

Adaptive Forward-Backward Greedy Algorithm for Learning Sparse Representations

Tong Zhang *
Statistics Department
Rutgers University, NJ
tzhang@stat.rutgers.edu

Abstract

Given a large number of basis functions that can be potentially more than the number of samples, we consider the problem of learning a sparse target function that can be expressed as a linear combination of a small number of these basis functions. We are interested in two closely related themes

- feature selection, or identifying the basis functions with non-zero coefficients;
- estimation accuracy, or reconstructing the target function from noisy observations.

Two heuristics that are widely used in practice are forward and backward greedy algorithms. First, we show that neither idea is adequate. Second, we propose a novel combination that is based on the forward greedy algorithm but takes backward steps adaptively whenever beneficial. For least squares regression, we develop strong theoretical results for the new procedure showing that it can effectively solve this problem under some assumptions. Experimental results support our theory.

1 Introduction

Consider a set of input vectors $\mathbf{x}_1, \dots, \mathbf{x}_n \in R^d$, with corresponding desired output variables y_1, \dots, y_n . The task of supervised learning is to estimate the functional relationship $y \approx f(\mathbf{x})$ between the input \mathbf{x} and the output variable y from the training examples $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$. The quality of prediction is often measured through a loss function $\phi(f(\mathbf{x}), y)$. In this paper, we consider linear prediction model $f(\mathbf{x}) = \mathbf{w}^T \mathbf{x}$. A commonly used estimation method is empirical risk minimization

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w} \in R^d} \sum_{i=1}^n \phi(\mathbf{w}^T \mathbf{x}_i, y_i). \quad (1)$$

Note that in this paper, we are mainly interested in the least squares problem where $\phi(\mathbf{w}^T \mathbf{x}_i, y_i) = (\mathbf{w}^T \mathbf{x}_i - y_i)^2$.

In modern machine learning applications, one is typically interested in the scenario that $d \gg n$. That is, there are many more features than the number of samples. In this case, a direct application

*This work is partially supported by NSF grant DMS-0706805. A shorter version of the paper appears in NIPS 2008.

of (1) is inadequate because the solution of $\hat{\mathbf{w}}$ may not be unique (which is often referred to as *ill-posed* in the numerical computation literature). Statistically, the solution $\hat{\mathbf{w}}$ overfits the data. The standard remedy for this problem is to impose a regularization condition of \mathbf{w} to obtain a *well-posed* problem. For computational reasons, one often employs a convex regularization condition which leads to a convex optimization problem of the following form:

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w} \in R^d} \sum_{i=1}^n \phi(\mathbf{w}^T \mathbf{x}_i, y_i) + \lambda g(\mathbf{w}), \quad (2)$$

where $\lambda > 0$ is a tuning parameter, and $g(\mathbf{w})$ is a regularization condition, such as $g(\mathbf{w}) = \|\mathbf{w}\|_p^p$.

One view of this additional regularization condition is that it constrains the target function space, which we assume can be approximated by some $\bar{\mathbf{w}}$ with small L_p $\|\bar{\mathbf{w}}\|_p$. An important target constraint is sparsity, which corresponds to the (non-convex) L_0 regularization, where we let $\|\bar{\mathbf{w}}\|_0 = |\{j : \bar{w}_j \neq 0\}| = k$. If we know the sparsity parameter k , a good learning method is L_0 regularization:

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w} \in R^d} \frac{1}{n} \sum_{i=1}^n \phi(\mathbf{w}^T \mathbf{x}_i, y_i) \quad (3)$$

subject to $\|\mathbf{w}\|_0 \leq k$.

If k is not known, then one may regard k as a tuning parameter, which can be selected through cross-validation. Sparse learning is an essential topic in machine learning, which has attracted considerable interests recently. Generally speaking, one is interested in two closely related themes:

- feature selection, or identifying the basis functions with non-zero coefficients;
- estimation accuracy, or reconstructing the target function from noisy observations.

If we can solve the first problem, that is, if we can perform feature selection well, then we can also solve the second problem. This is because after feature selection, we only need to perform empirical risk minimization (1) with the selected features. However, it is possible to obtain good prediction accuracy without solving the feature selection problem.

This paper focuses on the situation that approximate feature selection is possible. Under this scenario, we obtain results both on feature selection accuracy and on parameter estimation accuracy. Our main assumption is the *restricted isometry condition* (RIC) of [5], which we shall also refer to as the *sparse eigenvalue condition* in this paper. This condition says that any small number (at the order of the desired sparsity level) of features are not highly correlated. In fact, if a small number of features are correlated, then it is impossible to achieve accurate feature selection because a sparse target may be represented using more than one set of sparse features. Therefore the effectiveness of any feature selection algorithm requires such a condition.

It can be shown that under RIC, the solution of the L_0 regularization problem in (3) solves the feature selection problem, and achieves good prediction accuracy if the target function can be approximated by a sparse $\bar{\mathbf{w}}$. However, a fundamental difficulty with this method is the computational cost, because the number of subsets of $\{1, \dots, d\}$ of cardinality k (corresponding to the nonzero components of \mathbf{w}) is exponential in k . There are no efficient algorithms to solve (3) in the general case.

Due to the computational difficulty, in practice, there are several standard methods for learning sparse representations by solving approximations of (3). Their effectiveness has been recently analyzed under various assumptions.

- L_1 -regularization (Lasso): the idea is to replace the L_0 regularization in (3) by L_1 regularization:

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w} \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n \phi(\mathbf{w}^T \mathbf{x}_i, y_i),$$

subject to $\|\mathbf{w}\|_1 \leq s$,

or equivalently, solving (2) with $p = 1$. This is the closest convex approximation to (3). It is known that L_1 regularization often leads to sparse solutions. Its performance has been studied recently. For example, if the target is truly sparse, then it was shown in [14, 22] that under some restricted conditions referred to as *irrepresentable conditions*, L_1 regularization solves the feature selection problem. However, such conditions are much stronger than RIC considered here. The prediction performance of L_1 regularization has been considered in [11, 2, 3]. Performance bounds can also be obtained when the target function is only approximately sparse (e.g., [21, 4]). Despite its popularity, there are several problems with L_1 regularization: first, the sparsity (L_0 complexity) is only implicitly controlled through L_1 approximation, and good feature selection property requires relatively strong assumptions; second, in order to obtain very sparse solution, one has to use a large regularization parameter λ in (2) that leads to suboptimal prediction accuracy because the L_1 penalty not only shrinks irrelevant features to zero, but also shrinks relevant features to zero. A sub-optimal remedy is to threshold the resulting coefficients as suggested in [21] and to use two stage procedures. However, this approach requires additional tuning parameters, making the resulting procedures more complex and less robust.

- Forward greedy algorithm, which we will describe in details in Section 2. The method has been widely used by practitioners. For least squares regression, this method is referred to as *matching pursuit* [13] in the signal processing community (also see [10, 1]). In machine learning, the method is often known as boosting. The algorithm is analyzed in [17, 7] without considering stochastic noise, while its feature selection performance with stochastic noise has recently been studied in [20]. It was shown that the irrepresentable condition of [22] for L_1 regularization is also necessary for the greedy algorithm to effectively select features.
- Backward greedy algorithm, which we will describe in details in Section 2. Although this method is widely used by practitioners, there isn't much theoretical analysis in the literature when $n \ll d$. The reason will be discussed in Section 2. When $n \gg d$, backward greedy may be successful under some assumptions [6].

We shall point out if we are only interested in prediction performance instead of feature selection, then the problem of learning sparse representation is also related to learning a sparse target function under many irrelevant features, which has long been studied in the online learning literature. In particular, exponentiated gradient descent methods such as Winnow are also effective [12]. However, this class of methods do not lead to sparse solutions.

Although there have been considerable interests in learning sparse representations, and multiple algorithms have been proposed to solve the problem, satisfactory theoretical understanding (mainly

for L_1 regularization) has only appeared very recently. In this paper, we are particularly interested in greedy algorithms because they have been widely used but the effectiveness has not been well analyzed. Moreover, they do not suffer from some shortcomings of L_1 regularization which we have pointed out earlier.

As we shall explain later, neither the standard forward greedy idea nor the standard backward greedy idea is adequate for our purpose. However, the flaws of these methods can be fixed by a simple combination of the two ideas. This leads to a novel adaptive forward-backward greedy algorithm which we present in Section 3. The general idea works for all loss functions. For least squares loss, we obtain strong theoretical results showing that the method can solve the feature selection problem under moderate conditions.

For clarity, this paper only considers fixed design. To simplify notations in our description, we will replace the optimization problem in (1) and (3) with a more general formulation. Instead of working with n input data vectors $\mathbf{x}_i \in R^d$, we work with d feature vectors $\mathbf{f}_j \in R^n$ ($j = 1, \dots, d$), and $\mathbf{y} \in R^n$. Each \mathbf{f}_j corresponds to the j -th feature component of \mathbf{x}_i for $i = 1, \dots, n$. That is, $\mathbf{f}_{j,i} = \mathbf{x}_{i,j}$. Using this notation, we can generally rewrite (3) with the problem of the following form:

$$\begin{aligned} \hat{\mathbf{w}} &= \arg \min_{\mathbf{w} \in R^d} R(\mathbf{w}), \\ &\text{subject to } \|\mathbf{w}\|_0 \leq k, \end{aligned}$$

where weight $\mathbf{w} = [\mathbf{w}_1, \dots, \mathbf{w}_d] \in R^d$.

In the following, we also let $\mathbf{e}_j \in R^d$ be the vector of zeros, except for the j -component which is one. Throughout the paper, we consider only the least squares loss

$$R(\mathbf{w}) = \frac{1}{n} \left\| \mathbf{y} - \sum_{j=1}^d \mathbf{w}_j \mathbf{f}_j \right\|_2^2, \quad (4)$$

where $\mathbf{y} = [y_1, \dots, y_n] \in R^n$.

For convenience, we also introduce the following notations.

Definition 1.1 Define $\text{supp}(\mathbf{w}) = \{j : \mathbf{w}_j \neq 0\}$ as the set of nonzero coefficients of a vector $\mathbf{w} = [\mathbf{w}_1, \dots, \mathbf{w}_d] \in R^d$. For a weight vector $\mathbf{w} \in R^d$, we define mapping $f : R^d \rightarrow R^n$ as:

$$f(\mathbf{w}) = \sum_{j=1}^d \mathbf{w}_j \mathbf{f}_j.$$

Given $\mathbf{f} \in R^d$ and $F \subset \{1, \dots, d\}$, let

$$\hat{\mathbf{w}}(F, \mathbf{f}) = \min_{\mathbf{w} \in R^d} \frac{1}{n} \|\mathbf{f}(\mathbf{w}) - \mathbf{f}\|_2^2 \text{ subject to } \text{supp}(\mathbf{w}) \subset F,$$

and let $\hat{\mathbf{w}}(F) = \hat{\mathbf{w}}(F, \mathbf{y})$ be the solution of the least squares problem using feature set F .

Note that from the definition, $f(\hat{\mathbf{w}}(F, \mathbf{f}))$ is simply the projection of \mathbf{f} to the subspace spanned by $\{\mathbf{f}_j : j \in F\}$.

2 Forward and Backward Greedy Algorithms

Forward greedy algorithms have been widely used in applications. It can be used to improve an arbitrary prediction method or select relevant features. In the former context, it is often referred to as boosting, and in the latter context, forward feature selection. Although a number of variations exist, they all share the basic form of greedily picking an additional feature at every step to aggressively reduce the squared error. The intention is to make most significant progress at each step in order to achieve sparsity. In this regard, the method can be considered as an approximation algorithm for solving (3). An example algorithm is presented in Figure 1. This particular algorithm performs a full optimization using the selected basis function at each step, and is often referred to as orthogonal matching pursuit or OMP. This per-step optimization is important in our analysis.

```

Input:  $\mathbf{f}_1, \dots, \mathbf{f}_d, \mathbf{y} \in R^n$  and  $\epsilon > 0$ 
Output:  $F^{(k)}$  and  $\mathbf{w}^{(k)}$ 
let  $F^{(0)} = \emptyset$  and  $\mathbf{w}^{(0)} = 0$ 
for  $k = 1, 2, \dots$ 
  let  $i^{(k)} = \arg \min_i \min_\alpha R(\mathbf{w}^{(k-1)} + \alpha \mathbf{e}_i)$ 
  let  $F^{(k)} = \{i^{(k)}\} \cup F^{(k-1)}$ 
  let  $\mathbf{w}^{(k)} = \hat{\mathbf{w}}(F^{(k)})$ 
  if  $(R(\mathbf{w}^{(k-1)}) - R(\mathbf{w}^{(k)})) \leq \epsilon$  break
end

```

Figure 1: Forward Greedy Algorithm

A major flaw of this method is that it can never correct mistakes made in earlier steps. As an illustration, we consider the situation plotted in Figure 2 with least squares regression. In the figure, \mathbf{y} can be expressed as a linear combination of \mathbf{f}_1 and \mathbf{f}_2 but \mathbf{f}_3 is closer to \mathbf{y} . Therefore using the forward greedy algorithm, we will find \mathbf{f}_3 first, then \mathbf{f}_1 and \mathbf{f}_2 . At this point, we have already found all good features as \mathbf{y} can be expressed by \mathbf{f}_1 and \mathbf{f}_2 , but we are not able to remove \mathbf{f}_3 selected in the first step.

The above argument implies that forward greedy method is inadequate for feature selection. The method only works when small subsets of the basis functions $\{\mathbf{f}_j\}$ are near orthogonal. For

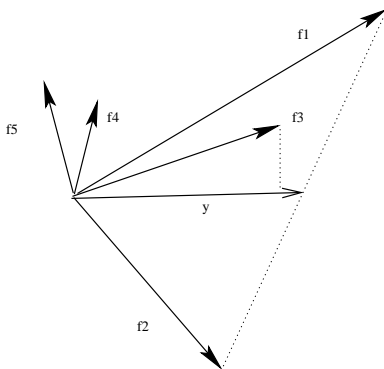


Figure 2: Failure of Forward Greedy Algorithm

example, see [17, 7] for analysis of greedy algorithm under such assumptions without statistical noise¹. Its feature selection performance with stochastic noise has been recently studied in [20]. In general, Figure 2 shows that even when the variables are not completely correlated (which is the case we consider in this paper), forward greedy algorithm will make errors that are not corrected later on. In fact, results in [17, 20] showed that in addition to the sparse eigenvalue condition, a stronger irrepresentable condition (also see [22]) is necessary for forward greedy algorithm to be successful.

