Learning with Structured Sparsity

Junzhou Huang
Department of Computer Science, Rutgers University

Tong Zhang
Department of Statistics, Rutgers University

Dimitris Metaxas
Department of Computer Science, Rutgers University

Abstract

This paper investigates a new learning formulation called ***structured sparsity***, which is a natural extension of the standard sparsity concept in statistical learning and compressive sensing. By allowing arbitrary structures on the feature set, this concept generalizes the group sparsity idea that has become popular in recent years. A general theory is developed for learning with structured sparsity, based on the notion of coding complexity associated with the structure. It is shown that if the coding complexity of the target signal is small, then one can achieve improved performance by using coding complexity regularization methods, which generalize the standard sparse regularization. Moreover, a structured greedy algorithm is proposed to efficiently solve the structured sparsity problem. It is shown that the greedy algorithm approximately solves the coding complexity optimization problem under appropriate conditions. Experiments are included to demonstrate the advantage of structured sparsity over standard sparsity on some real applications.

1 Introduction

We are interested in the sparse learning problem under the fixed design condition. Consider a fixed set of p basis vectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_p\}$ where $\mathbf{x}_j \in \mathbb{R}^n$ for each j. Here, n is the sample size. Denote by X the $n \times p$ data matrix, with column j of X being $\mathbf{x}_j$. Given a random observation $\mathbf{y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n] \in \mathbb{R}^n$ that depends on an underlying coefficient vector $\bar{\beta} \in \mathbb{R}^p$, we are interested in the problem of estimating $\bar{\beta}$ under the assumption that the target coefficient $\bar{\beta}$ is sparse. Throughout the paper, we consider fixed design only. That is, we assume X is fixed, and randomization is with respect to the noise in the observation $\mathbf{y}$.

We consider the situation that $\mathbb{E}\mathbf{y}$ can be approximated by a sparse linear combination of the basis vectors:

$$\mathbb{E}\mathbf{y} \approx X\bar{\beta},$$

where we assume that $\bar{\beta}$ is sparse. Define the support of a vector $\beta \in \mathbb{R}^p$ as

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

and $\|\beta\|_0 = |\mathrm{supp}(\beta)|$. A natural method for sparse learning is L_0 regularization:

$$\hat{\beta}_{L0} = \arg\min_{\beta \in \mathbb{R}^p} \hat{Q}(\beta) \quad \text{subject to } \|\beta\|_0 \leq s,$$

where s is the desired sparsity. For simplicity, unless otherwise stated, the objective function considered throughout this paper is the least squares loss

$$\hat{Q}(\beta) = \|X\beta - \mathbf{y}\|_2^2,$$

although other objective functions for generalized linear models (such as logistic regression) can be similarly analyzed.

Since this optimization problem is generally NP-hard, in practice, one often considers approximate solutions. A standard approach is convex relaxation of L_0 regularization to L_1 regularization, often referred to as Lasso [22]. Another commonly used approach is greedy algorithms, such as the orthogonal matching pursuit (OMP) [23].

In practical applications, one often knows a structure on the coefficient vector $\bar{\beta}$ in addition to sparsity. For example, in group sparsity, one assumes that variables in the same group tend to be zero or nonzero simultaneously. The purpose of this paper is to study the more general estimation problem under structured sparsity. If meaningful structures exist, we show that one can take advantage of such structures to improve the standard sparse learning.

2 Related Work

The idea of using structure in addition to sparsity has been explored before. An example is group structure, which has received much attention recently. For example, group sparsity has been considered for simultaneous sparse approximation [24] and multi-task compressive sensing [14] from the Bayesian hierarchical modeling point of view. Under the Bayesian hierarchical model framework, data from all sources contribute to the estimation of hyper-parameters in the sparse prior model. The shared prior can then be inferred from multiple sources. He et al. recently extend the idea to the tree sparsity in the Bayesian framework [11, 12]. Although the idea can be justified using standard Bayesian intuition, there are no theoretical results showing how much better (and under what kind of conditions) the resulting algorithms perform. In the statistical literature, Lasso has been extended to the group Lasso when there exist group/block structured dependences among the sparse coefficients [25].

However, none of the above mentioned work was able to show advantage of using group structure. Although some theoretical results were developed in [1, 18], neither showed that group Lasso is superior to the standard Lasso. The authors of [15] showed that group Lasso can be superior to standard Lasso when each group is an infinite dimensional kernel, by relying on the fact that meaningful analysis can be obtained for kernel methods in infinite dimension. In [19], the authors consider a special case of group Lasso in the multi-task learning scenario, and show that the number of samples required for recovering the exact support set is smaller for group Lasso under appropriate conditions. In [13], a theory for group Lasso was developed using a concept called strong group sparsity, which is a special case of the general structured sparsity idea considered here. It was shown in [13] that group Lasso is superior to standard Lasso for strongly group-sparse signals, which provides a convincing theoretical justification for using group structured sparsity.

While group Lasso works under the strong group sparsity assumption, it doesn't handle the more general structures considered in this paper. Several limitations of group Lasso were mentioned in [13]. For example, group Lasso does not correctly handle overlapping groups (in that overlapping components are over-counted); that is, a given coefficient should not belong to different groups. This requirement is too rigid for many practical applications. To address this issue, a method called composite absolute penalty (CAP) is proposed in [27] which can handle overlapping groups. Unfortunately, no theory is established to demonstrate the effectiveness of the approach. In a related development [16], Kowalski et al. generalized the mixed norm penalty to structured shrinkage, which can identify the structured significance maps and thus can handle the case of the overlapping groups, However, the structured shrinkage operations do not necessarily convergence to a fixed point. There were no additional theory to justify their methods.

Other structures have also been explored in the literature. For example, so-called tonal and transient structures were considered for sparse decomposition of audio signals in [8], but again without any theory. Grimm et al. [10] investigated positive polynomials with structured sparsity from an optimization perspective. The theoretical result there did not address the effectiveness of such methods in comparison to standard sparsity. The closest work to ours is a recent paper [2], which we learned after finishing this paper. In that paper, a specific case of structured sparsity, referred to as model based sparsity, was considered. It is important to note that some theoretical results were obtained there to show the effectiveness of their method in compressive sensing. However, they do not provide a generic framework for structured sparsity. Thus, different schemes have to be specifically designed for different data models. It remains as an open issue how to develop a general theory for structured sparsity, together with a general algorithm that can be applied to a wide class of such problems.

We see from the above discussion that there exists extensive literature on structured sparsity, with empirical evidence showing that one can achieve better performance by imposing additional structures. However, none of the previous work was able to establish a general theoretical framework for structured sparsity that can quantify its effectiveness. The goal of this paper is to develop such a general theory that addresses the following issues, where we pay special attention to the benefit of structured sparsity over the standard non-structured sparsity:

- quantifying structured sparsity;
- the minimal number of measurements required in compressive sensing;
- estimation accuracy under stochastic noise;
- an efficient algorithm that can solve a wide class of structured sparsity problems.

3 Structured Sparsity

In structured sparsity, not all sparse patterns are equally likely. For example, in group sparsity, coefficients within the same group are more likely to be zeros or nonzeros simultaneously. This means that if a sparse coefficient vector's support set is consistent with the underlying group structure, then it is more likely to occur, and hence incurs a smaller penalty in learning. One contribution of this work is to formulate how to define structure on top of sparsity, and how to penalize each sparsity pattern.

In order to formalize the idea, we denote by $\mathcal{I} = \{1, \ldots, p\}$ the index set of the coefficients. Consider any sparse subset $F \subset \{1, \ldots, p\}$, we assign a cost $\mathrm{cl}(F)$. In structured sparsity, the cost of F is an upper bound of the coding length of F (number of bits needed to represent F by a computer program) in a pre-chosen prefix coding scheme. It is a well-known fact in information theory (e.g. [7]) that mathematically, the existence of such a coding scheme is equivalent to

$$\sum_{F \subset \mathcal{I}} 2^{-\mathrm{cl}(F)} \leq 1.$$

From the Bayesian statistics point of view, $2^{-\mathrm{cl}(F)}$ can be regarded as a lower bound of the probability of F. The probability model of structured sparse learning is thus: first generate the sparsity pattern F according to probability $2^{-\mathrm{cl}(F)}$; then generate the coefficients in F.

Definition 3.1 *A cost function* $\mathrm{cl}(F)$ *defined on subsets of* $\mathcal{I}$ *is called a coding length (in base-2) if*

$$\sum_{F \subset \mathcal{I}, F \neq \emptyset} 2^{-\mathrm{cl}(F)} \leq 1.$$

We give $\emptyset$ *a coding length 0. The corresponding structured sparse coding complexity of* F *is defined as*

$$c(F) = |F| + \mathrm{cl}(F).$$

A coding length $\mathrm{cl}(F)$ *is sub-additive if*

$$\mathrm{cl}(F \cup F') \leq \mathrm{cl}(F) + \mathrm{cl}(F'),$$

and a coding complexity $c(F)$ *is sub-additive if*

$$c(F \cup F') \leq c(F) + c(F').$$

Clearly if $\mathrm{cl}(F)$ is sub-additive, then the corresponding coding complexity $c(F)$ is also sub-additive. Based on the structured coding complexity of subsets of $\mathcal{I}$, we can now define the structured coding complexity of a sparse coefficient vector $\bar{\beta} \in \mathbb{R}^p$.

Definition 3.2 *Giving a coding complexity* $c(F)$*, the structured sparse coding complexity of a coefficient vector* $\bar{\beta} \in \mathbb{R}^p$ *is*

$$c(\bar{\beta}) = \min\{c(F) : \mathrm{supp}(\bar{\beta}) \subset F\}.$$

Later in the paper, we will show that if a coefficient vector $\bar{\beta}$ has a small coding complexity $c(\bar{\beta})$, then $\bar{\beta}$ can be effectively learned, with good in-sample prediction performance (in statistical learning) and reconstruction performance (in compressive sensing). In order to see why the definition requires adding $|F|$ to $\mathrm{cl}(F)$, we consider the generative model for structured sparsity mentioned earlier. In this model, the number of bits to encode a sparse coefficient vector is the sum of the number of bits to encode F (which is $\mathrm{cl}(F)$) and the number of bits to encode nonzero coefficients in F (this requires $O(|F|)$ bits up to a fixed precision). Therefore the total number of bits required is $\mathrm{cl}(F) + O(|F|)$. This information theoretical result translates into a statistical estimation result: without additional regularization, the learning complexity for least squares regression within any fixed support set F is $O(|F|)$. By adding the model selection complexity $\mathrm{cl}(F)$ for each support set F, we obtain an overall statistical estimation complexity of $O(\mathrm{cl}(F) + |F|)$.

While the idea of using coding based penalization is clearly motivated by the minimum description length (MDL) principle, the actual penalty we obtain for structured sparsity problems is different from the standard MDL penalty for model selection. This difference is important in sparse learning. Therefore in order to prevent confusion, we avoid using MDL in our terminology. Nevertheless, one may consider our framework as a natural combination of the MDL idea and the modern sparsity analysis.

4 Structured Sparsity Examples

Before giving detailed examples, we introduce a general coding scheme called *block coding*. The basic idea of block coding is to define a coding scheme on a small number of base blocks (a block is a subset of $\mathcal{I}$), and then define a coding scheme on all subsets of $\mathcal{I}$ using these base blocks.

Consider a subset $\mathcal{B} \subset 2^{\mathcal{I}}$. That is, each element (a block) of $\mathcal{B}$ is a subset of $\mathcal{I}$. We call $\mathcal{B}$ a block set if $\mathcal{I} = \cup_{B \in \mathcal{B}} B$ and all single element sets $\{j\}$ belong to $\mathcal{B}$ ($j \in \mathcal{I}$). Note that $\mathcal{B}$ may contain additional non single-element blocks. The requirement of $\mathcal{B}$ containing all single element sets is for convenience, as it implies that every subset $F \subset \mathcal{I}$ can be expressed as the union of blocks in $\mathcal{B}$.

Let cl_0 be a code length on $\mathcal{B}$:

$$\sum_{B \in \mathcal{B}} 2^{-\mathrm{cl}_0(B)} \leq 1,$$

we define $\mathrm{cl}(B) = \mathrm{cl}_0(B) + 1$ for $B \in \mathcal{B}$. It not difficult to show that the following cost function on $F \subset \mathcal{I}$ is a coding length

$$\mathrm{cl}(F) = \min \left\{ \sum_{j=1}^{b} \mathrm{cl}(B_j) : F = \bigcup_{j=1}^{b} B_j \quad (B_j \in \mathcal{B}) \right\}.$$

This is because

$$\sum_{F \subset \mathcal{I}, F \neq \emptyset} 2^{-\mathrm{cl}(F)} \leq \sum_{b \geq 1} \sum_{\{B_\ell\} \in \mathcal{B}^b} 2^{-\sum_{\ell=1}^{b} \mathrm{cl}(B_\ell)} \leq \sum_{b \geq 1} \prod_{\ell=1}^{b} \sum_{B_\ell \in \mathcal{B}} 2^{-\mathrm{cl}(B_\ell)} \leq \sum_{b \geq 1} 2^{-b} = 1.$$

It is clear from the definition that block coding is sub-additive.

The main purpose of introducing block coding is to design computationally efficient algorithms based on the block structure. In particular, this paper considers a structured greedy algorithm that can take advantage of block structures. In the structured greedy algorithm, instead of searching over all subsets of $\mathcal{I}$ up to a fixed coding complexity s (the number of such subsets can be exponential in s), we greedily add blocks from $\mathcal{B}$ one at a time. Each search problem over $\mathcal{B}$ can be efficiently performed because we require that $\mathcal{B}$ contains only a computationally manageable number of base blocks. Therefore the algorithm is computationally efficient.

