $s+1 \le j \le n$

and $\tilde{c}_p = c_p + \nu\Delta$ with

$$\Delta = \frac{s\gamma^2(1-c_p) + \sum_{j=s+1}^{n}(c_j - c_p)}{(\gamma^2-1)s + n}.$$

The parameter $\tilde{c}_p = c_p + \nu\Delta$ has the following properties:

(i) Since

$$\Delta \leq \frac{s\gamma^2(1-c_p) + (n-s)}{(\gamma^2-1)s+n} = 1 - c_p\frac{s\gamma^2}{s\gamma^2+n-s} \leq 1 - c_p,$$

it holds that $\tilde{c}_p \leq c_p + \nu(1-c_p) \leq 1$.

(ii) If $\gamma(1-c_p) \geq \max_{s+1\leq j\leq n}|c_j - c_p|$, then

$$\Delta \geq \gamma(1-c_p)\frac{s\gamma - n + s}{(\gamma^2-1)s+n} = \gamma(1-c_p)\frac{n^{3/2} - n + n^{1/2}}{n^{5/2}+n-n^{1/2}} \geq 0,$$

which implies that $\tilde{c}_p \geq c_p \geq 0$.

(iii) If $\gamma(1-c_p) < \max_{s+1\leq j\leq n}|c_j - c_p|$, then $1-c_p < \gamma^{-1} = n^{-1}$ and thus $c_p > 1-n^{-1}$. Noticing that $\Delta \geq -(n-s)/([\gamma^2-1]s+n)$, we obtain that for any $n \geq 2$,

$$\tilde{c}_p \geq 1 - \frac{1}{n} - \nu\underbrace{\frac{n-s}{(\gamma^2-1)s+n}}_{\leq 1/2 \text{ for any } n\geq 2} \geq 0.$$

Taken together, (i)–(iii) imply that $\tilde{c}_p \in [0,1]$. Hence, in case (A), $\boldsymbol{\beta}^{[k+1]}$ has the form (3.4) with $\tilde{c}_j \in [0,1]$ for all $s+1 \leq j \leq p$. We now turn to case (B). Assuming without loss of generality that $j_{k+1} = s+1$, we obtain that $\boldsymbol{\beta}^{[k+1]} = (0,\ldots,0,-\tilde{c}_{s+1},\ldots,-\tilde{c}_n,\tilde{c}_p)^\top$ with $\tilde{c}_{s+1} = (1-\nu)c_{s+1} + \nu c_p$ and $\tilde{c}_j = c_j$ for $s+2 \leq j \leq p$. Since $c_{s+1} \in [0,1]$ and $c_p \in [0,1]$ by assumption, it immediately follows that $\tilde{c}_{s+1} \in [0,1]$. Hence, in case (B), $\boldsymbol{\beta}^{[k+1]}$ has the desired form (3.4) as well with parameters $\tilde{c}_j \in [0,1]$ for all $s+1 \leq j \leq p$. □

3.3 Proof of Corollary 1

Let $\boldsymbol{X}$ and $\boldsymbol{\beta}$ be defined as in the proof of Theorem 1. We make use of the following two facts:

(i) According to the proof of Theorem 1, for any $k \geq 0$, the vector $\boldsymbol{\beta}^{[k]}$ has the form $\boldsymbol{\beta}^{[k]} = (0,\ldots,0,-c_{s+1}^{[k]},\ldots,-c_n^{[k]},c_p^{[k]})^\top$ with $c_j^{[k]} \in [0,1]$ for $s+1 \leq j \leq p$, which implies that $\Delta^{[k]} = \boldsymbol{\beta}^{[k]} - \boldsymbol{\beta} = (-1,\ldots,-1,-c_{s+1}^{[k]},\ldots,-c_n^{[k]},c_p^{[k]})^\top$.

(ii) Since $\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}^{[k]}\|_2 = \|\boldsymbol{X}\Delta^{[k]}\|_2 \to 0$ as $k \to \infty$ by Theorem 5.1 in Temlyakov [21], the vector $\Delta^{[k]}$ must converge to an element of the nullspace $\mathcal{N}(\boldsymbol{X})$ as $k \to \infty$, in particular to the vector $\Delta^* = (-1, \ldots, -1, 1)^\top \in \mathbb{R}^p$.