For feature selection, the main problem of forward greedy algorithm is the lack of ability to correct errors made in earlier steps. In order to remedy the problem, the so-called backward greedy algorithm has been widely used by practitioners. The idea is to train a full model with all the features, and greedily remove one feature (with the smallest increase of squared error) at a time. The basic algorithm can be described in Figure 3.

Although at the first sight, backward greedy method appears to be a reasonable idea that addresses the problem of forward greedy algorithm, it is computationally very costly because it starts with a full model with all features. Moreover, there are no theoretical results showing that this procedure is effective. In fact, under our setting, the method may only work when $d \ll n$ (see, for example, [6]), which is not the case we are interested in. In the case $d \gg n$, during the first step, $\mathbf{w}^{(d)}$ can immediately overfit the data with perfect prediction. Moreover, removing any feature does not help: that is, $\mathbf{w}^{(d-1)}$ still completely overfits the data no matter which feature (either relevant or irrelevant) is removed. Therefore the method will completely fail when $d \gg n$, which explains why there is no theoretical result for this method.

It should be pointed out that the fundamental problem of backward greedy is that we cannot start with an overfitted model. To this end, one may replace (*) in Figure 3 by a solution procedure that does not overfit, for example, via L_1 regularization. Backward greedy combined with L_1 regularization is potentially beneficial because we can more effectively control the sparsity of the resulting L_1 regularization solution. However, such a procedure will be computationally costly, and its benefit is unknown. In this paper, we propose an alternative solution by combining the strength of both forward and backward greedy methods while avoiding their shortcomings.

```

Input:  $\mathbf{f}_1, \dots, \mathbf{f}_d, \mathbf{y} \in R^n$ 
Output:  $F^{(k)}$  and  $\mathbf{w}^{(k)}$ 
let  $F^{(d)} = \{1, \dots, d\}$ 
for  $k = d, d - 1, \dots$ 
  let  $\mathbf{w}^{(k)} = \hat{\mathbf{w}}(F^{(k)})$ 
  let  $j^{(k)} = \arg \min_{j \in F^{(k)}} R(\hat{\mathbf{w}}(F^{(k)} - \{j\}))$   (*)
  let  $F^{(k-1)} = F^{(k)} - \{j^{(k)}\}$ 
end

```

Figure 3: Backward Greedy Algorithm

¹Although the title in [7] claimed a treatment of noise, their definition of noise is not random, and thus different from ours. In our terminology, their noise only means that the target function is approximately sparse, which we also handle. It is different from stochastic noise considered in this paper.

3 Adaptive Forward-Backward Greedy Algorithm

As we have pointed out earlier, the main strength of forward greedy algorithm is that it always works with a sparse solution explicitly, and thus computationally efficient. Moreover, it does not significantly overfit the data due to the explicit sparsity. However, a major problem is its inability to correct any error made by the algorithm. On the other hand, backward greedy steps can potentially correct such an error, but need to start with a good model that does not completely overfit the data — it can only correct errors with a small amount of overfitting. Therefore a combination of the two can solve the fundamental flaws of both methods. However, a key design issue is how to implement a backward greedy strategy that is provably effective. Some heuristics exist in the literature, although without any effectiveness proof. For example, the standard heuristics, described in [9] and implemented in SAS, includes another threshold ϵ' in addition to ϵ : a feature is deleted if the squared error increase by performing the deletion is no more than ϵ' . Unfortunately we cannot provide an effectiveness proof for this heuristics: if the threshold ϵ' is too small, then it cannot delete any spurious features introduced in the forward steps; if it is too large, then one cannot make progress because good features are also deleted. In practice it can be hard to pick a good ϵ' , and even the best choice may be ineffective.

This paper takes a more principled approach, where we specifically design a forward-backward greedy procedure with *adaptive* backward steps that are carried out automatically. The procedure has provably good performance and fixes the drawbacks of forward greedy algorithm illustrated in Figure 2. There are two main considerations in our approach.

- We want to take reasonably aggressive backward steps to remove any errors caused by earlier forward steps, and to avoid maintaining a large number of basis functions.
- We want to take backward step *adaptively* and make sure that any backward greedy step does not erase the gain made in the forward steps. This ensures that we are always making progress.

Our algorithm, which we refer to as *FoBa*, is listed in Figure 4. It is designed to balance the above two aspects. Note that we only take a backward step when the squared error increase (d^-) is no more than half of the squared error decrease in the earlier corresponding forward step (d^+). This implies that if we take ℓ forward steps, then no matter how many backward steps are performed, the squared error is decreased by at least an amount of $\ell\epsilon/2$. It follows that if $R(\mathbf{w}) \geq 0$ for all $\mathbf{w} \in R^d$, then the algorithm terminates after no more than $2R(0)/\epsilon$ steps. This means that the procedure is computationally efficient.

Proposition 3.1 *When the FoBa procedure terminates in Figure 4, the total number of forward steps is no more than $1 + 2R(0)/\epsilon$. Moreover, the total number of backward steps is no more than the total number of forward steps.*

Note that the claim of Proposition 3.1 (as well as the later theoretical analysis) still holds if we employ a more aggressive backward strategy as follows: set $d^- = d^+ = 0$ at the beginning of the backward step, and update the quantities as $d^- = d^- + [R(\mathbf{w}^{(k)}) - \mathbf{w}_{j^{(k)}}^{(k)} \mathbf{e}_{j^{(k)}} - R(\mathbf{w}^{(k)})]$ and $d^+ = d^+ + \delta^{(k)}$. That is, d^- and d^+ are cumulative changes of squared error.

Now, consider an application of FoBa to the example in Figure 2. Again, in the first three forward steps, we will be able to pick \mathbf{f}_3 , followed by \mathbf{f}_1 and \mathbf{f}_2 . After the third step, since we are

```

Input:  $\mathbf{f}_1, \dots, \mathbf{f}_d, \mathbf{y} \in R^n$  and  $\epsilon > 0$ 
Output:  $F^{(k)}$  and  $\mathbf{w}^{(k)}$ 
let  $\nu = 0.5$  (it can also be set to another number in  $(0, 1)$ )
let  $F^{(0)} = \emptyset$  and  $\mathbf{w}^{(0)} = 0$ 
let  $k = 0$ 
while (true)
  let  $k = k + 1$ 
  // forward step
  let  $i^{(k)} = \arg \min_i \min_{\alpha} R(\mathbf{w}^{(k-1)} + \alpha \mathbf{e}_i)$ 
  let  $F^{(k)} = \{i^{(k)}\} \cup F^{(k-1)}$ 
  let  $\mathbf{w}^{(k)} = \hat{\mathbf{w}}(F^{(k)})$ 
  let  $\delta^{(k)} = R(\mathbf{w}^{(k-1)}) - R(\mathbf{w}^{(k)})$ 
  if ( $\delta^{(k)} \leq \epsilon$ )
     $k = k - 1$ 
    break
  end
  // backward step
  while (true)
    let  $j^{(k)} = \arg \min_{j \in F^{(k)}} R(\mathbf{w}^{(k)} - \mathbf{w}_j^{(k)} \mathbf{e}_j)$ 
    let  $d^- = [R(\mathbf{w}^{(k)} - \mathbf{w}_{j^{(k)}}^{(k)} \mathbf{e}_{j^{(k)}}) - R(\mathbf{w}^{(k)})]$ 
    let  $d^+ = \delta^{(k)}$ 
    if ( $d^- > \nu d^+$ )
      break
    end
    let  $k = k - 1$ 
    let  $F^{(k)} = F^{(k+1)} - \{j^{(k+1)}\}$ 
    let  $\mathbf{w}^{(k)} = \hat{\mathbf{w}}(F^{(k)})$ 
  end
end

```

Figure 4: FoBa: Adaptive Forward-Backward Greedy Algorithm

able to express \mathbf{y} using \mathbf{f}_1 and \mathbf{f}_2 only, by removing \mathbf{f}_3 in the backward step, we do not increase the squared error. Therefore at this stage, we are able to successfully remove the incorrect basis \mathbf{f}_3 while keeping the good features \mathbf{f}_1 and \mathbf{f}_2 . This simple illustration demonstrates the effectiveness of FoBa.

In the following, we will formally characterize this intuitive example, and prove results for the effectiveness of FoBa for feature selection as well as parameter estimation, under the condition that the target is either truly sparse or approximately sparse. Since the situation in Figure 2 is covered by our analysis, one cannot derive a similar result for the forward greedy algorithm. That is, the condition of our results do not exclude forward steps from making errors. Therefore it is essential to include backward steps in our theoretical analysis.

We introduce the following definition, which characterizes how linearly independent small subsets of $\{\mathbf{f}_j\}$ of size k are. For $k \ll n$, the number $\rho(k)$ defined below can be bounded away from zero

even when $d \gg n$. For example, for random basis functions \mathbf{f}_j , we may take $\ln d = O(n/k)$ and still have $\rho(k)$ to be bounded away from zero. This quantity is the smallest eigenvalue of the $k \times k$ diagonal blocks of the $d \times d$ design matrix $[\mathbf{f}_i^T \mathbf{f}_j]_{i,j=1,\dots,d}$, and has appeared in recent analysis of L_1 regularization methods such as in [3, 18, 21], etc. It was first introduced in [5], and was referred to as the restricted isometry condition. In this paper, we shall also call it *sparse eigenvalue condition*. This condition is the least restrictive condition when compared to other conditions in the literature such as the irrepresentable condition [22] or the mutual coherence condition [7] (also see discussions in [2, 18, 21]).

Definition 3.1 Define for all $\bar{F} \subset \{1, \dots, d\}$:

$$\begin{aligned}\rho(\bar{F}) &= \inf \left\{ \frac{1}{n} \|f(\mathbf{w})\|_2^2 / \|\mathbf{w}\|_2^2 : \text{supp}(\mathbf{w}) \subset \bar{F} \right\}, \\ \lambda(\bar{F}) &= \sup \left\{ \frac{1}{n} \|f(\mathbf{w})\|_2^2 / \|\mathbf{w}\|_2^2 : \text{supp}(\mathbf{w}) \subset \bar{F} \right\},\end{aligned}$$

and for all $1 \leq k \leq d$:

$$\rho(k) = \inf_{|\bar{F}| \leq k} \rho(\bar{F}), \quad \lambda(k) = \sup_{|\bar{F}| \leq k} \lambda(\bar{F}).$$

Assumption 3.1 Assume that the basis functions are normalized such that $\frac{1}{n} \|\mathbf{f}_j\|_2^2 = 1$ for all $j = 1, \dots, d$, and assume that $\{y_i\}_{i=1,\dots,n}$ are independent (but not necessarily identically distributed) sub-Gaussians: there exists $\sigma \geq 0$ such that $\forall i$ and $\forall t \in \mathbb{R}$,

$$\mathbf{E}_{y_i} e^{t(y_i - \mathbf{E}y_i)} \leq e^{\sigma^2 t^2 / 2}.$$

Both Gaussian and bounded random variables are sub-Gaussian using the above definition. For example, we have the following well-known Hoeffding's inequality.

Proposition 3.2 If a random variable $\xi \in [a, b]$, then $\mathbf{E}_\xi e^{t(\xi - \mathbf{E}\xi)} \leq e^{(b-a)^2 t^2 / 8}$. If a random variable is Gaussian: $\xi \sim N(0, \sigma^2)$, then $\mathbf{E}_\xi e^{t\xi} \leq e^{\sigma^2 t^2 / 2}$.

The following theorem is stated with an explicit ϵ for convenience. In real applications, one can always run the algorithm with a smaller ϵ and use cross-validation to determine the optimal stopping point.

Theorem 3.1 Consider the FoBa algorithm in Figure 4, where Assumption 3.1 holds. Consider any approximate target vector $\bar{\mathbf{w}} \in \mathbb{R}^d$ with $\bar{F} = \text{supp}(\bar{\mathbf{w}})$, and define

$$\Delta = \frac{1}{n} \|\mathbf{E}\mathbf{y} - f(\bar{\mathbf{w}})\|_2^2, \quad \bar{k} = |\bar{F}| + \lfloor 4\Delta/\epsilon \rfloor,$$

where $\epsilon > 0$ is the stopping criterion in Figure 4. let $s \leq d$ be an integer which satisfies one of the following conditions:

$$8\bar{k} \leq s\rho(s)^2 \quad \text{or} \quad 2\bar{k} \ln(2\bar{k} + 2n^{-1} \|f(\bar{\mathbf{w}})\|_2^2 / \epsilon + 2\Delta/\epsilon) + \bar{k} + 1 \leq s\rho(s) \quad \text{or} \quad s = d.$$

Define

$$k(\epsilon) = |\{j \in \bar{F} : |\bar{\mathbf{w}}_j|^2 \leq 12\epsilon/\rho(s)^2\}| + 2\lfloor 4\Delta/\epsilon \rfloor.$$

Assume that for some $\eta \in (0, 1/3)$, we have $\epsilon \geq 256\rho(s)^{-2}\sigma^2 \ln(2d/\eta)/n$. Then with probability larger than $1 - 3\eta$:

- The FoBa algorithm terminates after at most

$$\frac{8}{n\epsilon} \|f(\bar{\mathbf{w}})\|_2^2 + 2\bar{k} + \frac{4\bar{k}}{n\epsilon} \sigma^2 \left[1 + \sqrt{20 \ln(1/\eta)}\right]^2 + \frac{256k(\epsilon)\lambda(\bar{k})^2}{9\rho(s)^3}$$

forward iterations.