We will show that under appropriate conditions, a target coefficient vector with a small block coding complexity can be approximately learned using the structured greedy algorithm. This means that the block coding scheme has important algorithmic implications. That is, if a coding scheme can be approximated by block coding with a small number of base blocks, then the corresponding estimation problem can be approximately solved using the structured greedy algorithm. For this reason, we shall pay special attention to block coding approximation schemes for examples discussed below.

Standard sparsity

A simple coding scheme is to code each subset $F \subset \mathcal{I}$ of cardinality k using $k \log_2(2p)$ bits, which corresponds to block coding with $\mathcal{B}$ consisted only of single element sets, and each base block has a coding length $\mathrm{cl}_0 = \log_2 p$. This corresponds to the complexity for the standard sparse learning.

A more general version is to consider single element blocks $\mathcal{B} = \{\{j\} : j \in \mathcal{I}\}$, with a non-uniform coding scheme $\mathrm{cl}_0(\{j\}) = c_j$, such that $\sum_j 2^{-c_j} \leq 1$. It leads to a non-uniform coding length on $\mathcal{I}$ as

$$\mathrm{cl}(B) = |B| + \sum_{j \in B} c_j.$$

In particular, if a feature j is likely to be nonzero, we should give it a smaller coding length c_j, and if a feature j is likely to be zero, we should give it a larger coding length.

Group sparsity

The concept of group sparsity has appeared in various recent work, such as the group Lasso in [25]. Consider a partition of $\mathcal{I} = \cup_{j=1}^{m} G_j$ into m disjoint groups. Let $\mathcal{B}_G$ contain the m groups $\{G_j\}$, and $\mathcal{B}_1$ contain p single element blocks. The strong group sparsity coding scheme is to give each element in $\mathcal{B}_1$ a code-length cl_0 of ∞, and each element in $\mathcal{B}_G$ a code-length cl_0 of $\log_2 m$. Then the block coding scheme with blocks $\mathcal{B} = \mathcal{B}_G \cup \mathcal{B}_1$ leads to group sparsity, which only looks for signals consisted of the groups. The resulting coding length is: $\mathrm{cl}(B) = g \log_2(2m)$ if B can be represented as the union of g disjoint groups G_j; and $\mathrm{cl}(B) = \infty$ otherwise.

Note that if the signal can be expressed as the union of g groups, and each group size is k_0, then the group coding length $g \log_2(2m)$ can be significantly smaller than the standard sparsity coding length of $gk_0 \log_2(p)$. As we shall see later, the smaller coding complexity implies better learning behavior, which is essentially the advantage of using group sparse structure. It was shown in [13] that strong group sparsity defined above also characterizes the performance group Lasso. Therefore if a signal has a pre-determined group structure, then group Lasso is superior to standard Lasso.

An extension of this idea is to allow more general block coding length for $\mathrm{cl}_0(G_j)$ and $\mathrm{cl}_0(\{j\})$ so that

$$\sum_{j=1}^{m} 2^{-\mathrm{cl}_0(G_j)} + \sum_{j=1}^{p} 2^{-\mathrm{cl}_0(\{j\})} \leq 1.$$

This leads to non-uniform coding of the groups, so that a group that is more likely to be nonzero is given a smaller coding length.

Hierarchical sparsity

One may also create a hierarchical group structure. A simple example is wavelet coefficients of a signal [17]. Another simple example is a binary tree with the variables as leaves, which we describe below. Each internal node in the tree is associated with three options: left child only, right child only, and both children; each option can be encoded in $\log_2 3$ bits.

Given a subset $F \subset \mathcal{I}$, we can go down from the root of the tree, and at each node, decide whether only left child contains elements of F, or only right child contains elements of F, or both children contain elements of F. Therefore the coding length of F is $\log_2 3$ times the total number of internal nodes leading to elements of F. Since each leaf corresponds to no more than $\log_2 p$ internal nodes, the total coding length is no worse than $\log_2 3 \log_2 p|F|$. However, the coding length can be significantly smaller if nodes are close to each other or are clustered. In the extreme case, when the nodes are consecutive, we have $O(|F| + \log_2 p)$ coding length. More generally, if we can order elements in F as $F = \{j_1, \ldots, j_q\}$, then the coding length can be bounded as $\mathrm{cl}(F) = O(|F| + \log_2 p + \sum_{s=2}^{q} \log_2 \min_{\ell < s} |j_s - j_\ell|)$.

If all internal nodes of the tree are also variables in $\mathcal{I}$ (for example, in the case of wavelet decomposition), then one may consider feature set F with the following property: if a node is selected, then its parent is also selected. This requirement is very effective in wavelet compression, and often referred to as the zero-tree structure [21]. Similar requirements have also been applied in statistics [27] for variable selection. The argument presented in this section shows that if we require F to satisfy the zero-tree structure, then its coding length is at most $O(|F|)$, without any explicit dependency on the dimensionality p. This is because one does not have to reach a leave node.

The tree-based coding scheme discussed in this section can be approximated by block coding using no more than $p^{1+\delta}$ base blocks ($\delta > 0$). The idea is similar to that of the image coding example in the more general graph sparsity scheme which we discuss next.

Graph sparsity

We consider a generalization of the hierarchical and group sparsity idea that employs a (directed or undirected) graph structure G on $\mathcal{I}$. To the best of our knowledge, this general structure has not been considered in any previous work.

In graph sparsity, each variable (an element of $\mathcal{I}$) is a node of G but G may also contain additional nodes that are not variables. For simplicity, we assume G contains a starting node (this requirement is not critical).

At each node $v \in G$, we define coding length $\mathrm{cl}_v(S)$ on all subsets S of the neighborhood N_v of v including the empty set, as well as any other single node $u \in G$ with $\mathrm{cl}_v(u)$, such that

$$\sum_{S \subset N_v} 2^{-\mathrm{cl}_v(S)} + \sum_{u \in G} 2^{-\mathrm{cl}_v(u)} \leq 1.$$

To encode $F \subset G$, we start with the active set containing only the starting node, and finish when the set becomes empty. At each node v before termination, we may either pick a subset $S \subset N_v$, with coding length $\mathrm{cl}_v(S)$, or a node in $u \in G$, with coding length $\mathrm{cl}_v(u)$, and then put the selection into the active set. We then remove v from the active set (once a node v is removed, it does not return to the active set anymore). This process is continued until the active set becomes empty.

As a concrete example, we consider image processing, where each image is a rectangle of pixels (nodes); each pixel is connected to four adjacent pixels, which forms the underlying graph structure. At each pixel, the number of subsets in its neighborhood is $2^4 = 16$ (including the empty set), and each subset is given a coding length $\mathrm{cl}_v(S) = 5$; we also encode all other pixels in the image with random jumping, each with a coding length $1 + \log_2 p$. Using this scheme, we can encode each connected region F by no more than $\log_2 p + 5|F|$ bits by growing the region from a single point in the region. Therefore if F is composed of g connected regions, then the coding length is $g \log_2 p + 5|F|$, which can be significantly better than standard sparse coding length of $|F| \log_2 p$.

This example shows that the general graph coding scheme presented here favors connected regions (that is, nodes that are grouped together with respect to the graph structure). In particular, it proves the following more general result.

Proposition 4.1 *Given a graph G, there exists a constant C_G such that for any probability distribution q on G ($\sum_{v \in G} q(v) = 1$ and $q(v) \geq 0$ for $v \in G$), the following quantity is a coding length on 2^G:*

$$\mathrm{cl}(F) = C_G|F| - \sum_{j=1}^{g} \max_{v \in F_j} \log_2 q(v),$$

where $F \subset 2^G$ can be decomposed into the union of g connected components $F = \cup_{j=1}^{g} F_j$.

Our simple and suboptimal coding scheme described for the image processing example gives an upper bound $C_G \leq 1 + d_G$, where d_G is the maximum degree of G. However, for many graphs, one can improve this constant to $O(\log d_G)$ with a slightly more complicated argument.

As a simple application of graph coding, we consider the special case where we have only one connected component that contains the starting node v_0. We can simply let $q(v_0) = 1$, and the coding length is $O(|F|)$, which is independent of the dimensionality p. This generalizes the similar claim for the zero-tree structure described earlier.

The graph coding scheme can be approximated with block coding. The idea is to consider relatively small sized base blocks consisted of nodes that are close together with respect to the graph structure, and then use the induced block coding scheme to approximate the graph coding.

For example, for the previously discussed image coding example, we can use connected blocks of size upto $\delta \log_2 p/5$ as base blocks of $\mathcal{B}$ ($\delta > 0$). Since each base block can be encoded with $(1+\delta)\log_2 p$ bits by earlier discussion, we know that the total number of base blocks can be no more than $p^{1+\delta}$. We can give each of such blocks a coding length $(1+\delta)\log_2 p$. For a connected region F that can be covered by $O(1 + |F|/\log_2 p)$ of such blocks, the corresponding block coding length for a subset F is $\mathrm{cl}(F) = O(|F| + \log_2 p)$, which is the same as the complexity of the original graph coding length (up to a constant). This means that graph coding length can be approximated with block coding scheme. As we have pointed out, such an approximation is useful because the latter is required in the structured greedy algorithm which we propose in this paper.

Random field sparsity

Let $z_j \in \{0, 1\}$ be a random variable for $j \in \mathcal{I}$ that indicates whether j is selected or not. The most general coding scheme is to consider a joint probability distribution of $z = [z_1, \ldots, z_p]$. The coding length for F can be defined as $-\log_2 p(z_1, \ldots, z_p)$ with $z_j = I(j \in F)$ indicating whether $j \in F$ or not.

Such a probability distribution can often be conveniently represented as a binary random field on an underlying graph. In order to encourage sparsity, on average, the marginal probability $p(z_j)$ should take 1 with probability close to $O(1/p)$, so that the expected number of j's with $z_j = 1$ is $O(1)$. For disconnected graphs (z_j are independent), the variables z_j are iid Bernoulli random variables with probability $1/p$ being one. In this case, the coding length of a set F is $|F|\log_2(p) - (p - |F|)\log_2(1 - 1/p) \approx |F|\log_2(p) + 1$. This is essentially the probability model for the standard sparsity scheme.

In many cases, it is possible to approximate a general random field coding scheme with block coding by using approximation methods in the graphical model literature. However, the details of such approximations are beyond the scope of this paper.

5 Algorithms for Structured Sparsity

The following algorithm is a natural extension of L_0 regularization to structured sparsity problems. It penalizes the coding complexity instead of the cardinality (sparsity) of the feature set.

$$\hat{\beta}_{constr} = \arg\min_{\beta \in \mathbb{R}^p} \hat{Q}(\beta) \quad \text{subject to } c(\beta) \leq s. \tag{1}$$

Alternatively, we may consider the formulation

$$\hat{\beta}_{pen} = \arg\min_{\beta \in \mathbb{R}^p} \left[\hat{Q}(\beta) + \lambda c(\beta) \right]. \tag{2}$$

The optimization of either (1) or (2) is generally hard. For related problems, there are two common approaches to alleviate this difficulty. One is convex relaxation (L_1 regularization to replace L_0 regularization for standard sparsity); the other is forward greedy selection (also called orthogonal matching pursuit or OMP). We do not know any extensions of L_1 regularization like convex relaxation that can handle general structured sparsity formulations. However, one can extend greedy algorithm by using a block structure. We call the resulting procedure structured greedy algorithm or StructOMP, which can approximately solve (1).

We have discussed the relationship of this greedy algorithm and block coding in Section 4. It is important to understand that the block structure is only used to limit the search space in the greedy algorithm. The actual coding scheme does not have to be the corresponding block coding. However, our theoretical analysis assumes that the underlying coding scheme can be approximated with block coding using base blocks employed in the greedy algorithm. Although one does not need to know the specific approximation in order to use the greedy algorithm, knowing its existence (which can be shown for the examples discussed in Section 4) guarantees the effectiveness of the algorithm. It is also useful to understand that our result does not imply that the algorithm won't be effective if the actual coding scheme cannot be approximated by block coding.

```
Input: $(X, \mathbf{y})$, $\mathcal{B} \subset 2^{\mathcal{I}}$, $s > 0$
Output: $F^{(k)}$ and $\beta^{(k)}$
let $F^{(0)} = \emptyset$ and $\beta^{(0)} = 0$
**for** $k = 1, 2, \ldots$
  select $B^{(k)} \in \mathcal{B}$ to maximize progress        $(*)$
  let $F^{(k)} = B^{(k)} \cup F^{(k-1)}$
  let $\beta^{(k)} = \arg\min_{\beta \in \mathbb{R}^p} \hat{Q}(\beta)$ subject to $\mathrm{supp}(\beta) \subset F^{(k)}$
  **if** $(c(\beta^{(k)}) > s)$ **break**
**end**
```

Figure 1: Structured Greedy Algorithm

In Figure 1, we are given a set of blocks $\mathcal{B}$ that contains subsets of $\mathcal{I}$. Instead of searching all subsets $F \subset \mathcal{I}$ up to a certain complexity $|F| + c(F)$, which is computationally infeasible, we search only the blocks restricted to $\mathcal{B}$. It is assumed that searching over $\mathcal{B}$ is computationally manageable.