Taken together, (i) and (ii) yield that $\|\Delta_S^{[k]}\|_1 = s = \sqrt{n}$ for any $k \geq 0$ and $\|\Delta_{S^c}^{[k]}\|_1 \geq (p-s)/2 = (n+1-\sqrt{n})/2$ for k large enough. Consequently,

$$\frac{n+1-\sqrt{n}}{2\sqrt{n}}\|\Delta_S^{[k]}\|_1 \leq \|\Delta_{S^c}^{[k]}\|_1$$

for sufficiently large k, or put differently, $\Delta^{[k]} \notin \mathbb{C}(S, L)$ for any $L < (n+1-\sqrt{n})/(2\sqrt{n})$ and sufficiently large k. From this, the statement of Corollary 1 easily follows.

References

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

[2] Bühlmann, P. (2006). Boosting for high-dimensional linear models. *Ann. Statist.*, **34** 559–583.

[3] Bühlmann, P. and Yu, B. (2006). Sparse boosting. *Journal of Machine Learning Research*, **7** 1001–1024.

[4] Candès, E. and Tao, T. (2005). Decoding by linear programming. *IEEE Trans. Info Theory*, **51** 4203–4215.

[5] Candès, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, **35** 2313–2351.

[6] Chen, S., Donoho, D. and Saunders, M. (1998). Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, **20** 33–61.

[7] Cohen, A., Dahmen, W. and DeVore, R. A. (2009). Compressed sensing and best k-term approximation. *Journal of the American Mathematical Society*, **22** 211–231.

[8] Donoho, D. and Elad, M. (2003). Maximal sparsity representation via ℓ_1 minimization. *Proc. Natl. Acad. Sci.*, **100** 2197–2202.

[9] Donoho, D., Elad, M. and Temlyakov, V. (2006). Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. Info Theory*, **52** 6–18.

[10] Donoho, D. and Huo, X. (2001). Uncertainty principles and ideal atomic decomposition. *IEEE Trans. Info Theory*, **47** 2845–2862.

[11] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Ann. Statist.*, **32** 407–499.

[12] Feuer, A. and Nemirovski, A. (2003). On sparse representation in pairs of bases. *IEEE Trans. Info Theory*, **49** 1579–1581.

[13] Freund, Y. and Schapire, R. (1996). Experiments with a new boosting algorithm. In *Proceedings of the Thirteenth International Conference on Machine Learning*. Morgan Kaufmann, San Francisco, 148–156.

[14] Friedman, J. (2001). Greedy function approximation: a gradient boosting machine. *Ann. Statist.*, **29** 1189–1232.

[15] Gribonval, R. and Nielsen, M. (2003). Sparse representations in unions of bases. *IEEE Trans. Inform. Theory*, **49** 3320–3325.

[16] Hastie, T., Taylor, J., Tibshirani, R. and Walther, G. (2007). Forward stagewise regression and the monotone lasso. *Electronic Journal of Statistics*, **1** 1–29.

[17] Hepp, T., Schmid, M., Gefeller, O., Waldmann, E. and Mayr, A. (2016). Approaches to regularized regression – a comparison between Gradient Boosting and the Lasso. *Methods Inf Med*, **55** 422–430.

[18] Mallat, S. and Zhang, Z. (1993). Matching pursuits with time-frequency dictionaries. *IEEE Trans. Signal Process.*, **41** 3397–3415.

[19] Qian, S. and Chen, D. (1994). Signal representation using adaptive normalized Gaussian functions. *Signal Process.*, **36** 1–11.

[20] Schapire, R. (1990). The strength of weak learnability. *Machine Learning*, **5** 197–227.

[21] Temlyakov, V. (2000). Weak greedy algorithms. *Adv. Comput. Math.*, **12** 213–227.

[22] Temlyakov, V. (2011). *Greedy approximation*, vol. 20 of *Cambridge Monographs on Applied and Computational Mathematics*. Cambridge University Press.

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

[24] Tropp, J. (2004). Greed is good: algorithmic results for sparse approximation. *IEEE Trans. Info Theory*, **50** 2231–2242.