- When the algorithm terminates, we have:

$$\begin{aligned} |\bar{F} - F^{(k)}| &\leq 2k(\epsilon), \quad |F^{(k)} - \bar{F}| \leq 4\Delta/\epsilon + \frac{256k(\epsilon)\lambda(\bar{k})}{9\rho(s)^3}, \\ \sum_{j \in \bar{F} - F^{(k)}} |\bar{\mathbf{w}}_j|^2 &\leq 16k(\epsilon)\epsilon/\rho(s)^2 + 2\Delta/\rho(\bar{k}), \\ \|\mathbf{w}^{(k)} - \bar{\mathbf{w}}\|_2 &\leq \frac{8\sqrt{2k(\epsilon)\lambda(\bar{k})}\epsilon}{3\rho(s)^{3/2}} + \sigma\sqrt{\bar{k}/(n\rho(\bar{k}))} \left[1 + \sqrt{20 \ln(1/\eta)}\right] + \sqrt{\Delta/\rho(\bar{k})}. \end{aligned}$$

Note that the choice of s given in the theorem is an upper bound of $|\bar{F} \cup F^{(k)}|$; it can be replaced with other upper bounds without changing the other claims. The proof of the theorem depends on an analysis of the stopping criterion of the FoBa procedure. In particular, we do not attempt to show that any particular backward step is effective. Instead, the proof shows the following three properties of the FoBa procedure: if a particular backward step deletes a good feature, then FoBa does not stop; FoBa does not stop if many good features in \bar{F} are missing from $F^{(k)}$; the backward step will start to delete (either bad or good) features if $F^{(k)}$ contains too many features not in \bar{F} . By combining these properties, we can deduce that once the FoBa procedure stops, the feature set $F^{(k)}$ approximately equals \bar{F} . Since our analysis does not require the effectiveness of a specific backward step, our results do not hold for a simpler procedure that performs a sequence of forward steps, followed by a sequence of backward steps. Such a method is unreliable because good features can be deleted in the backward steps. Our adaptive forward-backward approach does not suffer from this problem.

Since the form of Theorem 3.1 is complicated, it is useful to examine simplified bounds. For simplicity, we can assume that the condition $8\bar{k} \leq s\rho(s)^2$ holds, with $\rho(s)$ bounded away from zero. We will absorb it into the $O(\cdot)$ notation.

The statement

$$|\bar{F} - F^{(k)}| = O(k(\epsilon)), \quad |F^{(k)} - \bar{F}| = O(k(\epsilon))$$

implies that the estimated feature set differs from \bar{F} by no more than $O(k(\epsilon))$ elements. This difference includes below noise level small coefficients $|\bar{\mathbf{w}}_j|^2 = O(\epsilon)$ that cannot be differentiated from zero, and missing features caused by the approximation error Δ . Therefore $F^{(k)} \approx \bar{F}$ when $k(\epsilon) \ll \bar{k}$. That is, if Δ is small and only a small number of weights $\bar{\mathbf{w}}_j$ ($j \in \bar{F}$) are below $O(\epsilon)$. In particular, if $k(\epsilon) = 0$, then we can recover the true feature set \bar{F} with large probability. The statement on $\sum_{j \in \bar{F} - F^{(k)}} |\bar{\mathbf{w}}_j|^2$ also implies that any $j \in \bar{F}$ such that $|\bar{\mathbf{w}}_j|^2 > 17(k(\epsilon) + 1)\epsilon/\rho(s)^2$ is included in $F^{(k)}$. That is, the final feature set $F^{(k)}$ will contain all large coefficients of $\bar{\mathbf{w}}_j$.

In summary, the above discussion shows that in the general case, even if we cannot reliably identify the correct feature set \bar{F} , the difference is small as long as $k(\epsilon)$ is small. Moreover, all large coefficients can be reliably identified.

The theorem also implies a parameter estimation error bound of the following form:

$$\|\mathbf{w}^{(k)} - \bar{\mathbf{w}}\|_2 = O\left(\sigma\sqrt{k/n} + \|\mathbf{E}\mathbf{y} - f(\bar{\mathbf{w}})\|_2/\sqrt{n} + \sqrt{k'(\epsilon)\epsilon}\right), \quad (5)$$

where $k'(\epsilon) = |\{j \in \bar{F} : |\bar{\mathbf{w}}_j|^2 \leq 12\epsilon/\rho(s)^2\}|$. The first term is the parametric rate that is independent of d ; the second term is caused by the approximation error, which cannot be removed; the third term is caused by the noise. Since we need to take ϵ at the order $O(\sigma\sqrt{\ln d/n})$, the third term is of the order $O(\sigma\sqrt{k'(\epsilon)\ln d/n})$. However, this term is small when $k'(\epsilon)$ is small: this happens when only a small number of $j \in \bar{F}$ have small coefficients $|\bar{\mathbf{w}}_j|^2 \leq O(\epsilon)$. Intuitively, this term indicates that better feature selection improves parameter estimation in the following sense: only features that cannot be reliably identified induce a $\ln d$ dependency. Therefore parameter estimation accuracy can be improved through better feature identification, which reflects our discussion in the introduction. However, in the worst case, even when all coefficients are small (and thus reliable feature selection is impossible), the parameter estimation bound in (5) is still meaningful (with $\epsilon = O(\sigma\sqrt{\ln d/n})$):

$$\|\mathbf{w}^{(k)} - \bar{\mathbf{w}}\|_2 = O\left(\|\mathbf{E}\mathbf{y} - f(\bar{\mathbf{w}})\|_2/\sqrt{n} + \sqrt{k \ln d/n}\right).$$

That is, we can still obtain a relatively small parameter estimation error even if we cannot reliably select any features. On the other hand, the bound in (5) shows that the parameter estimation accuracy can be improved through better feature selection, when $k'(\epsilon)$ is small.

A useful application of the theorem is when the target $\mathbf{E}\mathbf{y}$ is truly sparse: that is $\bar{\mathbf{w}}^T \mathbf{x}_i = \mathbf{E}y_i$ for $i = 1, \dots, n$. In this case, if we further assume that $k(\epsilon) = 0$, then $\bar{F} = F^{(k)}$ when the algorithm terminates. That is, we can identify the correct set of features with large probability. This particular problem has drawn significant interests in recent years, such as [22, 21].

Corollary 3.1 *Consider the FoBa algorithm in Figure 4, where Assumption 3.1 holds. Consider a target $\bar{\mathbf{w}} \in R^d$ such that $\mathbf{E}\mathbf{y} = f(\bar{\mathbf{w}})$. Let $s \leq d$ be an integer such that $8\bar{k} \leq s\rho(s)^2$, and assume that $|\bar{\mathbf{w}}_j|^2 > 12\epsilon/\rho(s)^2$ for all $j \in \bar{F}$. Assume that for some $\eta \in (0, 1/3)$, we have $\epsilon \geq 256\rho(s)^{-2}\sigma^2 \ln(2d/\eta)/n$. Then with probability larger than $1 - 3\eta$: when the algorithm terminates, we have: $\bar{F} = F^{(k)}$ and $\|\mathbf{w}^{(k)} - \bar{\mathbf{w}}^*\|_2 \leq \sigma\sqrt{k/(n\rho(\bar{k}))} \left[1 + \sqrt{20 \ln(1/\eta)}\right]$.*

The corollary says that we can identify the correct set of features \bar{F} as long as the coefficients $\bar{\mathbf{w}}_j$ ($j \in \bar{F}$) are larger than the noise level $O(\sigma\sqrt{\ln d/n})$. Such a requirement is quite natural, and occurs in other work on the effectiveness of feature selection [22, 21]. In fact, if any nonzero weight is below the noise level, then no algorithm can distinguish it from zero with large probability. That is, it is impossible to reliably perform feature selection due to the noise. Therefore FoBa is near optimal in term of its ability to perform reliable feature selection, except for the constant hiding in $O(\cdot)$ (as well as its dependency on $\rho(s)$). Moreover, Theorem 3.1 shows that if only a small number of coefficients $\bar{\mathbf{w}}_j$ are below the noise level, then FoBa can still select part of the features reliably, with good parameter estimation accuracy.

The result can be applied as long as eigenvalues of small $s \times s$ diagonal blocks of the design matrix $[\mathbf{f}_i^T \mathbf{f}_j]_{i,j=1,\dots,d}$ are bounded away from zero (that is, the sparse eigenvalue condition holds). This is the situation under which the forward greedy step can make mistakes, but such mistakes can be corrected using FoBa. Because the conditions of the corollary do not prevent forward steps

from making errors, the example described in Figure 2 indicates that it is not possible to prove a similar result for the forward greedy algorithm. In fact, it was shown in [20] that the stronger irrepresentable condition of [22] is necessary for the forward greedy algorithm to be effective.

4 FoBa versus Lasso

Lasso can successfully select features under *irrepresentable conditions* of [22] (also see [17] which considered the noiseless case). It was shown that such a condition is necessary for feature selection using Lasso, when zero-threshold is used. It is known that the sparse eigenvalue condition considered in this paper is significantly weaker [18, 15, 2].

Although under the sparse eigenvalue condition, it is not possible to reliably select features using Lasso and zero-threshold, it was pointed out in [21] that it is possible to reliably select features using an appropriately chosen non-zero threshold (that is, a post-processing step is used to remove features with coefficients smaller than a certain threshold). There are two problems for this approach. First, this requires tuning two parameters: one is the L_1 regularization parameter, and the other non-zero threshold. Second, even if one can tune the threshold parameter successfully, the result in [21] requires that the condition $\min_{j \in \bar{F}} |\bar{\mathbf{w}}_j|^2 \geq c\bar{k}\sigma^2 \ln(d/\eta)/n$ for some $c > 0$ under the sparse eigenvalue condition considered here (although the \bar{k} -dependence can be removed under stronger assumptions). This is due to the inclusion of L_1 regularization which introduces a “bias”. In comparison, Theorem 3.1 only requires $\min_{j \in \text{supp}(\bar{\mathbf{w}})} |\bar{\mathbf{w}}_j|^2 \geq c\sigma^2 \ln(d/\eta)/n$ for some $c > 0$. The difference can be significant when \bar{k} is large.

There is no counterpart of Theorem 3.1 for L_1 regularization under the sparse eigenvalue conditions. The closest analogy is a parameter estimation bound for a two-stage procedure investigated in [21], but the result there was weaker because it requires the stronger mutual coherence assumption of [7]. The bound of Theorem 3.1 is thus superior to all existing bounds for Lasso. The underlying reason is due to the better feature selection ability of FoBa. As we have pointed out, better feature selection implies better parameter estimation (under the sparse eigenvalue condition).

Finally, we shall point out that the forward-backward greedy procedure is closely related to the path-following algorithm for solving Lasso, such as the LARS algorithm for solving Lasso in [8], where one starts with a very large (infinity) regularization parameter, which is gradually decreased. Similar to FoBa, LARS also has forward and backward steps. However, unlike FoBa, which tracks the insertion and deletion with unbiased least squares error, LARS tracks the path through L_1 penalized least squares error, by gradually decreasing the regularization parameter. Initially, to obtain a very sparse set of features, one has to set a very large L_1 regularization parameter, which causes a significant bias. The added bias implies that LARS deviates significantly from subset selection at least initially. When the algorithm progresses, the regularization parameter is reduced, and thus the extra L_1 bias becomes smaller. Since the theory of L_1 regularization requires a regularization parameter larger than the noise level, the resulting bias is not negligible. Similar to FoBa, some mistakes made in earlier steps can be potentially removed later on.

It follows from the above discussion that FoBa is related to the path-following view of L_1 regularization. However, unlike Lasso, FoBa does not introduce a bias. Instead, it generates the path by directly keeping track of the objective function. Therefore FoBa is closer to subset selection than L_1 regularization, especially when a highly sparse solution is desired (as far as training error is concerned). This claim is conformed by our experiments.

5 Experiments

We compare FoBa to forward-greedy and L_1 -regularization on artificial and real data. They show that in practice, FoBa is closer to subset selection than the other two approaches, in the sense that FoBa achieves smaller training error given any sparsity level. In order to compare with Lasso, we use the LARS [8] package in R, which generates a path of actions for adding and deleting features, along the L_1 solution path. For example, a path of $\{1, 3, 5, -3, \dots\}$ means that in the first three steps, feature 1, 3, 5 are added; and the next step removes feature 3.

Using such a solution path, we can compare Lasso to Forward-greedy and FoBa under the same framework. Similar to the Lasso path, FoBa also generates a path with both addition and deletion operations, while forward-greedy algorithm only adds features without deletion.

Our experiments compare the performance of the three algorithms using the corresponding feature addition/deletion paths. We are interested in features selected by the three algorithms at any sparsity level k , where k is the desired number of features presented in the final solution. Given a path, we can keep an active feature set by adding or deleting features along the path. For example, for path $\{1, 3, 5, -3\}$, we have two potential active feature sets of size $k = 2$: $\{1, 3\}$ (after two steps) and $\{1, 5\}$ (after four steps). We then define the k best features as the active feature set of size k with the smallest least squares error because this is the best approximation to subset selection (along the path generated by the algorithm). For the forward-greedy algorithm, this is just the first k features in the path. For FoBa, it is the last time when the active set contains k features, because the squared error is always reduced in later stages. For Lasso, it is also like to be the last time when the active set contains k features — this is because it corresponds to the smallest regularization parameter (thus smallest bias). However, to be safe, for Lasso, we compute the least squares solutions for all active feature sets of size k , and then select the one with the smallest squared error.

From the above discussion, we do not have to set ϵ explicitly in the FoBa procedure. Instead, we just generate a solution path which is five times as long as the maximum desired sparsity k , and then generate the best k features for any sparsity level using the above described procedure.

5.1 Simulation Data

Since for real data, we do not know the true feature set \bar{F} , simulation is needed to compare feature selection performance. We generate $n = 100$ data points of dimension $d = 500$. The target vector $\bar{\mathbf{w}}$ is truly sparse with $\bar{k} = 5$ nonzero coefficients generated uniformly from 0 to 10. The noise level is $\sigma^2 = 0.1$. The basis functions \mathbf{f}_j are randomly generated with moderate correlation: that is, some basis functions are correlated to the basis functions spanning the true target. Note that if there is no correlation (i.e., \mathbf{f}_j are independent random vectors), then all three methods work well (this is the well-known case considered in the compressed sensing literature). If the correlation among \mathbf{f}_j is very strong, then it is not surprising that all three methods will fail. Therefore in this experiment, we generate moderate correlation so that the performance of the three methods can be differentiated. Such moderate correlation does not violate the sparse eigenvalue condition in our analysis, but violates the more restrictive conditions for forward-greedy method and Lasso.

Table 1 shows the performance of the three methods, where we repeat the experiments 50 times, and report the average \pm standard-deviation. We use the three methods to select five best features, using the procedure described above. We report three metrics. Training error is the squared error of the least squares solution with the selected five features. Parameter estimation error is the

| | FoBa | Forward-greedy | L_1 |
|------------------------------|------------------|------------------|-----------------|
| least squares training error | 0.093 ± 0.02 | 0.16 ± 0.089 | 0.25 ± 0.14 |
| parameter estimation error | 0.057 ± 0.2 | 0.52 ± 0.82 | 1.1 ± 1 |
| feature selection error | 0.76 ± 0.98 | 1.8 ± 1.1 | 3.2 ± 0.77 |

Table 1: Performance comparison on simulation data at sparsity level $k = 5$

2-norm of the estimated parameter (with the five features) minus the true parameter. Feature selection error is the number of incorrectly selected features. It is clear from the table that for this data, FoBa achieves significantly smaller training error than the other two methods, which implies that it is closest to subset selection. Moreover, the parameter estimation performance and feature selection performance are also better. Note that the signal to noise ratio is relatively small in this example, which ensures that feature selection can be performed reliably. If feature selection becomes unreliable, then L_1 regularization may achieve better performance because it is inherently more stable.