At each step $(*)$, we try to find a block from $\mathcal{B}$ to maximize progress. It is thus necessary to define a quantity that measures progress. Our idea is to approximately maximize the gain ratio:

$$\lambda^{(k)} = \frac{\hat{Q}(\beta^{(k-1)}) - \hat{Q}(\beta^{(k)})}{c(\beta^{(k)}) - c(\beta^{k-1})},$$

which measures the reduction of objective function per unit increase of coding complexity. This greedy criterion is a natural generalization of the standard greedy algorithm, and essential in our analysis. For least squares regression, we can approximate $\lambda^{(k)}$ using the following definition

$$\phi(B) = \frac{\|P_{B - F^{(k-1)}}(X\beta^{(k-1)} - \mathbf{y})\|_2^2}{c(B \cup F^{(k-1)}) - c(F^{(k-1)})}, \tag{3}$$

where

$$P_F = X_F(X_F^\top X_F)^{-1}X_F^\top$$

is the projection matrix to the subspaces generated by columns of X_F. We then select $B^{(k)}$ so that

$$\phi(B^{(k)}) \geq \gamma \max_{B\in\mathcal{B}} \phi(B),$$

where $\gamma \in (0,1]$ is a fixed approximation ratio that specifies the quality of approximate optimization. Alternatively, we may use a simpler definition

$$\tilde{\phi}(B) = \frac{\|X_{B-F^{(k-1)}}^\top(X\beta^{(k-1)} - \mathbf{y})\|_2^2}{c(B\cup F^{(k-1)}) - c(F^{(k-1)})},$$

which is easier to compute, especially when blocks are overlapping. Since the ratio

$$\|X_{B-F^{(k-1)}}^\top \mathbf{r}\|_2^2 / \|P_{B-F^{(k-1)}}\mathbf{r}\|_2^2$$

is bounded between $\rho_+(B)$ and $\rho_-(B)$ (these quantities are defined in Definition 6.1), we know that maximizing $\tilde{\phi}(B)$ would lead to approximate maximization of $\phi(B)$ with $\gamma \geq \rho_-(B)/\rho_+(B)$.

Note that we shall ignore $B \in \mathcal{B}$ such that $B \subset F^{(k-1)}$, and just let the corresponding gain to be 0. Moreover, if there exists a base block $B \not\subset F^{(k-1)}$ but $c(B\cup F^{(k-1)}) \leq c(F^{(k-1)})$, we can always select B and let $F^{(k)} = B \cup F^{(k-1)}$ (this is because it is always beneficial to add more features into $F^{(k)}$ without additional coding complexity). We assume this step is always performed if such a $B \in \mathcal{B}$ exists. The non-trivial case is $c(B\cup F^{(k-1)}) > c(F^{(k-1)})$ for all $B \in \mathcal{B}$; in this case both $\phi(B)$ and $\tilde{\phi}(B)$ are well defined.

6 Theory of Structured Sparsity

6.1 Assumptions

We assume sub-Gaussian noise as follows.

Assumption 6.1 *Assume that $\{\mathbf{y}_i\}_{i=1,\ldots,n}$ are independent (but not necessarily identically distributed) sub-Gaussians: there exists a constant $\sigma \geq 0$ such that $\forall i$ and $\forall t \in R$,*

$$\mathbb{E}_{\mathbf{y}_i}\, e^{t(\mathbf{y}_i - \mathbb{E}\mathbf{y}_i)} \leq e^{\sigma^2 t^2/2}.$$

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

The following property of sub-Gaussian noise is important in our analysis. Our simple proof yields a sub-optimal choice of the constants.

Proposition 6.1 *Let $P \in \mathbb{R}^{n\times n}$ be a projection matrix of rank k, and $\mathbf{y}$ satisfies Assumption 6.1. Then for all $\eta \in (0,1)$, with probability larger than $1-\eta$:*

$$\|P(\mathbf{y} - \mathbb{E}\mathbf{y})\|_2^2 \leq \sigma^2[7.4k + 2.7\ln(2/\eta)].$$

We also need to generalize sparse eigenvalue condition, used in the modern sparsity analysis. It is related to (and weaker than) the RIP (restricted isometry property) assumption [6] in the compressive sensing literature. This definition takes advantage of coding complexity, and can be also considered as (a weaker version of) structured RIP. We introduce a definition.

Definition 6.1 *For all $F \subset \{1, \ldots, p\}$, define*

$$\begin{aligned}\rho_-(F) =& \inf \left\{ \frac{1}{n} \|X\beta\|_2^2 / \|\beta\|_2^2 : \mathrm{supp}(\beta) \subset F \right\},\\ \rho_+(F) =& \sup \left\{ \frac{1}{n} \|X\beta\|_2^2 / \|\beta\|_2^2 : \mathrm{supp}(\beta) \subset F \right\}.\end{aligned}$$

Moreover, for all $s > 0$, define

$$\begin{aligned}\rho_-(s) =& \inf\{\rho_-(F) : F \subset \mathcal{I}, c(F) \leq s\},\\ \rho_+(s) =& \sup\{\rho_+(F) : F \subset \mathcal{I}, c(F) \leq s\}.\end{aligned}$$

In the theoretical analysis, we need to assume that $\rho_-(s)$ is not too small for some s that is larger than the signal complexity. Since we only consider eigenvalues for submatrices with small cost $c(\bar{\beta})$, the sparse eigenvalue $\rho_-(s)$ can be significantly larger than the corresponding ratio for standard sparsity (which will consider all subsets of $\{1, \ldots, p\}$ up to size s). For example, for random projections used in compressive sensing applications, the coding length $c(\mathrm{supp}(\bar{\beta}))$ is $O(k \ln p)$ in standard sparsity, but can be as low as $c(\mathrm{supp}(\bar{\beta})) = O(k)$ in structured sparsity (if we can guess $\mathrm{supp}(\bar{\beta})$ approximately correctly. Therefore instead of requiring $n = O(k \ln p)$ samples, we requires only $O(k + \mathrm{cl}(\mathrm{supp}(\bar{\beta})))$. The difference can be significant when p is large and the coding length $\mathrm{cl}(\mathrm{supp}(\bar{\beta})) \ll k \ln p$. An example for this is group sparsity, where we have p/k_0 even sized groups, and variables in each group are simultaneously zero or nonzero. The coding length of the groups are $(k/k_0) \ln(p/k_0)$, which is significantly smaller than $k \ln p$ when p is large.

More precisely, we have the following random projection sample complexity bound for the structured sparse eigenvalue condition. The theorem implies that the structured RIP condition is satisfied with sample size $n = O((k/k_0) \ln(p/k_0))$ in group sparsity rather than $n = O(k \ln(p))$ in standard sparsity. Therefore Theorem 6.2 shows that in the compressive sensing applications, it is possible to reconstruct signals with fewer number of random projections by using group sparsity (or more general structured sparsity).

Theorem 6.1 (Structured-RIP) *Suppose that elements in X are iid standard Gaussian random variables $N(0, 1)$. For any $t > 0$ and $\delta \in (0, 1)$, let*

$$n \geq \frac{8}{\delta^2} [\ln 3 + t + s \ln(1 + 8/\delta)].$$

Then with probability at least $1 - e^{-t}$, the random matrix $X \in \mathbb{R}^{n \times p}$ satisfies the following structured-RIP inequality for all vector $\bar{\beta} \in \mathbb{R}^p$ with coding complexity no more than s:

$$(1 - \delta)\|\bar{\beta}\|_2 \leq \frac{1}{\sqrt{n}} \|X\bar{\beta}\|_2 \leq (1 + \delta)\|\bar{\beta}\|_2. \tag{4}$$

Although in the theorem, we assume Gaussian random matrix in order to state explicit constants, it is clear that similar results hold for other sub-Gaussian random matrices.

6.2 Coding complexity regularization

The following result gives a performance bound for constrained coding complexity regularization.

Theorem 6.2 *Suppose that Assumption 6.1 is valid. Consider any fixed target $\bar\beta \in \mathbb{R}^p$. Then with probability exceeding $1-\eta$, for all $\epsilon \geq 0$ and $\hat\beta \in \mathbb{R}^p$ such that: $\hat Q(\hat\beta) \leq \hat Q(\bar\beta) + \epsilon$, we have*

$$\|X\hat\beta - \mathbb{E}\mathbf{y}\|_2 \leq \|X\bar\beta - \mathbb{E}\mathbf{y}\|_2 + \sigma\sqrt{2\ln(6/\eta)} + 2(7.4\sigma^2 c(\bar\beta) + 4.7\sigma^2 \ln(6/\eta) + \epsilon)^{1/2}.$$

Moreover, if the coding scheme $c(\cdot)$ is sub-additive, then

$$n\rho_-(c(\hat\beta) + c(\bar\beta))\|\hat\beta - \bar\beta\|_2^2 \leq 10\|X\bar\beta - \mathbb{E}\mathbf{y}\|_2^2 + 37\sigma^2 c(\bar\beta) + 29\sigma^2\ln(6/\eta) + 2.5\epsilon.$$

This theorem immediately implies the following result for (1): $\forall\bar\beta$ such that $c(\bar\beta) \leq s$,

$$\frac{1}{\sqrt{n}}\|X\hat\beta_{constr} - \mathbb{E}\mathbf{y}\|_2 \leq \frac{1}{\sqrt{n}}\|X\bar\beta - \mathbb{E}\mathbf{y}\|_2 + \frac{\sigma}{\sqrt{n}}\sqrt{2\ln(6/\eta)} + \frac{2\sigma}{\sqrt{n}}(7.4s + 4.7\ln(6/\eta))^{1/2},$$
$$\|\hat\beta_{constr} - \bar\beta\|_2^2 \leq \frac{1}{\rho_-(s + c(\bar\beta))n}\left[10\|X\bar\beta - \mathbb{E}\mathbf{y}\|_2^2 + 37\sigma^2 s + 29\sigma^2\ln(6/\eta)\right].$$

Note that we generally expect $\rho_-(s + c(\bar\beta)) = O(1)$. The result immediately implies that as sample size $n \to \infty$ and $s/n \to 0$, the root mean squared error prediction performance $\|X\hat\beta - \mathbb{E}\mathbf{y}\|_2/\sqrt{n}$ converges to the optimal prediction performance $\inf_{c(\bar\beta)\leq s}\|X\bar\beta - \mathbb{E}\mathbf{y}\|_2/\sqrt{n}$. This result is agnostic in that even if $\|X\bar\beta - \mathbb{E}\mathbf{y}\|_2/\sqrt{n}$ is large, the result is still meaningful because it says the performance of the estimator $\hat\beta$ is competitive to the best possible estimator in the class $c(\bar\beta) \leq s$.

In compressive sensing applications, we take $\sigma = 0$, and we are interested in recovering $\bar\beta$ from random projections. For simplicity, we let $X\bar\beta = \mathbb{E}\mathbf{y} = \mathbf{y}$, and our result shows that the constrained coding complexity penalization method achieves exact reconstruction $\hat\beta_{constr} = \bar\beta$ as long as $\rho_-(2c(\bar\beta)) > 0$ (by setting $s = c(\bar\beta)$). According to Theorem 6.1, this is possible when the number of random projections (sample size) reaches $n = O(c(\bar\beta))$. This is a generalization of corresponding results in compressive sensing [6]. As we have pointed out earlier, this number can be significantly smaller than the standard sparsity requirement of $n = O(\|\bar\beta\|_0 \ln p)$, if the structure imposed in the formulation is meaningful.

Similar to Theorem 6.2, we can obtain the following result for (2). A related result for standard sparsity under Gaussian noise can be found in [5].

Theorem 6.3 *Suppose that Assumption 6.1 is valid. Consider any fixed target $\bar\beta \in \mathbb{R}^p$. Then with probability exceeding $1-\eta$, for all $\lambda > 7.4\sigma^2$ and $a \geq 7.4\sigma^2/(\lambda - 7.4\sigma^2)$, we have*

$$\|X\hat\beta_{pen} - \mathbb{E}\mathbf{y}\|_2^2 \leq (1+a)^2\|X\bar\beta - \mathbb{E}\mathbf{y}\|_2^2 + (1+a)\lambda c(\bar\beta) + \sigma^2(10 + 5a + 7a^{-1})\ln(6/\eta).$$

Unlike the result for (1), the prediction performance $\|X\hat\beta_{pen} - \mathbb{E}\mathbf{y}\|_2$ of the estimator in (2) is competitive to $(1+a)\|X\bar\beta - \mathbb{E}\mathbf{y}\|_2$, which is a constant factor larger than the optimal prediction performance $\|X\bar\beta - \mathbb{E}\mathbf{y}\|_2$. By optimizing λ and a, it is possible to obtain a similar result as that of Theorem 6.2. However, this requires tuning λ, which is not as convenient as tuning s in (1). Note that both results presented here, and those in [5] are superior to the more traditional least squares regression results with fixed λ. This is because one can only obtain the form presented in Theorem 6.2 by tuning λ. Therefore results with fixed λ are not suitable for real applications.

6.3 Structured greedy algorithm

We shall introduce a definition before stating our main results.

Definition 6.2 *Given $\mathcal{B} \subset 2^{\mathcal{I}}$, define*

$$\rho_0(\mathcal{B}) = \max_{B\in\mathcal{B}} \rho_+(B), \qquad c_0(\mathcal{B}) = \max_{B\in\mathcal{B}} c(B)$$

and

$$c(\bar\beta, \mathcal{B}) = \min\left\{ \sum_{j=1}^{b} c(\bar B_j) : \mathrm{supp}(\bar\beta) \subset \bigcup_{j=1}^{b} \bar B_j \quad (\bar B_j \in \mathcal{B}) \right\}.$$

The following theorem shows that if $c(\bar\beta, \mathcal{B})$ is small, then one can use the structured greedy algorithm to find a coefficient vector $\beta^{(k)}$ that is competitive to $\bar\beta$, and the coding complexity $c(\beta^{(k)})$ is not much worse than that of $c(\bar\beta, \mathcal{B})$. This implies that if the original coding complexity $c(\bar\beta)$ can be approximated by block complexity $c(\bar\beta, \mathcal{B})$, then we can approximately solve (1).

Theorem 6.4 *Suppose the coding scheme is sub-additive. Consider $\bar\beta$ and ϵ such that*

$$\epsilon \in (0, \|\mathbf{y}\|_2^2 - \|X\bar\beta - \mathbf{y}\|_2^2]$$

and

$$s \geq \frac{\rho_0(\mathcal{B}) c(\bar\beta, \mathcal{B})}{\gamma \rho_-(s + c(\bar\beta))} \ln \frac{\|\mathbf{y}\|_2^2 - \|X\bar\beta - \mathbf{y}\|_2^2}{\epsilon}.$$

Then at the stopping time k, we have

$$\hat Q(\beta^{(k)}) \leq \hat Q(\bar\beta) + \epsilon.$$

By Theorem 6.2, the result in Theorem 6.4 implies that

$$\|X\beta^{(k)} - \mathbb{E}\mathbf{y}\|_2 \leq \|X\bar\beta - \mathbb{E}\mathbf{y}\|_2 + \sigma\sqrt{2\ln(6/\eta)} + 2\sigma(7.4(s + c_0(\mathcal{B})) + 4.7\ln(6/\eta) + \epsilon/\sigma^2)^{1/2},$$

$$\|\beta^{(k)} - \bar\beta\|_2^2 \leq \frac{1}{\rho_-(s + c_0(\mathcal{B}) + c(\bar\beta))n}\left[10\|X\bar\beta - \mathbb{E}\mathbf{y}\|_2^2 + 37\sigma^2(s + c_0(\mathcal{B})) + 29\sigma^2\ln(6/\eta) + 2.5\epsilon\right].$$

The result shows that in order to approximate a signal $\bar\beta$ up to accuracy ϵ, one needs to use coding complexity $O(\ln(1/\epsilon))c(\bar\beta, \mathcal{B})$. If $\mathcal{B}$ contains small blocks and their sub-blocks with equal coding length, and the coding scheme is block coding generated by $\mathcal{B}$, then $c(\bar\beta, \mathcal{B}) = c(\bar\beta)$. In this case we need $O(s\ln(1/\epsilon))$ to approximate a signal with coding complexity s.

In order to get rid of the $O(\ln(1/\epsilon))$ factor, backward greedy strategies can be employed, as shown in various recent work such as [26]. For simplicity, we will not analyze such strategies in this paper. However, in the following, we present an additional convergence result for structured greedy algorithm that can be applied to weakly sparse p-compressible signals common in practice. It is shown that the $\ln(1/\epsilon)$ can be removed for such weakly sparse signals.

Theorem 6.5 *Suppose the coding scheme is sub-additive. Given a sequence of targets $\bar\beta_j$ such that $\hat Q(\bar\beta_0) \leq \hat Q(\bar\beta_1) \leq \cdots$ and $c(\bar\beta_j, \mathcal{B}) \leq c(\bar\beta_0, \mathcal{B})/2^j$. If*

$$s \geq \frac{\rho_0(\mathcal{B})}{\gamma \min_j \rho_-(s + c(\bar\beta_j))} c(\bar\beta_0, \mathcal{B}) \left[3.4 + \sum_{j=0}^{\infty} 2^{-j} \ln \frac{\hat Q(\bar\beta_{j+1}) - \hat Q(\bar\beta_0) + \epsilon}{\hat Q(\bar\beta_j) - \hat Q(\bar\beta_0) + \epsilon}\right]$$

for some $\epsilon > 0$. *Then at the stopping time* k, *we have*

$$\hat{Q}(\beta^{(k)}) \leq \hat{Q}(\bar{\beta}_0) + \epsilon.$$

In the above theorem, we can see that if the signal is only weakly sparse, in that $(\hat{Q}(\bar{\beta}_{j+1}) - \hat{Q}(\bar{\beta}_0) + \epsilon)/(\hat{Q}(\bar{\beta}_j) - \hat{Q}(\bar{\beta}_0) + \epsilon)$ grows sub-exponentially in j, then we can choose $s = O(c(\bar{\beta}_0, \mathcal{B}))$. This means that we can find $\beta^{(k)}$ of complexity $s = O(c(\bar{\beta}_0, \mathcal{B}))$ to approximate a signal $\bar{\beta}_0$. The worst case scenario is when $\hat{Q}(\bar{\beta}_1) \approx \hat{Q}(0)$, which reduces to the $s = O(c(\bar{\beta}_0, \mathcal{B}) \log(1/\epsilon))$ complexity in Theorem 6.4.

As an application, we introduce the following concept of weakly sparse compressible target that generalizes the corresponding concept of compressible signal in standard sparsity from the compressive sensing literature [9].

Definition 6.3 *The target* $\mathbb{E}\mathbf{y}$ *is* (a, q)*-compressible with respect to block* $\mathcal{B}$ *if there exist constants* $a, q > 0$ *such that for each* $s > 0$, $\exists \bar{\beta}(s)$ *such that* $c(\bar{\beta}(s), \mathcal{B}) \leq s$ *and*

$$\frac{1}{n}\|X\bar{\beta}(s) - \mathbb{E}\mathbf{y}\|_2^2 \leq a s^{-q}.$$

Corollary 6.1 *Suppose that the target is* (a, q)*-compressible with respect to* $\mathcal{B}$. *Then with probability* $1 - \eta$, *at the stopping time* k, *we have*

$$\hat{Q}(\beta^{(k)}) \leq \hat{Q}(\bar{\beta}(s')) + 2na/s'^q + 2\sigma^2[\ln(2/\eta) + 1],$$

where

$$s' \leq \frac{s\,\gamma}{(10 + 3q)\rho_0(\mathcal{B})} \min_{u \leq s'} \rho_-(s + c(\bar{\beta}(u))).$$

If we assume the underlying coding scheme is block coding generated by $\mathcal{B}$, then $\min_{u \leq s'} \rho_-(s + c(\bar{\beta}(u))) \leq \rho_-(s + s')$. The corollary shows that we can approximate a compressible signal of complexity s' with complexity $s = O(qs')$ using greedy algorithm. This means the greedy algorithm obtains optimal rate for weakly-sparse compressible signals. The sample complexity suffers only a constant factor $O(q)$. Combine this result with Theorem 6.2, and take union bound, we have with probability $1 - 2\eta$, at stopping time k:

$$\frac{1}{\sqrt{n}}\|X\beta^{(k)} - \mathbb{E}\mathbf{y}\|_2 \leq \sqrt{\frac{a}{s'^q}} + \sigma\sqrt{\frac{2\ln(6/\eta)}{n}} + 2\sigma\sqrt{\frac{7.4(s + c_0(\mathcal{B})) + 6.7\ln(6/\eta)}{n} + \frac{2a}{\sigma^2 s'^q}},$$
$$\|\beta^{(k)} - \bar{\beta}(s')\|_2^2 \leq \frac{1}{\rho_-(s + s' + c_0(\mathcal{B}))}\left[\frac{15a}{s'^q} + \frac{37\sigma^2(s + c_0(\mathcal{B})) + 34\sigma^2\ln(6/\eta)}{n}\right].$$

Given a fixed n, we can obtain a convergence result by choosing s (and thus s') to optimize the right hand side. The resulting rate is optimal for the special case of standard sparsity, which implies that the bound has the optimal form for structured q-compressible targets. In particular, in compressive sensing applications where $\sigma = 0$, we obtain when sample size reaches $n = O(qs')$, the reconstruction performance is

$$\|\bar{\beta}^{(k)} - \bar{\beta}\|_2^2 = O(a/s'^q),$$

which matches that of the constrained coding complexity regularization method in (1) up to a constant $O(q)$. Since many real data involve weakly sparse signals, our result provides strong theoretical justification for the use of OMP in such problems. Our experiments are consistent with the theory.

7 Experiments

The purpose of these experiments is to demonstrate the advantage of structured sparsity over standard sparsity. We compare the proposed StructOMP to OMP and Lasso, which are standard algorithms to achieve sparsity but without considering structure. In our experiments, we use Lasso-modified least angle regression (LAS/Lasso) as the solver of Lasso [4]. We also compare StructOMP with group Lasso [25, 3]. In order to quantitatively compare performance of different algorithms, we use recovery error, defined as the relative difference in 2-norm between the estimated sparse coefficient vector $\hat{\beta}_{est}$ and the ground-truth sparse coefficient $\bar{\beta}$: $\|\hat{\beta}_{est} - \bar{\beta}\|_2 / \|\bar{\beta}\|_2$. Our experiments focus on graph sparsity, with several different underlying graph structures. Note that graph sparsity is more general than group sparsity; in fact connected regions may be regarded as dynamic groups that are not pre-defined.

7.1 Simulated 1D Signals with Line-Structured Sparsity

In the first experiment, we randomly generate a $1D$ structured sparse signal with values ± 1, where $p = 512$, $k = 64$ and $g = 4$. The support set of these signals is composed of g connected regions. Here, each component of the sparse coefficient is connected to two of its adjacent components, which forms the underlying graph structure. The graph sparsity concept introduced earlier is used to compute the coding length of sparsity patterns in StructOMP. The projection matrix X is generated by creating an $n \times p$ matrix with i.i.d. draws from a standard Gaussian distribution $N(0, 1)$. For simplicity, the rows of X are normalized to unit magnitude. Zero-mean Gaussian noise with standard deviation $\sigma = 0.01$ is added to the measurements. Our task is to compare the recovery performance of StructOMP to those of OMP, Lasso and group Lasso for these structured sparsity signals.

Figure 2 shows one instance of generated signal and the corresponding recovered results by different algorithms when $n = 160$. Since the sample size n is not big enough, OMP and Lasso do not achieve good recovery results, whereas the StructOMP algorithm achieves near perfect recovery of the original signal. As we do not know the predefined groups for group Lasso, we just try group Lasso with several different group sizes (gs=2, 4, 8, 16). Although the results obtained with group Lasso are better than those of OMP and Lasso, they are still inferior to the results with StructOMP. To study how the sample size n effects the recovery performance, we vary the sample size and record the recovery results by different algorithms. To reduce the randomness, we perform the experiment 100 times for each sample size. Figure 3(a) shows the recovery performance of the three algorithms, averaged over 100 random runs for each sample size. As expected, StructOMP is better than the group Lasso and far better than the OMP and Lasso. The results show that the proposed StructOMP can achieve better recovery performance for structured sparsity signals with less samples.

Note that Lasso performs better than OMP in the first example. This is because the signal is strongly sparse (that is, all nonzero coefficients are significantly different from zero). In the second experiment, we randomly generate a $1D$ structured sparse signal with weak sparsity, where the nonzero coefficients decay gradually to zero, but there is no clear cutoff. One instance of generated signal is shown in Figure 4 (a). Here, $p = 512$ and all coefficient of the signal are not zeros. We define the sparsity k as the number of coefficients that contain 95% of the image energy. The support set of these signals is composed of $g = 2$ connected regions. Again, each element of the sparse coefficient is connected to two of its adjacent elements, which forms the underlying 1D line graph

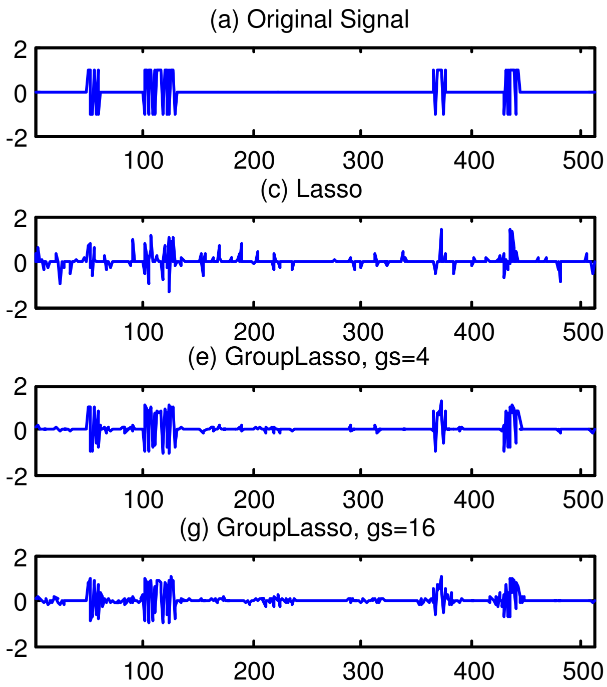
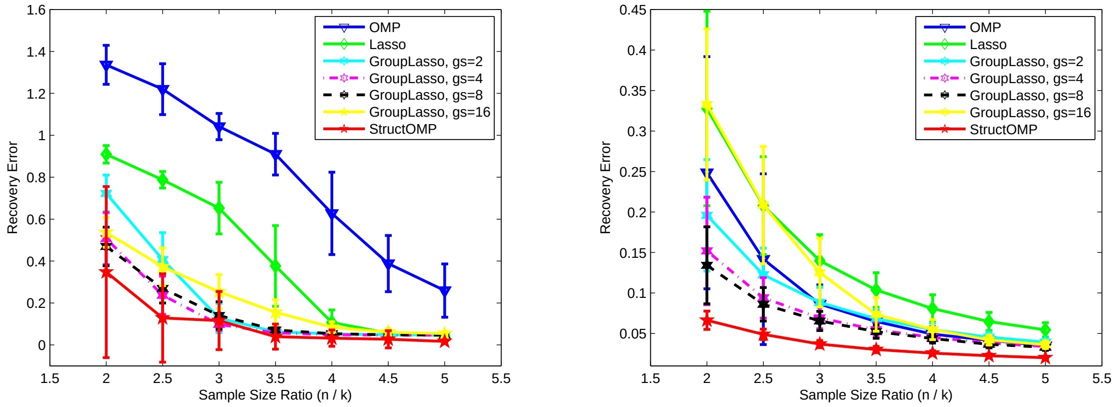


Figure 2: Recovery results of 1D signal with line-structured sparsity. (a) original data; (b) recovered results with OMP (error is 0.9921); (c) recovered results with Lasso (error is 0.8660);; (d) recovered results with Group Lasso (error is 0.4832 with group size gs=2); (e) recovered results with Group Lasso (error is 0.4832 with group size gs=4);(f) recovered results with Group Lasso (error is 0.2646 with group size gs=8);(g) recovered results with Group Lasso (error is 0.3980 with group size gs=16); (h) recovered results with StructOMP (error is 0.0246).