5.2 Real Data

Instead of listing results for many datasets, we consider two data sets that reflect typical behaviors of the algorithms. A careful analysis of the two datasets leads to better insights than a list of performance numbers for many data. The experiments show that FoBa does what it is designed to do well: that is, it gives a better approximation to subset selection than either forward-greedy or L_1 regularization. However, as well shall see, better sparsity does not always lead to better generalization on real data. This is because sparsity alone is not necessarily always the best complexity measure for real problems.

5.2.1 Boston Housing data

The first dataset we consider is the *Boston Housing* data, which is the housing data for 506 census tracts of Boston from the 1970 census, available from the *UCI Machine Learning Database Repository*: <http://archive.ics.uci.edu/ml/>. Each census tract is a data-point, with 13 features (we add a constant offset one as the 14th feature), and the desired output is the housing price. In the experiment, we randomly partition the data into 50 training plus 456 test points. We perform the experiments 50 times, and for each sparsity level from 1 to 10, we report the average training and test squared error. The results are plotted in Figure 5. From the results, we can see that FoBa achieves better training error for any given sparsity, which is consistent with the theory and the design goal of FoBa. Moreover, it achieves better test accuracy with small sparsity level (corresponding to a more sparse solution). With large sparsity level (corresponding to a less sparse solution), the test error increase more quickly with FoBa. This is because it searches a larger space by more aggressively mimicking subset selection, which makes it more prone to overfitting. However, at the best sparsity level of 3, FoBa achieves significantly better test error. Moreover, we can observe with small sparsity level (a more sparse solution), L_1 regularization performs poorly, due to the bias caused by using a large L_1 -penalty.

For completeness, we also compare FoBa to the backward-greedy algorithm and the classical heuristic forward-backward greedy algorithm as implemented in SAS (see its description at the beginning of Section 3). We still use the Boston Housing data, but plot the results separately, in

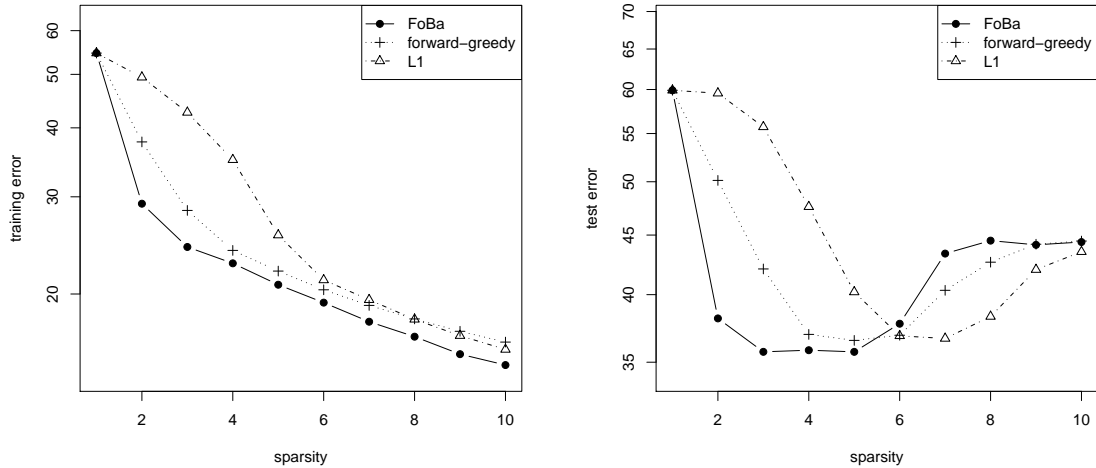


Figure 5: Performance of the algorithms on *Boston Housing* data Left: average training squared error versus sparsity; Right: average test squared error versus sparsity

order to avoid cluttering. As we have pointed out, there is no theory for the SAS version of forward-backward greedy algorithm. It is difficult to select an appropriate backward threshold ϵ' : a too small value leads to few backward steps, and a too large value leads to overly aggressive deletion, and the procedure terminates very early. In this experiment, we pick a value of 10, because it is a reasonably large quantity that does not lead to an extremely quick termination of the procedure. The performance of the algorithms are reported in Figure 6. From the results, we can see that backward greedy algorithm performs reasonably well on this problem. Note that for this data, $d \ll n$, which is the scenario that backward does not start with a completely overfitted full model. Still, it is inferior to FoBa at small sparsity level, which means that some degree of overfitting still occurs. Note that backward-greedy algorithm cannot be applied in our simulation data experiment, because $d \gg n$ which causes immediate overfitting. From the graph, we also see that FoBa is more effective than the SAS implementation of forward-backward greedy algorithm. The latter does not perform significant better than the forward-greedy algorithm with our choice of ϵ' . Unfortunately, using a larger backward threshold ϵ' will lead to an undesirable early termination of the algorithm. This intuition is why the provably effective adaptive backward strategy introduced in this paper is superior.

5.2.2 Ionosphere data

The second dataset we consider is the *Ionosphere* data, also available from the *UCI Machine Learning Database Repository*. It contains 351 data points, with 34 features (we add a constant offset one as the 35th feature), and the desired output is binary valued $\{0,1\}$. Although this is a classification problem, we treat it as regression. In the experiment, we randomly partition the data into 50 training plus 301 test points. We perform the experiments 50 times, and for each sparsity from 1 to 10, we report the average training and test squared error. The results are plotted in Figure 7.

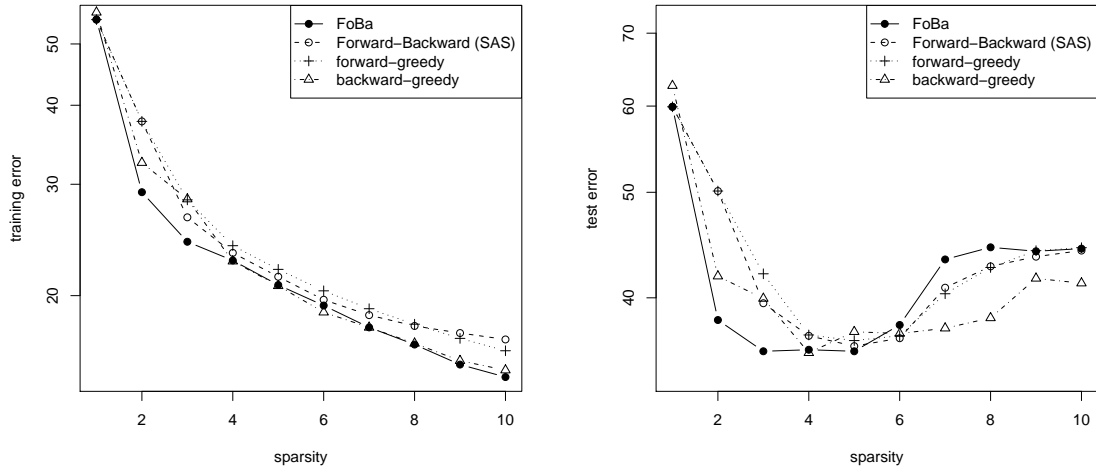


Figure 6: Performance of greedy algorithms on *Boston Housing* data. Left: average training squared error versus sparsity; Right: average test squared error versus sparsity

From the results, we can see that FoBa again achieves better training error for any given sparsity, which is consistent with the theory and the design goal of FoBa. However, it does not achieve better test accuracy. This suggests that sparsity alone is not the correct complexity measure for this data. Indeed by examining the results closely, we observe that the coefficients of Lasso solution tend to be much smaller (due to the added L_1 constraints) than those from FoBa or the forward-greedy algorithm (which do not favor small coefficients in their designs). From Figure 7, we can see that even with smaller coefficients, Lasso achieves similar training error at small sparsity level. This means that Lasso effectively searches a smaller space, and thus more stable and less prone to overfitting. Therefore, for this dataset, the added prior knowledge of small coefficients (in addition to sparsity) in L_1 regularization gives it an edge over the greedy approaches, which do not take the size of coefficients into consideration. For simplicity, we do not include a comparison with the backward greedy method.

6 Discussion

This paper investigates the problem of learning sparse representations using greedy algorithms. We showed that neither forward greedy nor backward greedy algorithms are adequate by themselves. However, through a novel combination of the two ideas, we proved that an adaptive forward-back greedy algorithm, referred to as FoBa, can effectively solve the problem under reasonable conditions.

FoBa is designed to be a better approximation to subset selection. In fact, backward step naturally appears in solving L_0 regularized optimization problems. Consider the penalization version of (3):

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w} \in R^d} \left[\frac{1}{n} \sum_{i=1}^n \phi(\mathbf{w}^T \mathbf{x}_i, y_i) + \lambda \|\mathbf{w}\|_0 \right].$$

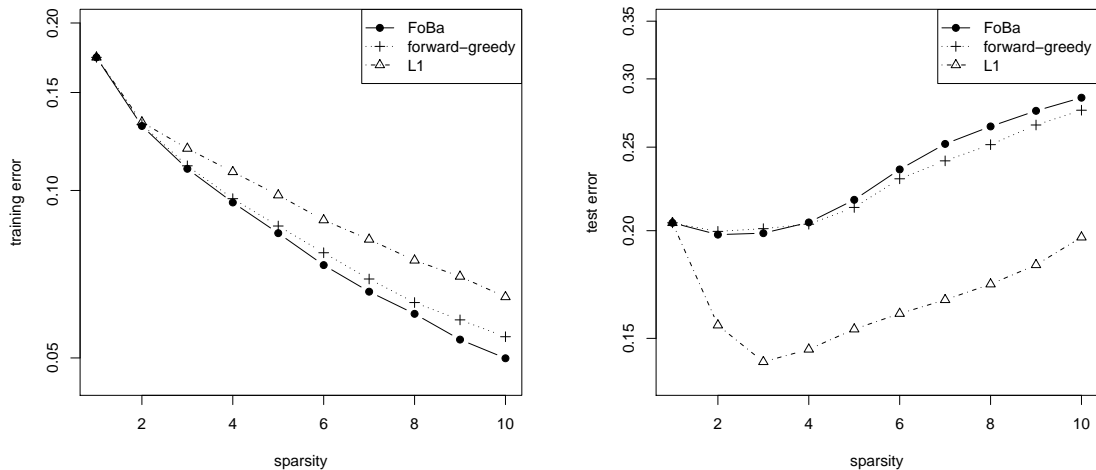


Figure 7: Performance of the algorithms on *Ionosphere* data. Left: average training squared error versus sparsity; Right: average test squared error versus sparsity

If we remove one nonzero component (backward step) from a tentative solution \mathbf{w} , then the term $\lambda\|\mathbf{w}\|_0$ is decreased by λ ; in the mean time, if $\frac{1}{n}\sum_{i=1}^n\phi(\mathbf{w}^T\mathbf{x}_i, y_i)$ is increased by an amount less than λ , then the overall regularized objective function is decreased. In this case, the backward step should be taken because it decreases the overall regularized objective value. However, a main problem of using a fixed λ is when λ is small (which is required to achieve good statistical performance), the backward step is ineffective because it can only occur after a significant number of forward steps, which have already overfitted the data. This problem is similar to that of the standard backward greedy algorithm. The FoBa algorithm, which chooses the λ threshold adaptively, fixes the problem. If we pick $\nu \rightarrow 1$, then the algorithm can be regarded as an approximate path-following scheme (similar to the LARS method for Lasso) with gradually decreasing λ . Moreover, the algorithm is also theoretically justified. Under the sparse eigenvalue condition, we obtained strong performance bounds for FoBa for feature selection and parameter estimation. In fact, to the author’s knowledge, in terms of sparsity, the bounds developed for FoBa in this paper are superior to earlier results in the literature for other methods.

Our experiments also showed that FoBa achieves its design goal: that is, it gives smaller training error than either forward-greedy or L_1 regularization for any given level of sparsity. Therefore the experiments are consistent with our theory. In simulation, better sparsity leads to better parameter estimation accuracy. In real data, better sparsity helps on some data (e.g. *Boston Housing*) but not always. This implies that sparsity may not be the best complexity measure for any given problem. In particular, as shown in the *Ionosphere* experiment, the prior knowledge of using small coefficients, which is encoded in the L_1 regularization formulation, can lead to better generalization performance (when such a prior is appropriate for the problem). The so called “bias” of L_1 regularization, which leads to suboptimal sparsity on the training data, can be advantageous on the test data.

The experiments also indicate that in order to design learning methods with the best possible generalization performance, one should consider factors beyond sparsity. In fact, for some data,

features cannot be reliably selected due to the high correlation among some key variables or the small signal to noise ratio. In this scenario, any algorithm (including FoBa) that mimics L_0 -regularization is unstable (that is, it often selects incorrect features), which hurts the prediction performance. Additional prior knowledge, such as small coefficients in certain norm, can be very important. We believe such a prior knowledge can be incorporated into FoBa by adding a suitable regularization into the objective function. In particular, it should be possible to design a FoBa like path-following algorithm that simultaneously achieves sparsity and small coefficients. This important extension of FoBa will be studied in a future work.

References

- [1] A.R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, 1993.
- [2] Peter Bickel, Yaacov Ritov, and Alexandre Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 2008. to appear.
- [3] Florentina Bunea, Alexandre Tsybakov, and Marten H. Wegkamp. Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics*, 1:169–194, 2007.
- [4] Florentina Bunea, Alexandre B. Tsybakov, and Marten H. Wegkamp. Aggregation for Gaussian regression. *Annals of Statistics*, 35:1674–1697, 2007.
- [5] Emmanuel J. Candes and Terence Tao. Decoding by linear programming. *IEEE Trans. on Information Theory*, 51:4203–4215, 2005.
- [6] Christophe Couvreur and Yoram Bresler. On the optimality of the backward greedy algorithm for the subset selection problem. *SIAM J. Matrix Anal. Appl.*, 21(3):797–808, 2000.
- [7] David L. Donoho, Michael Elad, and Vladimir N. Temlyakov. Stable recovery of sparse over-complete representations in the presence of noise. *IEEE Trans. Info. Theory*, 52(1):6–18, 2006.
- [8] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *Annals of Statistics*, 32(2):407–499, 2004.
- [9] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2001.
- [10] L.K. Jones. A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training. *Ann. Statist.*, 20(1):608–613, 1992.
- [11] Vladimir Koltchinskii. Sparsity in penalized empirical risk minimization. *Annales de l’Institut Henri Poincaré*, 2008.
- [12] N. Littlestone. Learning quickly when irrelevant attributes abound: a new linear-threshold algorithm. *Machine Learning*, 2:285–318, 1988.
- [13] S. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.