(a)

(b)

Figure 3: Recovery error vs. Sample size ratio (n/k): a) 1D signals; (b) 1D Weak sparse signal

structure. The graph sparsity concept introduced earlier is used to compute the coding length of sparsity patterns in StructOMP. The projection matrix X is generated by creating an $n \times p$ matrix with i.i.d. draws from a standard Gaussian distribution $N(0, 1)$. For simplicity, the rows of X are normalized to unit magnitude. Zero-mean Gaussian noise with standard deviation $\sigma = 0.01$ is added to the measurements.

Figure 4 shows one generated signal and its recovered results by different algorithms when $k = 32$ and $n = 48$. Again, we observe that OMP and Lasso do not achieve good recovery results, whereas the StructOMP algorithm achieves near perfect recovery of the original signal. As we do not know the predefined groups for group Lasso, we just try group Lasso with several different group sizes (gs=2, 4, 8, 16). Although the results obtained with group Lasso are better than those of OMP and Lasso, they are still inferior to the results with StructOMP. In order to study how the sample size n effects the recovery performance, we vary the sample size and record the recovery results by different algorithms. To reduce the randomness, we perform the experiment 100 times for each of the sample sizes. Figure 3(b) shows the recovery performance of different algorithms, averaged over 100 random runs for each sample size. As expected, StructOMP algorithm is superior in all cases. What's different from the first experiment is that the recovery error of OMP becomes smaller than that of Lasso. This result is consistent with our theory, which predicts that if the underlying signal is weakly sparse, then the relatively performance of OMP becomes comparable to Lasso.

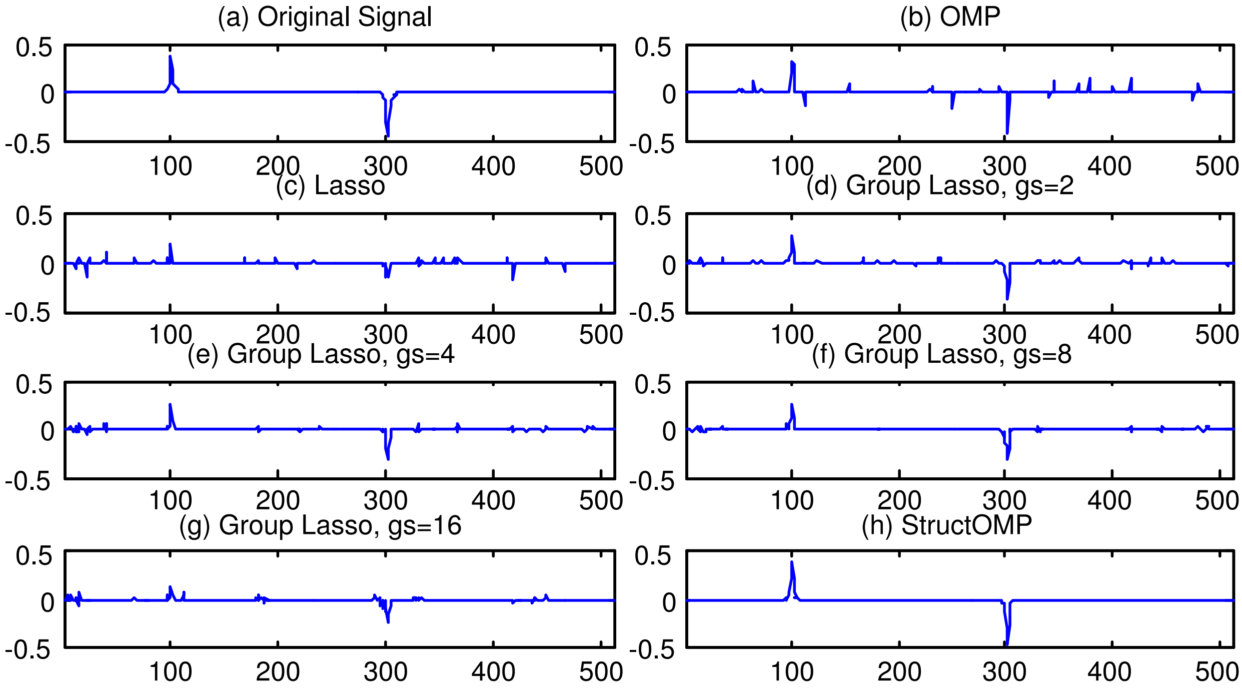


Figure 4: Recovery results of 1D weakly sparse signal with line-structured sparsity. (a) original data; (b) recovered results with OMP (error is 0.5599); (c) recovered results with Lasso (error is 0.6686);; (d) recovered results with Group Lasso (error is 0.4732 with group size gs=2); (e) recovered results with Group Lasso (error is 0.2893 with group size gs=4);(f) recovered results with Group Lasso (error is 0.2646 with group size gs=8);(g) recovered results with Group Lasso (error is 0.5459 with group size gs=16); (h) recovered results with StructOMP (error is 0.0846).

7.2 2D Image Compressive Sensing with Tree-structured Sparsity

It is well known that 2D natural images are sparse in a wavelet basis. Their wavelet coefficients have a hierarchical tree structure, which is widely used for wavelet-based compression algorithms [21]. Figure 5(a) shows a widely used example image with size 64×64: *cameraman*. Each 2D wavelet coefficient of this image is connected to its parent coefficient and child coefficients, which forms the underlying hierarchical tree structure (which is a special case of graph sparsity). In our experiment, we choose Haar-wavelet to obtain its tree-structured sparsity wavelet coefficients. The projection matrix X and noises are generated with the same method as that for 1D structured sparsity signals. OMP, Lasso and StructOMP are used to recover the wavelet coefficients from the random projection samples respectively. Then, the inverse wavelet transform is used to reconstruct the images with these recovered wavelet coefficients. Our task is to compare the recovery performance of the StructOMP to those of OMP and Lasso.

Figure 5 shows one example of the recovered results by different algorithms. It shows that StructOMP obtains the best recovered result. Figure 6(a) shows the recovery performance of the three algorithms, averaged over 100 random runs for each sample size. The StructOMP algorithm is better than both Lasso and OMP in this case. Since real image data are weakly sparse, the performance of standard OMP (without structured sparsity) is similar to that of Lasso.

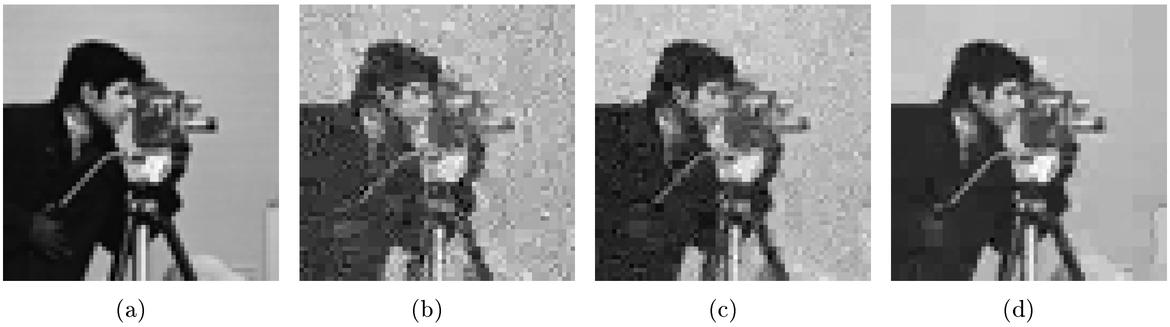


Figure 5: Recovery results with sample size $n = 2048$: (a) the background subtracted image, (b) recovered image with OMP (error is 0.21986), (c) recovered image with Lasso (error is 0.1670) and (d) recovered image with StructOMP (error is 0.0375)

7.3 Background Subtracted Images for Robust Surveillance

Background subtracted images are typical structure sparsity data in static video surveillance applications. They generally correspond to the foreground objects of interest. Unlike the whole scene, these images are not only spatially sparse but also inclined to cluster into groups, which correspond to different foreground objects. Thus, the StructOMP algorithm can obtain superior recovery from compressive sensing measurements that are received by a centralized server from multiple and randomly placed optical sensors. In this experiment, the testing video is downloaded from http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/. The background subtracted images are obtained with the software [28]. One sample image frame is shown in Figure 7(a). The support set of 2D images is thus composed of several connected regions. Here, each pixel of the 2D background subtracted image is connected to four of its adjacent pixels, forming the underlying graph structure in graph sparsity. The results shown in Figure 7 demonstrate that the StructOMP outperforms

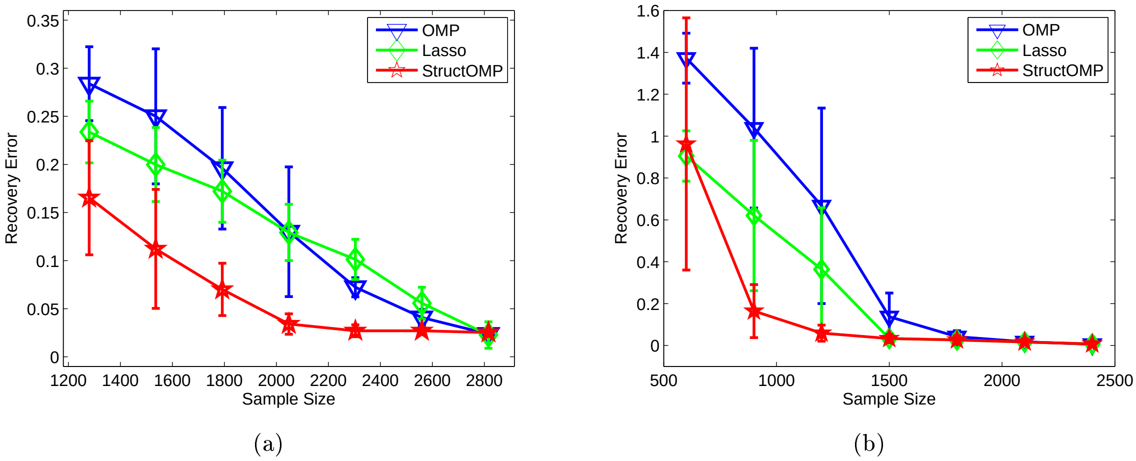


Figure 6: Recovery error vs. Sample size: a) 2D image with tree-structured sparsity in wavelet basis; (b) background subtracted images with structured sparsity

both OMP and Lasso in recovery. We randomly choose 100 background subtracted images as test images. Figure 6(b) shows the recovery performance as a function of increasing sample sizes. It demonstrates again that StructOMP significantly outperforms OMP and Lasso in recovery performance on video data. Comparing to the image compression example in the previous section, the background subtracted images have a more clearly defined sparsity pattern where nonzero coefficients are generally distinct from zero (that is, stronger sparsity); this explains why Lasso performs better than the standard (unstructured) OMP on this particular data. The result is again consistent with our theory.

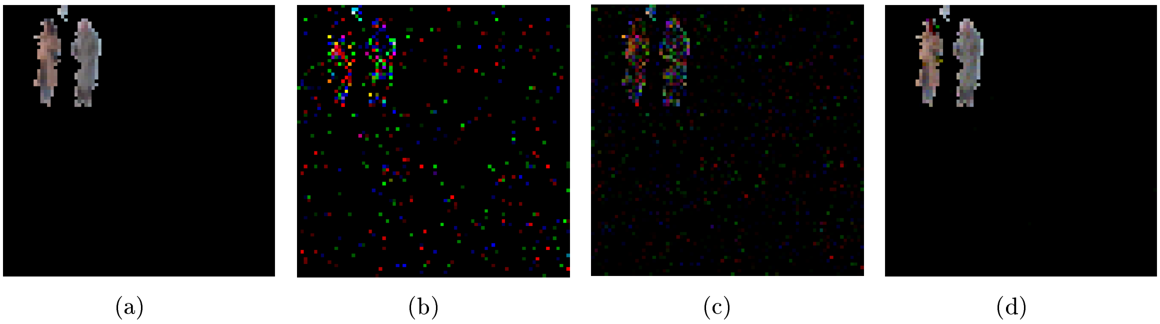


Figure 7: Recovery results with sample size $n = 900$: (a) the background subtracted image, (b) recovered image with OMP (error is 1.1833), (c) recovered image with Lasso (error is 0.7075) and (d) recovered image with StructOMP (error is 0.1203)

8 Discussion

This paper develops a theory for structured sparsity where prior knowledge allows us to prefer certain sparsity patterns to others. Some examples are presented to illustrate the concept. The general framework established in this paper includes the recently popularized group sparsity idea has a special case.

In structured sparsity, the complexity of learning is measured by the coding complexity $c(\bar{\beta}) \leq \|\bar{\beta}\|_0 + \mathrm{cl}(\mathrm{supp}(\bar{\beta}))$ instead of $\|\bar{\beta}\|_0 \ln p$ which determines the complexity in standard sparsity. Using this notation, a theory parallel to that of the standard sparsity is developed. The theory shows that if the coding length $\mathrm{cl}(\mathrm{supp}(\bar{\beta}))$ is small for a target coefficient vector $\bar{\beta}$, then the complexity of learning $\bar{\beta}$ can be significantly smaller than the corresponding complexity in standard sparsity. Experimental results demonstrate that significant improvements can be obtained on some real problems that have natural structures.