- [14] Nicolai Meinshausen and Peter Buhlmann. High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics*, 34:1436–1462, 2006.
- [15] Nicolai Meinshausen and Bin Yu. Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics*, 2008. to appear.
- [16] B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM J. Comput.*, 24(2):227–234, 1995.
- [17] Joel A. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Trans. Info. Theory*, 50(10):2231–2242, 2004.
- [18] Cun-Hui Zhang and Jian Huang. Model-selection consistency of the Lasso in high-dimensional linear regression. Technical report, Rutgers University, 2006.
- [19] Tong Zhang. Approximation bounds for some sparse kernel regression algorithms. *Neural Computation*, 14:3013–3042, 2002.
- [20] Tong Zhang. On the consistency of feature selection using greedy least squares regression. *Journal of Machine Learning Research*, 2008. to appear.
- [21] Tong Zhang. Some sharp performance bounds for least squares regression with L_1 regularization. *The Annals of Statistics*, 2009. to appear.
- [22] Peng Zhao and Bin Yu. On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7:2541–2567, 2006.

A Auxiliary Lemmas

The following lemma provides a bound on the squared error reduction of one forward greedy step. Some ingredients of the proof have appeared in [16, 19].

Lemma A.1 *Let Assumption 3.1 hold. Consider any $\mathbf{w}' \in R^d$ and $\mathbf{f}' = f(\mathbf{w}')$. Consider $\bar{F} = \text{supp}(\mathbf{w}')$ and $F \subset \{1, \dots, d\}$. Let $s = |\bar{F} \cup F|$ and $s' = |\bar{F} - F| > 0$. Let $\mathbf{w} = \hat{\mathbf{w}}(F)$ and $\mathbf{f} = f(\mathbf{w})$. If for some $\beta > -1$,*

$$R(\mathbf{w}) - R(\mathbf{w}') \geq \frac{\beta}{n} \|\mathbf{f}' - \mathbf{f}\|_2^2,$$

then

$$\inf_{\alpha \in R, j \in \bar{F} - F} R(\mathbf{w} + \alpha \mathbf{e}_j) \leq R(\mathbf{w}) - \frac{\rho(s)(1 + \beta)}{4s'} \left(\frac{1}{n} \|\mathbf{f} - \mathbf{f}'\|_2^2 + R(\mathbf{w}) - R(\mathbf{w}') \right).$$

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

$$\begin{aligned}
& (\mathbf{f} - \mathbf{y})^T \sum_{j \in \bar{F}-F} (\mathbf{w}'_j - \mathbf{w}_j) \mathbf{f}_j \\
&= (\mathbf{f} - \mathbf{y})^T \sum_{j \in \bar{F} \cup F} (\mathbf{w}'_j - \mathbf{w}_j) \mathbf{f}_j \\
&= (\mathbf{f} - \mathbf{y})^T (\mathbf{f}' - \mathbf{f}) \\
&= -\frac{1}{2}(\mathbf{f}' - \mathbf{f})^2 + \frac{1}{2}(\mathbf{f}' - \mathbf{y})^2 - \frac{1}{2}(\mathbf{f} - \mathbf{y})^2 \\
&= -\frac{1}{2}(\mathbf{f}' - \mathbf{f})^2 + \frac{n}{2}R(\mathbf{w}') - \frac{n}{2}R(\mathbf{w}).
\end{aligned}$$

This leads to the following derivation $\forall \eta > 0$:

$$\begin{aligned}
& s' \inf_{j \in \bar{F}-F} R(\mathbf{w} + \eta(\mathbf{w}'_j - \mathbf{w}_j) \mathbf{e}_j) \\
&\leq \sum_{j \in \bar{F}-F} R(\mathbf{w} + \eta(\mathbf{w}'_j - \mathbf{w}_j) \mathbf{e}_j) \\
&= s'R(\mathbf{w}) + \frac{\eta^2}{n} \sum_{j \in \bar{F}-F} (\mathbf{w}'_j - \mathbf{w}_j)^2 \|\mathbf{f}_j\|_2^2 + \frac{2\eta}{n} (\mathbf{f} - \mathbf{y})^T \sum_{j \in \bar{F}-F} (\mathbf{w}'_j - \mathbf{w}_j) \mathbf{f}_j \\
&= s'R(\mathbf{w}) + \eta^2 \sum_{j \in \bar{F}-F} (\mathbf{w}'_j - \mathbf{w}_j)^2 - \eta \left[\frac{1}{n}(\mathbf{f}' - \mathbf{f})^2 + R(\mathbf{w}) - R(\mathbf{w}') \right].
\end{aligned}$$

Note that in the last equation, we have used $\|\mathbf{f}_j\|_2^2 = n$ in Assumption 3.1. By optimizing over η , we obtain

$$\begin{aligned}
& s' \inf_{j \in \bar{F}-F} R(\mathbf{w} + \eta(\mathbf{w}'_j - \mathbf{w}_j) \mathbf{e}_j) \\
&\leq s'R(\mathbf{w}) - \frac{[\frac{1}{n}(\mathbf{f}' - \mathbf{f})^2 + R(\mathbf{w}) - R(\mathbf{w}')]^2}{4 \sum_{j \in \bar{F}} (\mathbf{w}'_j - \mathbf{w}_j)^2} \\
&\leq s'R(\mathbf{w}) - \frac{\rho(s)(1 + \beta)}{4} \left[\frac{1}{n}(\mathbf{f}' - \mathbf{f})^2 + R(\mathbf{w}) - R(\mathbf{w}') \right].
\end{aligned}$$

This leads to the lemma. ■

The following lemma provides a bound on the squared error increase of one backward greedy step.

Lemma A.2 *Let Assumption 3.1 hold. Consider $\mathbf{w}' \in \mathbb{R}^d$ and $\mathbf{f}' = f(\mathbf{w}')$. Consider $\bar{F} = \text{supp}(\mathbf{w}')$ and $F \subset \{1, \dots, d\}$. Let $s = |F \cup \bar{F}|$ and $s'' = |F - \bar{F}|$. Let $\mathbf{w} = \hat{\mathbf{w}}(F)$ and $\mathbf{f} = f(\mathbf{w})$. Then*

$$\inf_{j \in F} R(\mathbf{w} - \mathbf{w}_j \mathbf{e}_j) \leq R(\mathbf{w}) + \frac{1}{s''} \|\mathbf{w} - \mathbf{w}'\|_2^2 \leq R(\mathbf{w}) + \frac{1}{\rho(s)s''n} \|\mathbf{f} - \mathbf{f}'\|_2^2.$$

Proof For all $j \in F$, we have $R(\mathbf{w} + \alpha \mathbf{e}_j)$ achieves minimum at $\alpha = 0$. This implies that $(\mathbf{f} - \mathbf{y})^T \mathbf{f}_j = 0$ for $j \in F$. We thus have

$$\begin{aligned}
s'' \inf_{j \in F} nR(\mathbf{w} - \mathbf{w}_j \mathbf{e}_j) &\leq \sum_{j \in F - \bar{F}} nR(\mathbf{w} - \mathbf{w}_j \mathbf{e}_j) \\
&= \sum_{j \in F - \bar{F}} \|\mathbf{f} - \mathbf{y} - \mathbf{w}_j \mathbf{f}_j\|_2^2 \\
&= |F - \bar{F}| \|\mathbf{f} - \mathbf{y}\|_2^2 + \sum_{j \in F - \bar{F}} \mathbf{w}_j^2 \|\mathbf{f}_j\|_2^2 \\
&= s'' \|\mathbf{f} - \mathbf{y}\|_2^2 + n \sum_{j \in F - \bar{F}} \mathbf{w}_j^2 \\
&\leq s'' \|\mathbf{f} - \mathbf{y}\|_2^2 + \rho(s)^{-1} \|\mathbf{f} - \mathbf{f}'\|_2^2,
\end{aligned}$$

where $\|\mathbf{f}_j\|_2^2 = n$ in Assumption 3.1 is used. This leads to the lemma. \blacksquare

The following lemma can be used to obtain a more refined bound for the squared error reduction of the forward greedy step, when Lemma A.1 is not applicable.

Lemma A.3 For all $\mathbf{f}, \mathbf{f}', \mathbf{y} \in R^n$, we have

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

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

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

Lemma A.4 Consider n independent random variables ξ_1, \dots, ξ_n such that $\mathbf{E}e^{t(\xi_i - \mathbf{E}\xi_i)} \leq e^{\sigma_i^2 t^2 / 2}$ for all t and i . Consider $g_{i,j}$ for $i = 1, \dots, n$ and $j = 1, \dots, m$, we have for all $\eta \in (0, 1)$, with probability larger than $1 - \eta$:

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

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

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

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

This implies that

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

This implies the lemma. ■

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

Lemma A.5 *Let Assumption 3.1 hold. Consider any fixed $\bar{F} \subset \{1, \dots, d\}$ with $|\bar{F}| = \bar{k}$, and let $\hat{\mathbf{f}} = f(\hat{\mathbf{w}}(\bar{F}))$. Let $\bar{\mathbf{w}} = \hat{\mathbf{w}}(\bar{F}, \mathbf{E}\mathbf{y})$. For all $\eta \in (0, 1)$, with probability larger than $1 - \eta$, we have*

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

Proof For simplicity, let $G \in R^{n \times \bar{k}}$ be the matrix with columns \mathbf{f}_j for $j \in \bar{F}$. Let $\hat{\mathbf{w}}' \in R^{\bar{k}}$ and $\bar{\mathbf{w}}' \in R^{\bar{k}}$ be the restrictions of $\hat{\mathbf{w}}(\bar{F}) \in R^d$ and $\bar{\mathbf{w}} \in R^d$ to \bar{F} respectively. By definition, $\hat{\mathbf{w}}' = (G^T G)^{-1} G^T \mathbf{y}$ and $\bar{\mathbf{w}}' = (G^T G)^{-1} G^T \mathbf{E}\mathbf{y}$. It follows that

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

Therefore for $j = 1, \dots, \bar{k}$:

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

Lemma A.4 implies that with probability larger than $1 - \eta$, for all $j = 1, \dots, \bar{k}$:

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

According to Definition 3.1, $\rho(\bar{k})n$ is no larger than the smallest eigenvalue of $G^T G$. Therefore the desired inequality follows from the estimate

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

■

The following lemma gives a bound on the 2-norm of the difference between the estimated parameter $\hat{\mathbf{w}}(\bar{F})$ and the true parameter $\bar{\mathbf{w}}$ when the set of features \bar{F} are known in advance.

Lemma A.6 *Let Assumption 3.1 hold. Consider any fixed $\bar{F} \subset \{1, \dots, d\}$ with $|\bar{F}| = \bar{k}$, and let $\hat{\mathbf{f}} = f(\hat{\mathbf{w}}(\bar{F}))$. Let $\bar{\mathbf{w}} = \hat{\mathbf{w}}(\bar{F}, \mathbf{E}\mathbf{y})$ and $\bar{\mathbf{f}} = f(\bar{\mathbf{w}})$. For all $\eta \in (0, 1)$, with probability larger than $1 - \eta$, we have*

$$\sqrt{\rho(\bar{k})n} \|\hat{\mathbf{w}}(\bar{F}) - \bar{\mathbf{w}}\|_2 \leq \|\hat{\mathbf{f}} - \bar{\mathbf{f}}\|_2 \leq \sigma \sqrt{\bar{k}} + \sigma \sqrt{20\bar{k} \ln(1/\eta)}.$$

Proof The proof is based on a similar bound in [21]. We include it here for completeness. Similar to the proof of Lemma A.5, we have

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

Therefore $\hat{\mathbf{f}} - \bar{\mathbf{f}} = P(\mathbf{y} - \mathbf{E}\mathbf{y})$, where $P = G(G^T G)^{-1} G^T$ is the projection matrix onto the subspace spanned by basis functions \mathbf{f}_j ($j \in \bar{F}$). It follows that there are vectors $\mathbf{z}_i \in R^{\bar{k}}$ ($i = 1, \dots, n$) such that $\sum_{i=1}^n \|\mathbf{z}_i\|_2^2 = \bar{k}$ and

$$\|\hat{\mathbf{f}} - \bar{\mathbf{f}}\|_2^2 = \left\| \sum_{i=1}^n \xi_i \mathbf{z}_i \right\|_2^2, \tag{6}$$

where $\xi_i = y_i - \mathbf{E}y_i$. Note that we can simply let \mathbf{z}_i be the i -th column vector of the matrix $(G^T G)^{-0.5} G^T$.

For each i , let ξ'_i be an identically distributed and independent copy of ξ_i , and $h(\cdot)$ is any real-valued function such that $h(\xi_i) - h(\xi'_i) \leq |\xi_i| + |\xi'_i|$. Then

$$\begin{aligned}
\mathbf{E}_{\xi_i} e^{t(h(\xi_i) - \mathbf{E}_{\xi'_i} h(\xi'_i))} &= 1 + \sum_{k=2}^{\infty} \frac{t^k}{k!} \mathbf{E}_{\xi_i} (h(\xi_i) - \mathbf{E}_{\xi'_i} h(\xi'_i))^k \\
&\leq 1 + \sum_{k=2}^{\infty} \frac{t^k}{k!} \mathbf{E}_{\xi_i} (|\xi_i| + \mathbf{E}_{\xi'_i} |\xi'_i|)^k \leq 1 + \sum_{k=2}^{\infty} \frac{(2t)^k}{k!} \mathbf{E}_{\xi_i} |\xi_i|^k \\
&= 1 + \sum_{k=1}^{\infty} \left[\frac{1}{(2k)!} \mathbf{E}_{\xi_i} |2t\xi_i|^{2k} + \frac{1}{(2k+1)!} \mathbf{E}_{\xi_i} |2t\xi_i|^{2k+1} \right] \\
&\leq 1 + \sum_{k=1}^{\infty} \left[\frac{1}{(2k)!} \mathbf{E} |2t\xi_i|^{2k} + \frac{0.5}{(2k)!} \mathbf{E} |2t\xi_i|^{2k} + \frac{1}{(2k+2)!} \mathbf{E} |2t\xi_i|^{2k+2} \right] \\
&\leq 1 + 2.5 \sum_{k=1}^{\infty} \frac{1}{(2k)!} \mathbf{E} |2t\xi_i|^{2k} = 1 + 1.25(\mathbf{E} e^{2t\xi_i} + \mathbf{E} e^{-2t\xi_i} - 2) \\
&\leq 1 + 1.25(2e^{2t^2\sigma^2} - 2) \leq e^{5t^2\sigma^2}.
\end{aligned}$$

The second inequality is due to Jensen's inequality. In the third inequality, we have used $|a|^{2k+1}/(2k+1)! \leq 0.5|a|^{2k}/(2k)! + |a|^{2k+2}/(2k+2)!$. The last inequality can be obtained by comparing the Taylor expansion of the function e^x on both sides.

Now let $s_j = \mathbf{E}_{\xi_{j+1}, \dots, \xi_n} \|\sum_{i=1}^n \xi_i \mathbf{z}_i\|_2$. If we regard $h(\xi_j) = s_j / \|\mathbf{z}_j\|_2$ as a function of ξ_j (with variable s_1, \dots, s_{j-1} fixed), then $s_j - s_{j-1} = (h(\xi_j) - \mathbf{E}_{\xi'_j} h(\xi'_j)) \|\mathbf{z}_j\|_2$ and $h(\xi_j) - h(\xi'_j) \leq |\xi_j| + |\xi'_j|$. Therefore from the above inequality, we have $\mathbf{E}_{\xi_j} e^{t(s_j - s_{j-1})} \leq e^{5\|\mathbf{z}_j\|_2^2 t^2 \sigma^2}$, and

$$\mathbf{E}_{\xi_1, \dots, \xi_j} e^{ts_j} = \mathbf{E}_{\xi_1, \dots, \xi_{j-1}} e^{ts_{j-1}} \mathbf{E}_{\xi_j} e^{t(s_j - s_{j-1})} \leq e^{5\|\mathbf{z}_j\|_2^2 \sigma^2 t^2} \mathbf{E}_{\xi_1, \dots, \xi_{j-1}} e^{ts_{j-1}}.$$

By induction and $\sum_j \|\mathbf{z}_j\|_2^2 = \bar{k}$, we obtain $\mathbf{E}_{\xi_1, \dots, \xi_n} e^{ts_n} \leq e^{5\sigma^2 t^2 \bar{k}} e^{ts_0}$, which implies that $P(s_n \geq s_0 + \sqrt{\bar{k}\epsilon}) e^{t(s_0 + \sqrt{\bar{k}\epsilon})} \leq e^{5\bar{k}\sigma^2 t^2} e^{ts_0}$. Let $t = \epsilon / (10\sqrt{\bar{k}\sigma^2})$, we have $P(s_n \geq s_0 + \sqrt{\bar{k}\epsilon}) \leq e^{-\epsilon^2 / (20\sigma^2)}$.

Note that

$$\mathbf{E}\xi_i^2 = \lim_{t \rightarrow 0} \frac{2}{t^2} (\mathbf{E}_{\xi_i} e^{t\xi_i} - 1) \leq \lim_{t \rightarrow 0} \frac{2(e^{\sigma^2 t^2 / 2} - 1)}{t^2} = \sigma^2.$$

Therefore, $s_0 = \mathbf{E} \|\sum_{i=1}^n \xi_i \mathbf{z}_i\|_2 \leq (\sum_{i=1}^n \mathbf{E}\xi_i^2 \|\mathbf{z}_i\|_2^2)^{1/2} \leq \sqrt{\bar{k}\sigma}$. We thus obtain using (6) that

$$P\left(\left\|\sum_{i=1}^n \xi_i \mathbf{z}_i\right\|_2 \geq \sqrt{\bar{k}}(\sigma + \epsilon)\right) \leq e^{-\epsilon^2 / (20\sigma^2)},$$

which implies the desired bound. ■

The following lemma provides an immediate result that can be combined with Lemma A.3, and is used in the proof of Lemma A.8.

Lemma A.7 *Let Assumption 3.1 hold. Consider any fixed $\bar{F} \subset \{1, \dots, d\}$ and let $\hat{\mathbf{f}} = f(\hat{\mathbf{w}}(\bar{F}))$. Define*

$$\omega(\bar{F}) = \sup_j |(f(\hat{\mathbf{w}}(\bar{F}), \mathbf{E}\mathbf{y})) - \mathbf{E}\mathbf{y}|^T \mathbf{f}_j|/n.$$

For all $\eta \in (0, 1)$, with probability larger than $1 - \eta$, we have $\forall \mathbf{f} = f(\mathbf{w}) \in R^n$ with $s = |\text{supp}(\mathbf{w})|$ and $s' = |\text{supp}(\mathbf{w}) - \bar{F}|$:

$$|(\hat{\mathbf{f}} - \mathbf{y})^T \mathbf{f}| \leq \sqrt{s'/\rho(s)} \|\mathbf{f}\|_2 \left[\sigma \sqrt{2 \ln(2d/\eta)} + \sqrt{n} \omega(\bar{F}) \right].$$

Proof Let P be the projection operator to the subspace spanned by $\{\mathbf{f}_j : j \in \bar{F}\}$ on R^n . Lemma A.4 implies that with probability larger than $1 - \eta$:

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

Since $(\hat{\mathbf{f}} - \mathbf{y})^T \mathbf{f}_j = 0$ for all $j \in \bar{F}$, we have $\hat{\mathbf{f}} = P\mathbf{y}$. Moreover, let $\bar{\mathbf{f}} = f(\hat{\mathbf{w}}(\bar{F}), \mathbf{E}\mathbf{y})$, then $\bar{\mathbf{f}} = P\mathbf{E}\mathbf{y}$ and $(I - P)(\bar{\mathbf{f}} - \mathbf{E}\mathbf{y}) = \bar{\mathbf{f}} - \mathbf{E}\mathbf{y}$. Therefore

$$\begin{aligned} |(\hat{\mathbf{f}} - \mathbf{y})^T \mathbf{f}| &= |((I - P)\mathbf{y})^T \mathbf{f}| \\ &= |(\mathbf{y} - P\mathbf{E}\mathbf{y})^T (I - P) \mathbf{f}| \\ &= |(\mathbf{y} - \bar{\mathbf{f}})^T \sum_{j \notin \bar{F}} \mathbf{w}_j (I - P) \mathbf{f}_j| \\ &\leq \sum_{j \notin \bar{F}} |\mathbf{w}_j| \left[|(\mathbf{y} - \mathbf{E}\mathbf{y})^T (I - P) \mathbf{f}_j| + |(\bar{\mathbf{f}} - \mathbf{E}\mathbf{y})^T (I - P) \mathbf{f}_j| \right] \\ &\leq \sum_{j \notin \bar{F}} |\mathbf{w}_j| \left[\sigma \sqrt{2n \ln(2d/\eta)} + n\omega(\bar{F}) \right] \\ &\leq \sqrt{s'} \|\mathbf{w}\|_2 \left[\sigma \sqrt{2n \ln(2d/\eta)} + n\omega(\bar{F}) \right] \\ &\leq \sqrt{s'/\rho(s)} \|\mathbf{f}\|_2 \left[\sigma \sqrt{2 \ln(2d/\eta)} + \sqrt{n} \omega(\bar{F}) \right]. \end{aligned}$$

This proves the desired bound. ■

The following lemma can be used to simplify the bound in Lemma A.1.

Lemma A.8 *Under the conditions of Lemma A.7, and the same probability event as that of Lemma A.7, we have: if $\|\mathbf{f} - \hat{\mathbf{f}}\|_2 \geq 4\sqrt{s'/\rho(s)} \left[\sigma \sqrt{2 \ln(2d/\eta)} + \sqrt{n} \omega(\bar{F}) \right]$, then*

$$\frac{1}{2n} \|\mathbf{f} - \hat{\mathbf{f}}\|_2^2 \leq R(\mathbf{w}) - R(\hat{\mathbf{w}}(\bar{F})) \leq \frac{3}{2n} \|\mathbf{f} - \hat{\mathbf{f}}\|_2^2.$$

Proof We have from Lemma A.7:

$$\begin{aligned}
& \left| R(\mathbf{w}) - R(\hat{\mathbf{w}}) - \frac{1}{n} \|\mathbf{f} - \hat{\mathbf{f}}\|_2^2 \right| \\
&= \frac{1}{n} \left| \|\mathbf{f} - \mathbf{y}\|_2^2 - \|\hat{\mathbf{f}} - \mathbf{y}\|_2^2 - \|\mathbf{f} - \hat{\mathbf{f}}\|_2^2 \right| \\
&= \frac{2}{n} |(\hat{\mathbf{f}} - \mathbf{y})^T (\mathbf{f} - \hat{\mathbf{f}})| \\
&\leq \frac{2}{n} \sqrt{s'/\rho(s)} \|\mathbf{f} - \hat{\mathbf{f}}\|_2 \left[\sigma \sqrt{2 \ln(2d/\eta)} + \sqrt{n\omega(\bar{F})} \right] \\
&\leq \frac{1}{2n} \|\mathbf{f} - \hat{\mathbf{f}}\|_2^2.
\end{aligned}$$

By rearranging the inequality, we obtain the desired bounds. ■

The following lemma can be used to compare the empirical minimizer $\hat{\mathbf{w}}(F)$ with the estimated feature set F and that of $\hat{\mathbf{w}}(\bar{F})$ with a pre-selected feature set \bar{F} . This can be used to estimate the behavior of the estimator when FoBa terminates.

Lemma A.9 *Let Assumption 3.1 hold. Consider any fixed $\bar{F} \subset \{1, \dots, d\}$ and let $\hat{\mathbf{f}} = f(\hat{\mathbf{w}}(\bar{F}))$. Define $\omega(\bar{F})$ as in Lemma A.7. For all $\eta \in (0, 1)$, with probability larger than $1 - \eta$ (which shares the same probability event as that of Lemma A.7), we have*

$$\begin{aligned}
\sqrt{n\rho(|F \cup \bar{F}|)} \|\hat{\mathbf{w}}(F) - \hat{\mathbf{w}}(\bar{F})\|_2 &\leq \|f(\hat{\mathbf{w}}(F) - \hat{\mathbf{w}}(\bar{F}))\|_2 \\
&\leq \sqrt{|F - \bar{F}|/\rho(|F|)} \left[\sigma \sqrt{2 \ln(2d/\eta)} + \sqrt{n\omega(\bar{F})} \right] + \|\delta\hat{\mathbf{f}}\|_2,
\end{aligned}$$

where $\delta\hat{\mathbf{f}} = \sum_{j \notin F} \hat{\mathbf{w}}(\bar{F})_j \mathbf{f}_j$.

Proof Let $\mathbf{w} \in R^{|F|}$ be the restriction of $\hat{\mathbf{w}}(F) \in R^d$ on F , and $G \in R^{n \times |F|}$ be the matrix with columns \mathbf{f}_j for $j \in F$. Let $\hat{\mathbf{f}} = f(\hat{\mathbf{w}}(\bar{F}))$, then the assumption of the lemma implies that we may write it as $\hat{\mathbf{f}} = G\tilde{\mathbf{w}} + \delta\hat{\mathbf{f}}$ for some $\tilde{\mathbf{w}} \in R^{|F|}$. Since $G^T G \mathbf{w} = G^T \mathbf{y}$, we obtain

$$\mathbf{w} - \tilde{\mathbf{w}} = (G^T G)^{-1} G^T (\mathbf{y} - \hat{\mathbf{f}} + \delta\hat{\mathbf{f}}).$$

Therefore noting that $f(\hat{\mathbf{w}}(F)) = G\mathbf{w}$, we obtain

$$f(\hat{\mathbf{w}}(F)) - \hat{\mathbf{f}} = G(G^T G)^{-1} G^T (\mathbf{y} - \hat{\mathbf{f}} + \delta\hat{\mathbf{f}}) - \delta\hat{\mathbf{f}}. \quad (7)$$

Since by the definition of $\hat{\mathbf{f}}$, $\mathbf{f}_j^T (\mathbf{y} - \hat{\mathbf{f}}) = 0$ for $j \in \bar{F}$, and by Lemma A.7, with probability larger than $1 - \eta$, for all $j \in F - \bar{F}$:

$$|\mathbf{f}_j^T (\mathbf{y} - \hat{\mathbf{f}})| \leq \sigma \sqrt{2 \ln(2d/\eta)n} + n\omega(\bar{F}).$$

We obtain

$$\|G^T (\mathbf{y} - \hat{\mathbf{f}})\|_2 \leq \sqrt{|F - \bar{F}|} \|G^T (\mathbf{y} - \hat{\mathbf{f}})\|_\infty \leq \sqrt{|F - \bar{F}|} \left[\sigma \sqrt{2 \ln(2d/\eta)n} + n\omega(\bar{F}) \right].$$

According to Definition 3.1, $\rho(|F|)n$ is no larger than the smallest eigenvalue of $G^T G$. Therefore

$$\|G(G^T G)^{-1}G^T(\mathbf{y} - \hat{\mathbf{f}})\|_2 \leq \sqrt{|F - \bar{F}|/\rho(|F|)} \left[\sigma \sqrt{2 \ln(2d/\eta)} + \sqrt{n\omega(\bar{F})} \right].$$

In addition, since $I - G(G^T G)^{-1}G^T$ is a projection operator, we have

$$\|G(G^T G)^{-1}G^T \delta \hat{\mathbf{f}} - \delta \hat{\mathbf{f}}\|_2 \leq \|\delta \hat{\mathbf{f}}\|_2.$$

Therefore by combining the above two estimates with (7), we get

$$\|f(\hat{\mathbf{w}}(F)) - \hat{\mathbf{f}}\|_2 \leq \sqrt{|F - \bar{F}|/\rho(|F|)} \left[\sigma \sqrt{2 \ln(2d/\eta)} + \sqrt{n\omega(\bar{F})} \right] + \|\delta \hat{\mathbf{f}}\|_2.$$

This implies the second inequality of the lemma. Note that the first inequality is due to the definition of $\rho(k)$ in Definition 3.1. ■

B Proof of Theorem 3.1

We shall first proof a more general result, and then specializes to Theorem 3.1.