The structured greedy algorithm presented in this paper is the first efficient algorithm proposed to handle the general structured sparsity learning. It is shown that the algorithm is effective under appropriate conditions. Future work include additional computationally efficient methods such as convex relaxation methods (e.g. L_1 regularization for standard sparsity, and group Lasso for strong group sparsity) and backward greedy strategies to improve the forward greedy method considered in this paper.

References

[1] Francis R. Bach. Consistency of the group lasso and multiple kernel learning. *JMLR*, 9:1179–1225, 2008.

[2] R. Baraniuk, V. Cevher, M. Duarte, and C. Hegde. Model based compressive sensing. 2008. preprint.

[3] E. Berg, M. Schmidt, M. Friedlander, and K. Murphy. Group sparsity via linear-time projection. 2008. Preprint.

[4] Iain Johnstone Bradley Efron, Trevor Hastie and Robert Tibshirani. Least angle regression. *Annals of Statistics*, 32:407–499, 2004.

[5] Florentina Bunea, Alexandre B. Tsybakov, and Marten H. Wegkamp. Aggregation for Gaussian regression. *Annals of Statistics*, 35:1674–1697, 2007.

[6] Emmanuel J. Candes and Terence Tao. Decoding by linear programming. *IEEE Trans. on Information Theory*, 51:4203–4215, 2005.

[7] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory.* Wiley-Interscience, 1991.

[8] L. Daudet. Sparse and structured decomposition of audio signals in overcomplete spaces. In *International Conference on Digital Audio Effects*, 2004.

[9] D. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52:1289–1306, 2006.

[10] D. Grimm, T. Netzer, and M. Schweighofer. A note on the representation of positive polynomials with structured sparsity. *Arch. Math.*, 89:399–403, 2007.

[11] L. He and L. Carin. Exploiting structure in wavelet-based bayesian compressive sensing. In *Preprint*, 2008.

[12] L. He and L. Carin. Exploiting structure in compressive sensing with a jpeg basis. In *Preprint*, 2009.

[13] Junzhou Huang and Tong Zhang. The benefit of group sparsity. Technical report, Rutgers University, January 2009. Available from http://arxiv.org/abs/0901.2962.

[14] S. Ji, D. Dunson, and L. Carin. Multi-task compressive sensing. *IEEE Transactions on Signal Processing*, 2008. Accepted.

[15] Vladimir Koltchinskii and Ming Yuan. Sparse recovery in large ensembles of kernel machines. In *COLT'08*, 2008.

[16] M. Kowalski and B. Torresani. Structured sparsity: from mixed norms to structured shrinkage. In *Workshop on Signal Processing with Adaptive Sparse Representations*, 2009.

[17] S. Mallat. In *A Wavelet Tour of Signal Processing.* Academic Press.

[18] Yuval Nardi and Alessandro Rinaldo. On the asymptotic properties of the group lasso estimator for linear models. *Electronic Journal of Statistics*, 2:605–633, 2008.

[19] G. Obozinski, M. J. Wainwright, and M. I. Jordan. Union support recovery in high-dimensional multivariate regression. Technical Report 761, UC Berkeley, 2008.

[20] G. Pisier. The volume of convex bodies and banach space geometry. 1989. Cambridge University Press.

[21] Jerome M. Shapiro. Embedded image coding using zerotrees of wavelet coefficients. *IEEE Transactions on Signal Processing*, 41:3445–3462, 1993.

[22] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 58:267–288, 1996.

[23] J. Tropp and A. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on Information Theory*, 53(12):4655–4666, 2007.

[24] D. Wipf and B. Rao. An empirical Bayesian strategy for solving the simultaneous sparse approximation problem. *IEEE Transactions on Signal Processing*, 55(7):3704–3716, 2007.

[25] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of The Royal Statistical Society Series B*, 68(1):49–67, 2006.

[26] Tong Zhang. Adpative forward-backward greedy algorithm for learning sparse representations. Technical report, Rutgers Statistics Department, 2008. A short version appeared in NIPS 08.

[27] P. Zhao, G. Rocha, and B. Yu. Grouped and hierarchical model selection through composite absolute penalties. *The Annals of Statistics.* to appear.

[28] Z. Zivkovic and F. Heijden. Efficient adaptive density estimation per image pixel for the task of background subtraction. *Pattern Recognition Letters*, 27(7):773–780, 2006.

A Proof of Proposition 6.1

Lemma A.1 *Consider a fixed vector* $\mathbf{x} \in \mathbb{R}^n$*, and a random vector* $\mathbf{y} \in \mathbb{R}^n$ *with independent sub-Gaussian components:* $\mathbb{E}e^{t(\mathbf{y}_i - \mathbb{E}\mathbf{y}_i)} \leq e^{\sigma^2 t^2/2}$ *for all* t *and* i*, then* $\forall \epsilon > 0$*:*

$$\Pr\left(\left|\mathbf{x}^\top \mathbf{y} - \mathbb{E}\mathbf{x}^\top \mathbf{y}\right| \geq \epsilon\right) \leq 2e^{-\epsilon^2/(2\sigma^2\|\mathbf{x}\|_2^2)}.$$

Proof Let $s_n = \sum_{i=1}^n (\mathbf{x}_i\mathbf{y}_i - \mathbb{E}\mathbf{x}_i\mathbf{y}_i)$; then by assumption, $\mathbb{E}(e^{ts_n} + e^{-ts_n}) \leq 2e^{\sum_i \mathbf{x}_i^2\sigma^2t^2/2}$, which implies that $\Pr(|s_n| \geq \epsilon)e^{t\epsilon} \leq 2e^{\sum_i \mathbf{x}_i^2\sigma^2t^2/2}$. Now let $t = \epsilon/(\sum_i \mathbf{x}_i^2\sigma^2)$, we obtain the desired bound. ■

The following lemma is taken from [20]. Since the proof is simple, it is included for completeness.

Lemma A.2 *Consider the unit sphere* $S^{k-1} = \{x : \|x\|_2 = 1\}$ *in* $\mathbb{R}^k$ *(*$k \geq 1$*). Given any* $\varepsilon > 0$*, there exists an* ε*-cover* $Q \subset S^{k-1}$ *such that* $\min_{q\in Q}\|x - q\|_2 \leq \varepsilon$ *for all* $\|x\|_2 = 1$*, with* $|Q| \leq (1 + 2/\varepsilon)^k$*.*

Proof Let $B^k = \{x : \|x\|_2 \leq 1\}$ be the unit ball in $\mathbb{R}^k$. Let $Q = \{q_i\}_{i=1,\ldots,|Q|} \subset S^{k-1}$ be a maximal subset such that $\|q_i - q_j\|_2 > \varepsilon$ for all $i \neq j$. By maximality, Q is an ε-cover of S^{k-1}. Since the balls $q_i + (\varepsilon/2)B^k$ are disjoint and belong to $(1 + \varepsilon/2)B^k$, we have

$$\sum_{i\leq|Q|} vol(q_i + (\varepsilon/2)B^k) \leq vol((1 + \varepsilon/2)B^k).$$

Therefore,

$$|Q|(\varepsilon/2)^k vol(B^k) \leq (1 + \varepsilon/2)^k vol(B^k),$$

which implies that $|Q| \leq (1 + 2/\varepsilon)^k$. ■

Proof of Proposition 6.1

According to Lemma A.2, given $\epsilon_1 > 0$, there exists a finite set $Q = \{q_i\}$ with $|Q| \leq (1 + 2/\epsilon_1)^k$ such that $\|Pq_i\|_2 = 1$ for all i, and $\min_i \|P\beta - Pq_i\|_2 \leq \epsilon_1$ for all $\|P\beta\|_2 = 1$.

For each i, Lemma A.1 implies that $\forall\epsilon_2 > 0$:

$$\Pr\left(\left|q_i^\top P(\mathbf{y} - \mathbb{E}\mathbf{y})\right| \geq \epsilon_2\right) \leq 2e^{-\epsilon_2^2/(2\sigma^2)}.$$

Taking union bound for all $q_i \in Q$, we obtain with probability exceeding $1 - 2(1 + 2/\epsilon_1)^k e^{-\epsilon_2^2/2\sigma^2}$:

$$\left|q_i^\top P(\mathbf{y} - \mathbb{E}\mathbf{y})\right| \leq \epsilon_2$$

for all i.

Let $\beta = P(\mathbf{y} - \mathbb{E}\mathbf{y})/\|P(\mathbf{y} - \mathbb{E}\mathbf{y})\|_2$, then there exists i such that $\|P\beta - Pq_i\|_2 \leq \epsilon_1$. We have

$$\begin{aligned}\|P(\mathbf{y} - \mathbb{E}\mathbf{y})\|_2 =& \beta^\top(\mathbf{y} - \mathbb{E}\mathbf{y}) \\ \leq& \|P\beta - Pq_i\|_2\|P(\mathbf{y} - \mathbb{E}\mathbf{y})\|_2 + |q_i^\top P(\mathbf{y} - \mathbb{E}\mathbf{y})| \\ \leq& \epsilon_1\|P(\mathbf{y} - \mathbb{E}\mathbf{y})\|_2 + \epsilon_2.\end{aligned}$$

Therefore

$$\|P(\mathbf{y} - \mathbb{E}\mathbf{y})\|_2 \le \epsilon_2/(1-\epsilon_1).$$

Let $\epsilon_1 = 2/15$, and $\eta = 2(1+2/\epsilon_1)^k e^{-\epsilon_2^2/2\sigma^2}$, we have

$$\epsilon_2^2 = 2\sigma^2[(4k+1)\ln 2 - \ln\eta],$$

and thus

$$\|P(\mathbf{y} - \mathbb{E}\mathbf{y})\|_2 \le \frac{15}{13}\sigma\sqrt{2(4k+1)\ln 2 - 2\ln\eta}.$$

This simplifies to the desired bound.

B Proof of Theorem 6.1

We use the following lemma from [13].

Lemma B.1 *Suppose X is generated according to Theorem 6.1. For any fixed set $F \subset \mathcal{I}$ with $|F| = k$ and $0 < \delta < 1$, we have with probability exceeding $1 - 3(1+8/\delta)^k e^{-n\delta^2/8}$:*

$$(1-\delta)\|\beta\|_2 \le \frac{1}{\sqrt{n}}\|X_F\beta\|_2 \le (1+\delta)\|\beta\|_2 \tag{5}$$

for all $\beta \in \mathbb{R}^k$.

Proof of Theorem 6.1

Since $\mathrm{cl}(F)$ is a coding length, we have

$$\begin{aligned}\sum_{F:|F|+\mathrm{cl}(F)\le s}(1+8/\delta)^{|F|} \le &\sum_{F:|F|+\gamma\mathrm{cl}(F)\le s}(1+8/\delta)^{|F|}\\ \le &\sum_F (1+8/\delta)^{s-\gamma\mathrm{cl}(F)} = (1+8/\delta)^s\sum_F 2^{-\mathrm{cl}(F)} \le (1+8/\delta)^s,\end{aligned}$$

where we let $\gamma = 1/\log_2(1+8/\delta)$ in the above derivation.

For each F, we know from Lemma B.1 that for all β such that $\mathrm{supp}(\beta) \subset F$:

$$(1-\delta)\|\beta\|_2 \le \frac{1}{\sqrt{n}}\|X\beta\|_2 \le (1+\delta)\|\beta\|_2$$

with probability exceeding $1 - 3(1+8/\delta)^{|F|}e^{-n\delta^2/8}$.

We can thus take the union bound over $F : |F| + \mathrm{cl}(F) \le s$, which shows that with probability exceeding

$$1 - \sum_{F:|F|+\mathrm{cl}(F)\le s} 3(1+8/\delta)^{|F|}e^{-n\delta^2/8},$$

the structured RIP in Equation (4) holds. Since

$$\sum_{F:|F|+\mathrm{cl}(F)\le s} 3(1+8/\delta)^{|F|}e^{-n\delta^2/8} \le 3(1+8/\delta)^s e^{-n\delta^2/8} \le e^{-t},$$

we obtain the desired bound.

C Proof of Theorem 6.2 and Theorem 6.3

Lemma C.1 *Suppose that Assumption 6.1 is valid. For any fixed subset $F \subset \mathcal{I}$, we have with probability $1-\eta$, $\forall \beta$ such that $\mathrm{supp}(\beta) \subset F$, and $a > 0$, we have*

$$\|X\beta - \mathbb{E}\mathbf{y}\|_2^2 \leq (1+a)[\|X\beta - \mathbf{y}\|_2^2 - \|\mathbf{y} - \mathbb{E}\mathbf{y}\|_2^2] + (2+a+a^{-1})\sigma^2[7.4|F| + 4.7\ln(4/\eta)].$$

Proof Let

$$P_F = X_F(X_F^\top X_F)^{-1}X_F^\top$$

be projection matrix to the subspaces generated by columns of X_F.

Let $\tilde{a} = (I-P_F)\mathbb{E}\mathbf{y}/\|(I-P_F)\mathbb{E}\mathbf{y}\|_2$, $\delta_1 = \|P_F(\mathbf{y} - \mathbb{E}\mathbf{y})\|_2$ and $\delta_2 = |\tilde{a}^\top(\mathbf{y} - \mathbb{E}\mathbf{y})|$, we have

$$\begin{aligned}
&\|X\beta - \mathbb{E}\mathbf{y}\|_2^2\\
=&\|X\beta - \mathbf{y}\|_2^2 - \|\mathbf{y} - \mathbb{E}\mathbf{y}\|_2^2 + 2(\mathbf{y} - \mathbb{E}\mathbf{y})^\top(X\beta - \mathbb{E}\mathbf{y})\\
=&\|X\beta - \mathbf{y}\|_2^2 - \|\mathbf{y} - \mathbb{E}\mathbf{y}\|_2^2 + 2(\mathbf{y} - \mathbb{E}\mathbf{y})^\top(X\beta - P_F\mathbb{E}\mathbf{y}) - 2\tilde{a}^\top(\mathbf{y} - \mathbb{E}\mathbf{y})\|(I-P_F)\mathbb{E}\mathbf{y}\|_2\\
=&\|X\beta - \mathbf{y}\|_2^2 - \|\mathbf{y} - \mathbb{E}\mathbf{y}\|_2^2 + 2(\mathbf{y} - \mathbb{E}\mathbf{y})^\top P_F(X\beta - P_F\mathbb{E}\mathbf{y}) - 2\tilde{a}^\top(\mathbf{y} - \mathbb{E}\mathbf{y})\|(I-P_F)\mathbb{E}\mathbf{y}\|_2\\
\leq&\|X\beta - \mathbf{y}\|_2^2 - \|\mathbf{y} - \mathbb{E}\mathbf{y}\|_2^2 + 2\delta_1\|X\beta - P_F\mathbb{E}\mathbf{y}\|_2 + 2\delta_2\|(I-P_F)\mathbb{E}\mathbf{y}\|_2\\
\leq&\|X\beta - \mathbf{y}\|_2^2 - \|\mathbf{y} - \mathbb{E}\mathbf{y}\|_2^2 + 2\sqrt{\delta_1^2+\delta_2^2}\sqrt{\|X\beta - P_F\mathbb{E}\mathbf{y}\|_2^2 + \|(I-P_F)\mathbb{E}\mathbf{y}\|_2^2}\\
=&\|X\beta - \mathbf{y}\|_2^2 - \|\mathbf{y} - \mathbb{E}\mathbf{y}\|_2^2 + 2\sqrt{\delta_1^2+\delta_2^2}\|X\beta - \mathbb{E}\mathbf{y}\|_2.
\end{aligned}$$

Note that in the above derivation, we have used the fact that $P_F X\beta = X\beta$, and $\|X\beta - P_F\mathbb{E}\mathbf{y}\|_2^2 + \|(I-P_F)\mathbb{E}\mathbf{y}\|_2^2 = \|X\beta - \mathbb{E}\mathbf{y}\|_2^2$.

Now, by solving the above inequality, we obtain

$$\begin{aligned}
\|X\beta - \mathbb{E}\mathbf{y}\|_2^2 \leq& \left[\sqrt{\|X\beta - \mathbf{y}\|_2^2 - \|\mathbf{y} - \mathbb{E}\mathbf{y}\|_2^2 + \delta_1^2 + \delta_2^2} + \sqrt{\delta_1^2+\delta_2^2}\right]^2\\
\leq&(1+a)[\|X\beta - \mathbf{y}\|_2^2 - \|\mathbf{y} - \mathbb{E}\mathbf{y}\|_2^2] + (2+a+1/a)(\delta_1^2+\delta_2^2).
\end{aligned}$$

The desired bound now follows easily from Proposition 6.1 and Lemma A.1, where we know that with probability $1-\eta/2$,

$$\delta_1^2 = (\mathbf{y} - \mathbb{E}\mathbf{y})^\top P_F(\mathbf{y} - \mathbb{E}\mathbf{y}) \leq \sigma^2(7.4|F| + 2.7\ln(4/\eta)),$$

and with probability $1-\eta/2$,

$$\delta_2^2 = |\tilde{a}^\top(\mathbf{y} - \mathbb{E}\mathbf{y})|^2 \leq 2\sigma^2\ln(4/\eta).$$

We obtain the desired result by substituting the above two estimates and simplify. ■

Lemma C.2 *Suppose that Assumption 6.1 is valid. Then we have with probability $1-\eta$, $\forall\beta \in \mathbb{R}^p$ and $a > 0$:*

$$\|X\beta - \mathbb{E}\mathbf{y}\|_2^2 \leq (1+a)\left[\|X\beta - \mathbf{y}\|_2^2 - \|\mathbf{y} - \mathbb{E}\mathbf{y}\|_2^2\right] + (2+a+1/a)\sigma^2[7.4c(\beta) + 4.7\ln(4/\eta)].$$

Proof Note that for each F, with probability $2^{-\mathrm{cl}(F)}\eta$, we obtain from Lemma C.1 that $\forall \mathrm{supp}(\beta) \in F$,

$$\|X\beta - \mathbb{E}\mathbf{y}\|_2^2 \leq (1+a)\left[\|X\beta - \mathbf{y}\|_2^2 - \|\mathbf{y} - \mathbb{E}\mathbf{y}\|_2^2\right] + (2+a+1/a)\sigma^2[7.4(|F| + \mathrm{cl}(F)) + 4.7\ln(4/\eta)].$$

Since $\sum_{F\subset\mathcal{I},F\neq\emptyset} 2^{-\mathrm{cl}(F)}\eta \leq \eta$, the result follows from the union bound. ■

Lemma C.3 *Consider a fixed subset $\bar{F} \subset \mathcal{I}$. Given any $\eta \in (0,1)$, we have with probability $1-\eta$:*

$$|\|X\bar{\beta} - \mathbf{y}\|_2^2 - \|\mathbf{y} - \mathbb{E}\mathbf{y}\|_2^2| \leq \|X\bar{\beta} - \mathbb{E}\mathbf{y}\|_2^2 + 2\sigma\sqrt{2\ln(2/\eta)}\|X\bar{\beta} - \mathbb{E}\mathbf{y}\|_2.$$

Proof Let $\tilde{a} = (X\bar{\beta} - \mathbb{E}\mathbf{y})/\|X\bar{\beta} - \mathbb{E}\mathbf{y}\|_2$, we have

$$\begin{aligned}&|\|X\bar{\beta} - \mathbf{y}\|_2^2 - \|\mathbf{y} - \mathbb{E}\mathbf{y}\|_2^2| \\ =&| - 2(X\bar{\beta} - \mathbb{E}\mathbf{y})^\top(\mathbf{y} - \mathbb{E}\mathbf{y}) + \|X\bar{\beta} - \mathbb{E}\mathbf{y}\|_2^2| \\ \leq& 2\|X\bar{\beta} - \mathbb{E}\mathbf{y}\|_2|\tilde{a}^\top(\mathbf{y} - \mathbb{E}\mathbf{y})| + \|\mathbb{E}\mathbf{y} - X\bar{\beta}\|_2^2.\end{aligned}$$

The desired result now follows from Lemma A.1. ■

Lemma C.4 *Suppose that Assumption 6.1 is valid. Consider any fixed target $\bar{\beta} \in \mathbb{R}^p$. Then with probability exceeding $1-\eta$, for all $\lambda \geq 0, \epsilon \geq 0, \hat{\beta} \in \mathbb{R}^p$ such that: $\hat{Q}(\hat{\beta}) + \lambda c(\hat{\beta}) \leq \hat{Q}(\bar{\beta}) + \lambda c(\bar{\beta}) + \epsilon$, and for all $a > 0$, we have*

$$\begin{aligned}\|X\hat{\beta} - \mathbb{E}\mathbf{y}\|_2^2 \leq& (1+a)[\|X\bar{\beta} - \mathbb{E}\mathbf{y}\|_2^2 + 2\sigma\sqrt{2\ln(6/\eta)}\|X\bar{\beta} - \mathbb{E}\mathbf{y}\|_2] \\ &+ (1+a)\lambda c(\bar{\beta}) + a'c(\hat{\beta}) + b'\ln(6/\eta) + (1+a)\epsilon,\end{aligned}$$

where $a' = 7.4(2+a+a^{-1})\sigma^2 - (1+a)\lambda$ and $b' = 4.7\sigma^2(2+a+a^{-1})$. Moreover, if the coding scheme $c(\cdot)$ is sub-additive, then

$$n\rho_-(c(\hat{\beta}) + c(\bar{\beta}))\|\hat{\beta} - \bar{\beta}\|_2^2 \leq 10\|X\bar{\beta} - \mathbb{E}\mathbf{y}\|_2^2 + 2.5\lambda c(\bar{\beta}) + (37\sigma^2 - 2.5\lambda)c(\hat{\beta}) + 29\sigma^2\ln(6/\eta) + 2.5\epsilon.$$

Proof We obtain from the union bound of Lemma C.2 (with probability $1-\eta/3$) and Lemma C.3 (with probability $1-2\eta/3$) that with probability $1-\eta$:

$$\begin{aligned}&\|X\hat{\beta} - \mathbb{E}\mathbf{y}\|_2^2 \\ \leq&(1+a)\left[\|X\hat{\beta} - \mathbf{y}\|_2^2 - \|\mathbf{y} - \mathbb{E}\mathbf{y}\|_2^2\right] + (2+a+a^{-1})[7.4\sigma^2 c(\hat{\beta}) + 4.7\sigma^2\ln(6/\eta)] \\ \leq&(1+a)\left[\|X\bar{\beta} - \mathbf{y}\|_2^2 - \|\mathbf{y} - \mathbb{E}\mathbf{y}\|_2^2 + \lambda c(\bar{\beta}) + \epsilon\right] + a'c(\hat{\beta}) + b'\ln(6/\eta) \\ \leq&(1+a)[\|X\bar{\beta} - \mathbb{E}\mathbf{y}\|_2^2 + 2\sigma\sqrt{2\ln(6/\eta)}\|X\bar{\beta} - \mathbb{E}\mathbf{y}\|_2] + (1+a)\lambda c(\bar{\beta}) + a'c(\hat{\beta}) \\ &+ b'\ln(6/\eta) + (1+a)\epsilon.\end{aligned}$$

This proves the first claim of the theorem.

The first claim with $a = 1$ implies that

$$\begin{aligned}
&\|X\hat{\beta} - X\bar{\beta}\|_2^2 \leq [\|X\hat{\beta} - \mathbb{E}\mathbf{y}\|_2 + \|X\bar{\beta} - \mathbb{E}\mathbf{y}\|_2]^2 \\
\leq &1.25\|X\hat{\beta} - \mathbb{E}\mathbf{y}\|_2^2 + 5\|X\bar{\beta} - \mathbb{E}\mathbf{y}\|_2^2 \\
\leq &7.5\|X\bar{\beta} - \mathbb{E}\mathbf{y}\|_2^2 + 5\sigma\sqrt{2\ln(6/\eta)}\|X\bar{\beta} - \mathbb{E}\mathbf{y}\|_2 + 2.5\lambda c(\bar{\beta}) + 1.25(29.6\sigma^2 - 2\lambda)c(\hat{\beta}) \\
&\quad + 1.25 \times 18.8\sigma^2 \ln(6/\eta) + 2.5\epsilon \\
\leq &10\|X\bar{\beta} - \mathbb{E}\mathbf{y}\|_2^2 + 2.5\lambda c(\bar{\beta}) + (37\sigma^2 - 2.5\lambda)c(\hat{\beta}) + 29\sigma^2 \ln(6/\eta) + 2.5\epsilon.
\end{aligned}$$

Since $c(\hat{\beta} - \bar{\beta}) \leq c(\hat{\beta}) + c(\bar{\beta})$, we have $\|X\hat{\beta} - X\bar{\beta}\|_2^2 \geq n\rho_-(c(\hat{\beta}) + c(\bar{\beta}))\|\hat{\beta} - \bar{\beta}\|_2^2$. This implies the second claim. ■

Proof of Theorem 6.2

We take $\lambda = 0$ in Lemma C.4, and obtain:

$$\begin{aligned}
\|X\hat{\beta} - \mathbb{E}\mathbf{y}\|_2^2 \leq &(1 + a)[\|X\bar{\beta} - \mathbb{E}\mathbf{y}\|_2^2 + 2\sigma\sqrt{2\ln(6/\eta)}\|X\bar{\beta} - \mathbb{E}\mathbf{y}\|_2] \\
&+ 7.4(2 + a + a^{-1})\sigma^2 c(\hat{\beta}) + 4.7\sigma^2(2 + a + a^{-1})\ln(6/\eta) + (1 + a)\epsilon \\
= &(\|X\bar{\beta} - \mathbb{E}\mathbf{y}\|_2 + \sigma\sqrt{2\ln(6/\eta)})^2 + 14.8\sigma^2 c(\hat{\beta}) + 7.4\sigma^2 \ln(6/\eta) + \epsilon \\
&+ a[(\|X\bar{\beta} - \mathbb{E}\mathbf{y}\|_2 + \sigma\sqrt{2\ln(6/\eta)})^2 + 7.4\sigma^2 c(\hat{\beta}) + 2.7\sigma^2 \ln(6/\eta) + \epsilon] \\
&+ a^{-1}[7.4\sigma^2 c(\hat{\beta}) + 4.7\sigma^2 \ln(6/\eta)].
\end{aligned}$$

Now let $z = \|X\bar{\beta} - \mathbb{E}\mathbf{y}\|_2 + \sigma\sqrt{2\ln(6/\eta)}$, and we choose a to minimize the right hand side as:

$$\begin{aligned}
\|X\hat{\beta} - \mathbb{E}\mathbf{y}\|_2^2 \leq &z^2 + 14.8\sigma^2 c(\hat{\beta}) + 7.4\sigma^2 \ln(6/\eta) + \epsilon \\
&+ 2[z^2 + 7.4\sigma^2 c(\hat{\beta}) + 2.7\sigma^2 \ln(6/\eta) + \epsilon]^{1/2}[7.4\sigma^2 c(\hat{\beta}) + 4.7\sigma^2 \ln(6/\eta)]^{1/2} \\
\leq &[(z^2 + 7.4\sigma^2 c(\hat{\beta}) + 2.7\sigma^2 \ln(6/\eta) + \epsilon)^{1/2} + (7.4\sigma^2 c(\hat{\beta}) + 4.7\sigma^2 \ln(6/\eta))^{1/2}]^2 \\
\leq &[z + 2(7.4\sigma^2 c(\hat{\beta}) + 4.7\sigma^2 \ln(6/\eta) + \epsilon)^{1/2}]^2.
\end{aligned}$$

This proves the first inequality. The second inequality follows directly from Lemma C.4 with $\lambda = 0$.