B.1 A General Result

Theorem B.1 *Consider the FoBa algorithm in Figure 4, where Assumption 3.1 holds. Given any fixed set $\bar{F} \subset \{1, \dots, d\}$, let $\bar{\mathbf{w}} = \hat{\mathbf{w}}(\bar{F}, \mathbf{E}\mathbf{y})$, $\bar{\mathbf{f}} = f(\bar{\mathbf{w}})$, and*

$$\omega(\bar{F}) = \sup_j |(\bar{\mathbf{f}} - \mathbf{E}\mathbf{y})^T \mathbf{f}_j|/n.$$

Let $\bar{k} = |\bar{F}|$, and $\epsilon > 0$ be the stopping criterion in Figure 4. Let $s \leq d$ be an integer which satisfies one of the following conditions:

$$8\bar{k} \leq s\rho(s)^2 \quad \text{or} \quad 2\bar{k} \ln(2\bar{k} + 2n^{-1}\|\bar{\mathbf{f}}\|_2^2/\epsilon) + \bar{k} + 1 \leq s\rho(s) \quad \text{or} \quad s = d.$$

Assume that for some $\eta \in (0, 1/3)$, we have

$$\epsilon \geq 32\rho(s)^{-2} \left[\sigma \sqrt{2 \ln(2d/\eta)/n} + \epsilon(\bar{F}) \right]^2.$$

Let

$$k(\epsilon) = |\{j \in \bar{F} : |\bar{\mathbf{w}}_j|^2 \leq 8\epsilon/\rho(s)^2\}|.$$

With probability larger than $1 - 3\eta$:

- *The algorithm terminates after at most*

$$\frac{4}{n\epsilon} \|\bar{\mathbf{f}}\|_2^2 + \frac{4\bar{k}}{n\epsilon} \sigma^2 \left[1 + \sqrt{20 \ln(1/\eta)} \right]^2 + \frac{256k(\epsilon)\lambda(\bar{F})^2}{9\rho(s)^3}$$

forward iterations.

- When the algorithm terminates, we have:

$$\begin{aligned}
|\bar{F} - F^{(k)}| &\leq 2k(\epsilon), \\
\sum_{j \in \bar{F} - F^{(k)}} |\bar{\mathbf{w}}_j|^2 &\leq 8k(\epsilon)\epsilon/\rho(s)^2, \\
|F^{(k)} - \bar{F}| &\leq \frac{2}{\epsilon} \|\mathbf{w}^{(k)} - \hat{\mathbf{w}}(\bar{F})\|_2^2 \leq \frac{256k(\epsilon)\lambda(\bar{F})}{9\rho(s)^3}, \\
\|\hat{\mathbf{w}}(\bar{F}) - \bar{\mathbf{w}}\|_\infty &\leq \sigma \sqrt{2 \ln(2\bar{k}/\eta)/(n\rho(\bar{k}))} \\
\|\hat{\mathbf{w}}(\bar{F}) - \bar{\mathbf{w}}\|_2 &\leq \sigma \sqrt{\bar{k}/(n\rho(\bar{k}))} \left[1 + \sqrt{20 \ln(1/\eta)}\right].
\end{aligned}$$

This theorem is a trivial consequence of the following two lemmas. Note that when the algorithm terminates, Lemma A.5 and Lemma A.6 imply that

$$\|\mathbf{w}^{(k)} - \bar{\mathbf{w}}\|_\infty \leq \sigma \sqrt{2 \ln(2\bar{k}/\eta)/(n\rho(\bar{k}))}$$

and

$$\|\mathbf{w}^{(k)} - \bar{\mathbf{w}}\|_2 \leq \sigma \sqrt{\bar{k}/(n\rho(\bar{k}))} \left[1 + \sqrt{20 \ln(1/\eta)}\right].$$

Lemma B.1 *Under the conditions of Theorem B.1, we have with probability larger than $1 - 2\eta$: at the beginning of each forward step, we have:*

- The “if condition” of Lemma A.8 holds:

$$\|f(\mathbf{w}^{(k-1)} - \hat{\mathbf{w}}(\bar{F}))\|_2 \geq 4\sqrt{|F^{(k-1)} - \bar{F}|/\rho(s)} \left[\sigma \sqrt{2 \ln(2d/\eta)} + \sqrt{n}\omega(\bar{F})\right].$$

- If $s < d$, then $|F^{(k-1)} \cup \bar{F}| < s$.
- If $|F^{(k-1)} - \bar{F}| = 0$, then $F^{(k-1)} = \bar{F}$, and the algorithm terminates.
- If $|F^{(k-1)} - \bar{F}| > 0$, and the algorithm terminates after the current forward-step, then

$$\begin{aligned}
\frac{1}{2}|\bar{F} - F^{(k-1)}| &\leq k(\epsilon), \\
\sum_{j \in \bar{F} - F^{(k-1)}} |\bar{\mathbf{w}}_j|^2 &\leq 8k(\epsilon)\epsilon/\rho(s)^2.
\end{aligned}$$

Moreover,

$$|F^{(k-1)} - \bar{F}| \leq \frac{2}{\epsilon} \|\mathbf{w}^{(k-1)} - \hat{\mathbf{w}}(\bar{F})\|_2^2 \leq \frac{256k(\epsilon)\lambda(\bar{F})}{9\rho(s)^3}.$$

Proof

Note that with probability larger than $1 - 3\eta$, the probability events in Lemma A.5, Lemma A.6, and Lemma A.7 hold. Therefore their claims hold.

We prove the first two claims together by induction under the induction hypothesis that they hold at the current iteration with the additional assumption that $\bar{F} \not\subset F^{(k-1)}$.

In order to see that the first claim of the lemma is true in the next iteration, we note that after the loop in the current iteration's backward step, if the "if condition" of Lemma A.8 does not hold, then the squared error increase in a backward step is at most $16\rho(s)^{-2}(\sigma\sqrt{2\ln(2d/\eta)/n} + \omega(\bar{F}))^2$ according to Lemma A.2, which is smaller than $\epsilon/2$ by the assumption of the theorem. However, this is not possible because this means that another backward step should have been taken without exiting the backward step loop. Therefore after the loop in the backward step finishes, the "if condition" of Lemma A.8 has to hold. That is, the first claim holds at the next iteration.

In order to see that the second claim of the lemma is true at the next iteration, we note that by induction hypothesis, the condition of Lemma A.8 holds at the beginning of the current forward step. It implies that we may take $\beta = 1/2$ and $\mathbf{w}' = \hat{\mathbf{w}}(\bar{F})$ in Lemma A.1, which implies that the squared error reduction in this forward-step is at least

$$\begin{aligned} & \frac{\rho(s)(1+\beta)}{4|\bar{F} - F^{(k-1)}|} \left(\frac{1}{n} \|f(\mathbf{w}^{(k-1)} - \hat{\mathbf{w}}(\bar{F}))\|_2^2 + R(\mathbf{w}^{(k-1)}) - R(\hat{\mathbf{w}}(\bar{F})) \right) \\ & \geq \frac{5\rho(s)}{8|\bar{F} - F^{(k-1)}|} \left(R(\mathbf{w}^{(k-1)}) - R(\hat{\mathbf{w}}(\bar{F})) \right). \end{aligned} \quad (8)$$

The inequality uses Lemma A.8. Now, if the second claim fails in the next iteration, then the condition $|F^{(k)} \cup \bar{F}| = s$ holds after the current iteration's backward loop finishes with no backward step being taken. At this point, we have shown earlier that the claim of Lemma A.8 holds. One of the following two situations is valid.

- $8\bar{k} \leq s\rho(s)^2$: we obtain from Lemma A.2 and Lemma A.8 that the reduction is at most

$$\begin{aligned} \frac{1}{\rho(s)|F^{(k)} - \bar{F}|n} \|\mathbf{f}^{(k)} - \hat{\mathbf{f}}\|_2^2 & \leq \frac{2}{\rho(s)(s - \bar{k})} (R(\mathbf{w}^{(k)}) - R(\hat{\mathbf{w}}(\bar{F}))) \\ & < \frac{5\rho(s)}{16\bar{k}} (R(\mathbf{w}^{(k)}) - R(\hat{\mathbf{w}}(\bar{F}))), \end{aligned}$$

where in the last inequality, we have used the condition $\bar{k} \leq s\rho(s)^2/8 < 5s\rho(s)^2(32+5\rho(s)^2)^{-1}$. However, this squared error reduction is smaller than half of the squared error reduction in the preceding forward step, which is no less than the value given by (8), where we note that $R(\mathbf{w}^{(k-1)}) > R(\mathbf{w}^{(k)})$. This means a backward step should be taken, which is a contradiction.

- $2\bar{k} \ln(2\bar{k} + 2n^{-1}\|\hat{\mathbf{f}}\|_2^2/\epsilon) + \bar{k} + 1 \leq s\rho(s)$: since Lemma A.8 implies that $R(\mathbf{w}^{(k)}) - R(\hat{\mathbf{w}}(\bar{F})) \geq 0$, and no backward step is taken, $R(\mathbf{w}^{(k-1)}) - R(\hat{\mathbf{w}}(\bar{F})) \geq \epsilon$. We can apply (8) recursively to obtain:

$$\begin{aligned} \epsilon & \leq R(\mathbf{w}^{(k-1)}) - R(\hat{\mathbf{w}}(\bar{F})) \\ & < (1 - 0.5\rho(s)/\bar{k})[R(\mathbf{w}^{(k-2)}) - R(\hat{\mathbf{w}}(\bar{F}))] \\ & < \dots < (1 - 0.5\rho(s)/\bar{k})^{k-1}[R(0) - R(\hat{\mathbf{w}}(\bar{F}))] \\ & \leq e^{-0.5(k-1)\rho(s)/\bar{k}}[R(0) - R(\hat{\mathbf{w}}(\bar{F}))] = e^{-0.5(k-1)\rho(s)/\bar{k}}\|\hat{\mathbf{f}}(\bar{F})\|_2^2/n. \end{aligned}$$

The last equality is due to simple algebra, and its detailed derivation can be found in (11). Therefore the above inequality and $k + \bar{k} \geq s$ imply

$$(s - \bar{k} - 1)\rho(s) \leq (k - 1)\rho(s) < 2\bar{k} \ln(\|\hat{\mathbf{f}}(\bar{F})\|_2^2/(n\epsilon)) < 2\bar{k} \ln(2\|\hat{\mathbf{f}}\|_2^2/(\epsilon n) + 2\bar{k}),$$

where the last inequality uses $\|\hat{\mathbf{f}}(\bar{F})\|_2 < \|\hat{\mathbf{f}}\|_2 + \sqrt{n\bar{k}\epsilon}$, which is from Lemma A.6. However, this is a contradiction to the condition $2\bar{k} \ln(2\bar{k} + 2n^{-1}\|\hat{\mathbf{f}}\|_2^2/\epsilon) + \bar{k} + 1 \leq s\rho(s)$.

The above contradictions imply that the second claim of the lemma has to hold in the next iteration.

Through induction, we know that the first two claims hold until the iteration after $\bar{F} \subset F^{(k-1)}$ (at that time, the algorithm would have ended, according to the third claim of the lemma).

We now prove the third claim of the lemma, where we assume that $\bar{F} \subset F^{(k-1)}$ at the beginning of the forward iteration. If we further have $\bar{F} = F^{(k-1)}$, then Lemma A.7 and Lemma A.3 imply that the squared error reduction in the next forward step is at most

$$(\sigma\sqrt{2\ln(2d/\eta)/n} + \omega(\bar{F}))^2/\rho(s) < \epsilon,$$

and hence the algorithm terminates.

To finish the proof of claim 3, we only need to prove by contradiction that the first time $\bar{F} \subset F^{(k-1)}$, we must have $\bar{F} = F^{(k-1)}$. If this is false, then at the end of the previous backward step, the error increase has to be larger than $\epsilon/2$, which is a quantity smaller than half of δ_k in any earlier forward iterations. Since $\bar{F} \subset F^{(k-1)}$, we can apply Lemma A.9 with $\delta\hat{\mathbf{f}} = 0$ (because $\hat{\mathbf{w}}(\bar{F})_j = 0$ when $j \notin F^{(k-1)}$). However, the bound in Lemma A.9, together with Lemma A.2 imply that this error increase should have been at most $\rho(s)^{-2}(\sigma\sqrt{2\ln(2d/\eta)/n} + \omega(\bar{F}))^2 < \epsilon/2$, which is a contradiction. This proves the third claim.

We now prove claim 4, under the assumption that $\bar{F} \not\subset F^{(k-1)}$ and $|F^{(k-1)} \cup \bar{F}| < s$ at the beginning of the current iteration. Since the condition of Lemma A.8 holds at the beginning of the current forward step. It implies that we may take $\beta = 1/2$ and $\mathbf{w}' = \hat{\mathbf{w}}(\bar{F})$ in Lemma A.1, which implies that the squared error reduction in this forward-step is at least

$$\begin{aligned} & \frac{\rho(s)(1+\beta)}{4|\bar{F} - F^{(k-1)}|} \left(\frac{1}{n} \|f(\mathbf{w}^{(k-1)} - \hat{\mathbf{w}}(\bar{F}))\|_2^2 + R(\mathbf{w}^{(k-1)}) - R(\hat{\mathbf{w}}(\bar{F})) \right) \\ & \geq \frac{9\rho(s)}{16n|\bar{F} - F^{(k-1)}|} \|f(\mathbf{w}^{(k-1)} - \hat{\mathbf{w}}(\bar{F}))\|_2^2 \\ & \geq \frac{9\rho(s)^2}{16|\bar{F} - F^{(k-1)}|} \|\mathbf{w}^{(k-1)} - \hat{\mathbf{w}}(\bar{F})\|_2^2 \\ & \geq \frac{9\rho(s)^2}{16|\bar{F} - F^{(k-1)}|} \sum_{j \in \bar{F} - F^{(k-1)}} |\hat{\mathbf{w}}(\bar{F})_j|^2. \end{aligned}$$

In the above derivation, the first inequality uses Lemma A.8; the second inequality comes from the definition of $\rho(s)$ in Definition 3.1; the third inequality is simple algebra. Since the algorithm terminates, the squared error reduction in the forward-step is no more than ϵ . That is

$$\frac{9\rho(s)^2}{16|\bar{F} - F^{(k-1)}|} \sum_{j \in \bar{F} - F^{(k-1)}} |\hat{\mathbf{w}}(\bar{F})_j|^2 \leq \epsilon. \quad (9)$$

By Lemma A.5, we have $\|\hat{\mathbf{w}}(\bar{F}) - \bar{\mathbf{w}}\|_\infty^2 < 4\epsilon/(9\rho(s)^2)$. From these two inequalities, we obtain

$$\sum_{j \in \bar{F} - F^{(k-1)}} |\bar{\mathbf{w}}_j|^2 \leq 4|\bar{F} - F^{(k-1)}|\epsilon/\rho(s)^2. \quad (10)$$

This implies that

$$|\{j \in \bar{F} - F^{(k-1)} : |\bar{\mathbf{w}}_j|^2 > 8\epsilon/\rho(s)^2\}| \times 8\epsilon/\rho(s)^2 \leq 4|\bar{F} - F^{(k-1)}|\epsilon/\rho(s)^2,$$

which can be simplified to $|\{j \in \bar{F} - F^{(k-1)} : |\bar{\mathbf{w}}_j|^2 > 8\epsilon/\rho(s)^2\}| \leq |\bar{F} - F^{(k-1)}|/2$. Therefore

$$|\{j \in \bar{F} - F^{(k-1)} : |\bar{\mathbf{w}}_j|^2 \leq 8\epsilon/\rho(s)^2\}| \geq |\bar{F} - F^{(k-1)}|/2,$$

which leads to the first displayed inequality of claim 4. Plug this estimate into (10), we obtain the second displayed inequality of claim 4.