Proof of Theorem 6.3

The desired bound is a direct consequence of Lemma C.4, by noticing that

$$2\sigma\sqrt{2\ln(6/\eta)}\|X\bar{\beta} - \mathbb{E}\mathbf{y}\|_2 \leq a\|X\bar{\beta} - \mathbb{E}\mathbf{y}\|_2^2 + a^{-1}2\sigma^2 \ln(6/\eta),$$

$a' \leq 0$, and

$$b' + a^{-1}2\sigma^2 \leq (10 + 5a + 7a^{-1})\sigma^2.$$

D Proof of Theorem 6.4 and Theorem 6.5

The following lemma is an adaptation of a similar result in [26] on greedy algorithms for standard sparsity.

Lemma D.1 *Suppose the coding scheme is sub-additive. Consider any $\bar{\beta}$, and a cover of $\bar{\beta}$ by $\mathcal{B}$:*

$$\mathrm{supp}(\bar{\beta}) \subset \bar{F} = \cup_{j=1}^{b} \bar{B}_j \quad (\bar{B}_j \in \mathcal{B}).$$

Let $c(\bar{\beta}, \mathcal{B}) = \sum_{j=1}^{b} c(\bar{B}_j)$. Let $\rho_0 = \max_j \rho_+(\bar{B}_j)$. Then for all F such that $c(\bar{B}_j \cup F) \geq c(F)$,

$$\beta = \arg\min_{\beta' \in \mathbb{R}^p} \|X\beta' - \mathbf{y}\|_2^2 \quad \textit{subject to} \quad \mathrm{supp}(\beta') \subset F,$$

and $\|X\beta - \mathbf{y}\|_2^2 \geq \|X\bar{\beta} - \mathbf{y}\|_2^2$, we have

$$\max_j \phi(\bar{B}_j) \geq \frac{\rho_-(F \cup \bar{F})}{\rho_0 c(\bar{\beta}, \mathcal{B})} [\|X\beta - \mathbf{y}\|_2^2 - \|X\bar{\beta} - \mathbf{y}\|_2^2],$$

where as in (3), we define

$$\phi(B) = \frac{\|P_{B-F}(X\beta - \mathbf{y})\|_2^2}{c(B \cup F) - c(F)}.$$

Proof For all $\ell \in F$, $\|X\beta + \alpha X\mathbf{e}_\ell - \mathbf{y}\|_2^2$ achieves the minimum at $\alpha = 0$ (where $\mathbf{e}_\ell$ is the vector of zeros except for the ℓ-th component, which is one). This implies that

$$\mathbf{x}_\ell^\top (X\beta - \mathbf{y}) = 0$$

for all $\ell \in F$. Therefore we have

$$\begin{aligned} &(X\beta - \mathbf{y})^\top \sum_{\ell \in \bar{F} - F} (\bar{\beta}_\ell - \beta_\ell)\mathbf{x}_\ell \\ =&(X\beta - \mathbf{y})^\top \sum_{\ell \in \bar{F} \cup F} (\bar{\beta}_\ell - \beta_\ell)\mathbf{x}_\ell = (X\beta - \mathbf{y})^\top (X\bar{\beta} - X\beta) \\ =&-\frac{1}{2}\|X(\bar{\beta} - \beta)\|_2^2 + \frac{1}{2}\|X\bar{\beta} - \mathbf{y}\|_2^2 - \frac{1}{2}\|X\beta - \mathbf{y}\|_2^2. \end{aligned}$$

Now, let $\bar{B}'_j \subset \bar{B}_j - F$ be disjoint sets such that $\cup_j \bar{B}'_j = \bar{F} - F$. The above inequality leads to the following derivation $\forall \eta > 0$:

$$\begin{aligned} &-\sum_j \phi(\bar{B}_j)(c(\bar{B}_j \cup F) - c(F)) \\ \leq&\sum_j \left[\left\| X\beta + \eta \sum_{\ell \in \bar{B}'_j} (\bar{\beta}_\ell - \beta_\ell)\mathbf{x}_\ell - \mathbf{y} \right\|_2^2 - \|X\beta - \mathbf{y}\|_2^2 \right] \\ \leq&\eta^2 \sum_{\ell \in \bar{F} - F} (\bar{\beta}_\ell - \beta_\ell)^2 \rho_0 n + 2\eta(X\beta - \mathbf{y})^\top \sum_{\ell \in \bar{F} - F} (\bar{\beta}_\ell - \beta_\ell)\mathbf{x}_\ell \\ \leq&\eta^2 \sum_{\ell \in \bar{F} - F} (\bar{\beta}_\ell - \beta_\ell)^2 \rho_0 n - \eta\|X(\bar{\beta} - \beta)\|_2^2 + \eta\|X\bar{\beta} - \mathbf{y}\|_2^2 - \eta\|X\beta - \mathbf{y}\|_2^2. \end{aligned}$$

Note that we have used the fact that $\|P_{B-F}(X\beta - \mathbf{y})\|_2^2 \geq \|X\beta - \mathbf{y}\|_2^2 - \|X\beta - \mathbf{y} + X\Delta\beta\|_2^2$ for all $\Delta\beta$ such that $\mathrm{supp}(\Delta\beta) \subset B - F$. By optimizing over η, we obtain

$$\begin{aligned}
\max_j \phi(\bar{B}_j) \sum_j c(\bar{B}_j) \geq & \sum_j \phi(\bar{B}_j)(c(\bar{B}_j \cup F) - c(F)) \\
\geq & \frac{[\|X(\bar{\beta} - \beta)\|_2^2 + \|X\beta - \mathbf{y}\|_2^2 - \|X\bar{\beta} - \mathbf{y}\|_2^2]^2}{4\sum_{\ell \in \bar{F} - F}(\bar{\beta}_\ell - \beta_\ell)^2 \rho_0 n} \\
\geq & \frac{4\|X(\bar{\beta} - \beta)\|_2^2[\|X\beta - \mathbf{y}\|_2^2 - \|X\bar{\beta} - \mathbf{y}\|_2^2]}{4\sum_{\ell \in \bar{F} - F}(\bar{\beta}_\ell - \beta_\ell)^2 \rho_0 n} \\
\geq & \frac{\rho_-(F \cup \bar{F})}{\rho_0}[\|X\beta - \mathbf{y}\|_2^2 - \|X\bar{\beta} - \mathbf{y}\|_2^2].
\end{aligned}$$

This leads to the desired bound. ■

Proof of Theorem 6.4

Let

$$\gamma' = \frac{\gamma \rho_-(s + c(\bar{F}))}{\rho_0(\mathcal{B}) c(\bar{\beta}, \mathcal{B})}.$$

By Lemma D.1, we have at any step $k > 0$:

$$\|X\beta^{(k-1)} - \mathbf{y}\|_2^2 - \|X\beta^{(k)} - \mathbf{y}\|_2^2 \geq \gamma'[\|X\beta^{(k-1)} - \mathbf{y}\|_2^2 - \|X\bar{\beta} - \mathbf{y}\|_2^2](c(\beta^{(k)}) - c(\beta^{(k-1)}),$$

which implies that

$$\max[0, \|X\beta^{(k)} - \mathbf{y}\|_2^2 - \|X\bar{\beta} - \mathbf{y}\|_2^2] \leq \max[0, \|X\beta^{(k-1)} - \mathbf{y}\|_2^2 - \|X\bar{\beta} - \mathbf{y}\|_2^2] e^{-\gamma'(c(\beta^{(k)}) - c(\beta^{(k-1)})}.$$

Therefore at stopping, we have

$$\begin{aligned}
& \|X\beta^{(k)} - \mathbf{y}\|_2^2 - \|X\bar{\beta} - \mathbf{y}\|_2^2 \\
\leq & [\|\mathbf{y}\|_2^2 - \|X\bar{\beta} - \mathbf{y}\|_2^2] e^{-\gamma' c(\beta^{(k)})} \\
\leq & [\|\mathbf{y}\|_2^2 - \|X\bar{\beta} - \mathbf{y}\|_2^2] e^{-\gamma' s} \leq \epsilon.
\end{aligned}$$

This proves the theorem.

Proof of Theorem 6.5

For simplicity, let $f_j = \hat{Q}(\bar{\beta}_j)$. For each k, let j_k be the largest j such that

$$\hat{Q}(\beta^{(k)}) \geq f_j + f_j - f_0 + \epsilon.$$

Let $\gamma' = (\gamma \min_j \rho_-(s + c(\bar{\beta}_j)))/(\rho_0(\mathcal{B}) c(\bar{\beta}_0, \mathcal{B}))$.

We prove by contradiction. Suppose that the theorem does not hold, then for all k before stopping, we have $j_k \geq 0$.

For each $k > 0$ before stopping, if $j_k = j_{k-1} = j$, then we have from Lemma D.1 (with $\bar{\beta} = \bar{\beta}_j$)

$$c(\beta^{(k)}) \leq c(\beta^{(k-1)}) + \gamma'^{-1} 2^{-j} \ln \frac{\|X\beta^{(k-1)} - \mathbf{y}\|_2^2 - f_j}{\|X\beta^{(k)} - \mathbf{y}\|_2^2 - f_j}.$$

Therefore for each $j \geq 0$, we have:

$$\sum_{k: j_k = j_{k-1} = j} [c(\beta^{(k)}) - c(\beta^{(k-1)})] \leq \gamma'^{-1} 2^{-j} \ln \frac{2(f_{j+1} - f_0 + \epsilon)}{f_j - f_0 + \epsilon}.$$

Moreover, for each $j \geq 0$, Lemma D.1 (with $\bar{\beta} = \bar{\beta}_j$) implies that

$$\sum_{k: j_k = j, j_{k-1} > j} [c(\beta^{(k)}) - c(\beta^{(k-1)})] \leq \gamma'^{-1} 2^{-j}.$$

Therefore we have

$$\sum_{k: j_k = j} [c(\beta^{(k)}) - c(\beta^{(k-1)})] \leq \gamma'^{-1} 2^{-j} \left[1.7 + \ln \frac{f_{j+1} - f_0 + \epsilon}{f_j - f_0 + \epsilon}\right].$$

Now by summing over $j \geq 0$, we have

$$c(\beta^{(k)}) \leq 3.4\gamma'^{-1} + \gamma'^{-1} \sum_{j=0}^{\infty} 2^{-j} \ln \frac{f_{j+1} - f_0 + \epsilon}{f_j - f_0 + \epsilon} \leq s.$$

This is a contradiction because we know at stopping, we should have $c(\beta^{(k)}) > s$.

E Proof of Corollary 6.1

Given s', we consider $f_j = \min_{\ell \geq j} \hat{Q}(\bar{\beta}(s'/2^\ell))$. We may assume that f_0 is achieved with $\ell_0 = 0$. Note that by Lemma C.3, we have with probability $1 - 2^{-j-1}\eta$:

$$\begin{aligned} |\hat{Q}(\bar{\beta}(s'/2^j)) - \|\mathbf{y} - \mathbb{E}\mathbf{y}\|_2^2| \leq & 2\|X\bar{\beta}(s'/2^j) - \mathbb{E}\mathbf{y}\|_2^2 + 2\sigma^2[j + 1 + \ln(2/\eta)] \\ \leq & 2an2^{qj}/s'^q + 2\sigma^2[j + 1 + \ln(2/\eta)]. \end{aligned}$$

This means the above inequality holds for all j with probability $1 - \eta$. Now, by taking $\epsilon = 2an/s'^q + 2\sigma^2[\ln(2/\eta) + 1]$ in Theorem 6.5, we obtain

$$\begin{aligned} \sum_{j=0}^{\infty} 2^{-j} \ln \frac{f_{j+1} - f_0 + \epsilon}{f_j - f_0 + \epsilon} \leq & \sum_{j=\ell_0}^{\infty} 2^{-j} \ln(1 + (f_{j+1} - f_0)/\epsilon) \\ \leq & \sum_{j=\ell_0}^{\infty} 2^{-j} \ln(2 + 2(j + 2^{q(j+1)})) \\ \leq & \sum_{j=\ell_0}^{\infty} 2^{-j} (\ln 2 + 1 + j + q(j+1)\ln 2) \leq 2 + 4(1 + q\ln 2), \end{aligned}$$

where we have used the simple inequality $\ln(\alpha + \beta) \le \alpha + \ln(\beta)$ when $\alpha, \beta \ge 1$. Therefore,

$$\begin{aligned} s \ge & \frac{\rho_0(\mathcal{B})s'}{\gamma \min_{u \le s'} \rho_-(s + c(\bar{\beta}(u)))} (10 + 3q) \\ \ge & \frac{\rho_0(\mathcal{B})s'}{\gamma \min_{u \le s'} \rho_-(s + c(\bar{\beta}(u)))} \left[3.4 + \sum_{j=0}^{\infty} 2^{-j} \ln \frac{f_{j+1} - f_0 + \epsilon}{f_j - f_0 + \epsilon} \right]. \end{aligned}$$

This means that Theorem 6.5 can be applied to obtain the desired bound.