Moreover, after the backward loop in the previous iteration, the squared error increase of a potential backward operation is no less than $\epsilon/2$. We thus obtain:

$$\epsilon/2 \leq \inf_{j \in F^{(k-1)}} R(\mathbf{w}^{(k-1)} - \mathbf{w}_j^{(k-1)} \mathbf{e}_j) - R(\mathbf{w}^{(k-1)}).$$

Therefore

$$\begin{aligned} |F^{(k-1)} - \bar{F}| \epsilon/2 &\leq \sum_{j \in F^{(k-1)} - \bar{F}} |\mathbf{w}_j^{(k-1)}|^2 \\ &\leq \|\mathbf{w}^{(k-1)} - \hat{\mathbf{w}}(\bar{F})\|_2^2 \\ &\leq 2|F^{(k-1)} - \bar{F}| \rho(s)^{-2} \left[\sigma \sqrt{2 \ln(2d/\eta)/n} + \omega(\bar{F}) \right]^2 + 2 \frac{\|\delta \hat{\mathbf{f}}\|_2^2}{n\rho(s)} \\ &< |F^{(k-1)} - \bar{F}| \epsilon/4 + 2 \frac{\|\delta \hat{\mathbf{f}}\|_2^2}{n\rho(s)}, \end{aligned}$$

where we define $\|\delta \hat{\mathbf{f}}\|_2$ as in Lemma A.9:

$$\frac{1}{n} \|\delta \hat{\mathbf{f}}\|_2^2 = \frac{1}{n} \left\| \sum_{j \in \bar{F} - F^{(k-1)}} \hat{\mathbf{w}}(\bar{F})_j \mathbf{f}_j \right\|_2^2 \leq \lambda(\bar{F}) \sum_{j \in \bar{F} - F^{(k-1)}} |\hat{\mathbf{w}}(\bar{F})_j|^2.$$

In the above derivation, the first inequality is due to Lemma A.2. The second inequality is simple algebra. The third inequality is due to Lemma A.9. The fourth inequality uses the definition of ϵ and simple algebra. We thus obtain from the above that

$$|F^{(k-1)} - \bar{F}| \epsilon/4 < 2 \frac{\|\delta \hat{\mathbf{f}}\|_2^2}{n\rho(s)}.$$

Therefore we can rewrite the earlier derivation as

$$\begin{aligned} |F^{(k-1)} - \bar{F}| \epsilon/2 &\leq \|\mathbf{w}^{(k-1)} - \hat{\mathbf{w}}(\bar{F})\|_2^2 \\ &\leq 4 \|\delta \hat{\mathbf{f}}\|_2^2 / (n\rho(s)) \\ &\leq \frac{4\lambda(\bar{F})}{\rho(s)} \sum_{j \in \bar{F} - F^{(k-1)}} |\hat{\mathbf{w}}(\bar{F})_j|^2 \\ &\leq \frac{64\lambda(\bar{F})}{9\rho(s)^3} |\bar{F} - F^{(k-1)}| \epsilon. \end{aligned}$$

The third inequality uses the definition of $\|\delta \hat{\mathbf{f}}\|_2$ expressed earlier. The last inequality is due to (9). This finishes the proof of the third displayed inequality of claim 4. \blacksquare

Lemma B.2 *Under the conditions of Theorem B.1, with the same probability event of that of Lemma B.1, FoBa terminates after at most*

$$\frac{4}{n\epsilon} \|\bar{\mathbf{f}}\|_2^2 + \frac{4\bar{k}}{n\epsilon} \sigma^2 \left[1 + \sqrt{20 \ln(1/\eta)}\right]^2 + \frac{256k(\epsilon)\lambda(\bar{F})^2}{9\rho(s)^3}$$

forward iterations.

Proof We have

$$R(0) - R(\hat{\mathbf{w}}(\bar{F})) = \frac{1}{n} \|\mathbf{y}\|_2^2 - \frac{1}{n} \|\hat{\mathbf{f}} - \mathbf{y}\|_2^2 = \frac{1}{n} (\|\hat{\mathbf{f}}\|_2^2 + 2\hat{\mathbf{f}}^T(\mathbf{y} - \hat{\mathbf{f}})) = \frac{1}{n} \|\hat{\mathbf{f}}\|_2^2. \quad (11)$$

The last equality uses the fact that $\hat{\mathbf{f}}$ is a minimizer using features in \bar{F} , and thus $\hat{\mathbf{f}}^T(\mathbf{y} - \hat{\mathbf{f}}) = 0$.

Similarly, using $\mathbf{w}^{(k)} = \hat{\mathbf{w}}(F^{(k)})$, we obtain

$$R\left(\sum_{j \in \bar{F}} \mathbf{w}_j^{(k)} \mathbf{e}_j\right) - R(\mathbf{w}^{(k)}) = \left\| \sum_{j \in F^{(k)} - \bar{F}} \mathbf{w}_j^{(k)} \mathbf{f}_j \right\|^2 / n.$$

Therefore, at the termination of FoBa we have:

$$\begin{aligned} R(0) - R(\mathbf{w}^{(k)}) &= R(0) - R(\hat{\mathbf{w}}(\bar{F})) + R(\hat{\mathbf{w}}(\bar{F})) - R(\mathbf{w}^{(k)}) \\ &\leq R(0) - R(\hat{\mathbf{w}}(\bar{F})) + R\left(\sum_{j \in \bar{F}} \mathbf{w}_j^{(k)} \mathbf{e}_j\right) - R(\mathbf{w}^{(k)}) \\ &= \frac{1}{n} \|\hat{\mathbf{f}}\|_2^2 + \frac{1}{n} \left\| \sum_{j \in F^{(k)} - \bar{F}} \mathbf{w}_j^{(k)} \mathbf{f}_j \right\|^2 \\ &\leq \frac{2}{n} \|\hat{\mathbf{f}} - \bar{\mathbf{f}}\|_2^2 + \frac{2}{n} \|\bar{\mathbf{f}}\|_2^2 + \lambda(\bar{F}) \sum_{j \in F^{(k)} - \bar{F}} |\mathbf{w}_j^{(k)}|^2 \\ &\leq \frac{2}{n} \|\bar{\mathbf{f}}\|_2^2 + \frac{2\bar{k}}{n} \sigma^2 \left[1 + \sqrt{20 \ln(1/\eta)}\right]^2 + \frac{128k(\epsilon)\lambda(\bar{F})^2}{9\rho(s)^3} \epsilon. \end{aligned}$$

Lemma B.1 and Lemma A.6 are used to obtain the third inequality.

The lemma is now a direct consequence of the fact that each forward iteration contributes at least $\epsilon/2$ to the total squared error reduction. ■

B.2 Proof of Theorem 3.1

We need the following lemma.

Lemma B.3 *Under the assumptions of Theorem 3.1, there exists \bar{F}^* such that*

- $\bar{F} \subset \bar{F}^*$
- $|\bar{F}^* - \bar{F}| \leq 4\Delta/\epsilon$.
- $\omega(\bar{F}^*) \leq \sqrt{\epsilon}/2$

- $\rho(\bar{k})\|\bar{\mathbf{w}} - \bar{\mathbf{w}}^*\|_2^2 \leq \|f(\bar{\mathbf{w}} - \bar{\mathbf{w}}^*)\|_2^2/n \leq \Delta$, where $\bar{\mathbf{w}}^* = \hat{\mathbf{w}}(\bar{F}^*, \mathbf{E}\mathbf{y})$.

Proof We start with $\mathbf{w}^{*(0)} = \bar{\mathbf{w}}$ and $\bar{F}^{*(0)} = \bar{F}$ and run forward-greedy steps by adding one feature at a time until the condition $\omega(F^{*(k)}) \leq \sqrt{\epsilon}/2$ is satisfied:

- $(\alpha^{(k)}, i^{(k)}) = \arg \min_{(\alpha, i)} \|f(\mathbf{w}^{(k)} + \alpha \mathbf{e}_i) - \mathbf{E}\mathbf{y}\|_2^2$
- $\bar{F}^{*(k+1)} = \bar{F}^{*(k)} \cup \{i^{(k)}\}$
- $\mathbf{w}^{*(k+1)} = \hat{\mathbf{w}}(\bar{F}^{*(k+1)}, \mathbf{E}\mathbf{y})$

Similar to Lemma A.3 (with \mathbf{y} replaced by $\mathbf{E}\mathbf{y}$), we have at any iteration k before stopping:

$$\|f(\mathbf{w}^{*(k)}) - \mathbf{E}\mathbf{y}\|_2^2 - \|f(\mathbf{w}^{*(k+1)}) - \mathbf{E}\mathbf{y}\|_2^2 \geq \sup_j |(f(\mathbf{w}^{*(k)}) - \mathbf{E}\mathbf{y})^T \mathbf{f}_j|^2/n \geq n\epsilon/4.$$

Therefore by aggregating the inequality from 0 to k , we obtain at stopping:

$$n\Delta = \|f(\mathbf{w}^{*(0)}) - \mathbf{E}\mathbf{y}\|_2^2 \geq nk\epsilon/4,$$

and thus $k \leq 4\Delta/\epsilon$. We let $\bar{F}^* = F^{*(k)}$ at the time k when the procedure stops, then $|\bar{F}^*| \leq |\bar{F}| + 4\Delta/\epsilon$ and the condition $\omega(\bar{F}^*) \leq \sqrt{\epsilon}/2$ is satisfied. This proves the first three claims.

Because $\bar{F} \subset \bar{F}^*$:

$$\|f(\bar{\mathbf{w}} - \bar{\mathbf{w}}^*)\|_2^2 = \|f(\bar{\mathbf{w}}) - \mathbf{E}\mathbf{y}\|_2^2 - \|f(\bar{\mathbf{w}}^*) - \mathbf{E}\mathbf{y}\|_2^2 \leq n\Delta.$$

We obtain the fourth claim. ■

Now, to prove the theorem, we can apply Theorem B.1 with \bar{F}^* of Lemma B.3 as \bar{F} in Theorem B.1, and $\bar{\mathbf{w}}^* = \hat{\mathbf{w}}(\bar{F}^*, \mathbf{E}\mathbf{y})$ as $\bar{\mathbf{w}}$ in Theorem B.1. Note that the third claim of Lemma B.3 implies that the required lower bound for ϵ is satisfied in Theorem B.1. Moreover, it follows from the last claim of Lemma B.3 that

$$\rho(\bar{k}) \left| \left\{ j \in \bar{F}^* : |\bar{\mathbf{w}}_j - \bar{\mathbf{w}}_j^*| \geq 0.5\sqrt{\epsilon/\rho(\bar{k})} \right\} \right| \left[0.5\sqrt{\epsilon/\rho(\bar{k})} \right]^2 \leq \Delta.$$

This implies that

$$\begin{aligned} & \left| \{j \in \bar{F}^* : |\bar{\mathbf{w}}_j^*|^2 \leq 8\epsilon/\rho(s)^2\} \right| \\ & \leq \left| \{j \in \bar{F} : |\bar{\mathbf{w}}_j|^2 \leq 12\epsilon/\rho(s)^2\} \right| + \left| \left\{ j \in \bar{F} : |\bar{\mathbf{w}}_j - \bar{\mathbf{w}}_j^*| \geq 0.5\sqrt{\epsilon/\rho(\bar{k})} \right\} \right| + 2\lfloor 4\Delta/\epsilon \rfloor \\ & \leq \left| \{j \in \bar{F} : |\bar{\mathbf{w}}_j|^2 \leq 12\epsilon/\rho(s)^2\} \right| + 2\lfloor 4\Delta/\epsilon \rfloor. \end{aligned}$$

Therefore the choice of $k(\epsilon)$ in Theorem 3.1 is an upperbound of $k(\epsilon)$ as defined in Theorem B.1.

Now the first claim of the theorem follows from Theorem B.1 and the last claim of Lemma B.3. The first two inequalities of the second claim follow directly from Theorem B.1. The third inequality uses Theorem B.1, the last claim of Lemma B.3: $\|\bar{\mathbf{w}} - \bar{\mathbf{w}}^*\|_2^2 \leq \Delta/\rho(\bar{k})$, and the simple inequality of the form $\|\mathbf{w}_1 + \mathbf{w}_2\|_2^2 \leq 2\|\mathbf{w}_1\|_2^2 + 2\|\mathbf{w}_2\|_2^2$. The last inequality of the second claim follows from the last claim of Lemma B.3:

$$\|\bar{\mathbf{w}} - \bar{\mathbf{w}}^*\|_2 \leq \sqrt{\Delta/\rho(\bar{k})},$$

and the following bounds in Theorem B.1:

$$\begin{aligned}\|\hat{\mathbf{w}}(\bar{F}^*) - \bar{\mathbf{w}}^*\|_2 &\leq \sigma \sqrt{\bar{k}/(n\rho(\bar{k}))} \left[1 + \sqrt{20 \ln(1/\eta)}\right], \\ \|\mathbf{w}^{(k)} - \hat{\mathbf{w}}(\bar{F}^*)\|_2 &\leq \sqrt{\frac{128k(\epsilon)\lambda(\bar{k})\epsilon}{9\rho(s)^3}}.\end{aligned}$$

By summing the above three inequalities, we obtain the desired result.