Forward-Backward Greedy Algorithms for General Convex Smooth Functions over A Cardinality Constraint

Ji Liu JI-LIU@CS.WISC.EDU
Department of Computer Sciences, University of Wisconsin-Madison

Ryohei Fujimaki RFUJIMAKI@NEC-LABS.COM
Department of Media Analytics, NEC Lab America, Inc.

Jieping Ye JIEPING.YE@ASU.EDU
Department of Computer Science and Engineering, Arizona State University

Abstract

We consider forward-backward greedy algorithms for solving sparse feature selection problems with general convex smooth functions. A state-of-the-art greedy method, the Forward-Backward greedy algorithm (FoBa-obj) requires to solve a large number of optimization problems, thus it is not scalable for large-size problems. The FoBa-gdt algorithm, which uses the gradient information for feature selection at each forward iteration, significantly improves the efficiency of FoBa-obj. In this paper, we systematically analyze the theoretical properties of both algorithms. Our main contributions are: 1) We derive better theoretical bounds than existing analyses regarding FoBa-obj for general smooth convex functions; 2) We show that FoBa-gdt achieves the same theoretical performance as FoBa-obj under the same condition: restricted strong convexity condition. Our new bounds are consistent with the bounds of a special case (least squares) and fills a previously existing theoretical gap for general convex smooth functions; 3) We show that the restricted strong convexity condition is satisfied if the number of independent samples is more than $\bar{k}\log d$ where $\bar{k}$ is the sparsity number and d is the dimension of the variable; 4) We apply FoBa-gdt (with the conditional random field objective) to the sensor selection problem for human indoor activity recognition and our results show that FoBa-gdt outperforms other methods based on forward greedy selection and L1-regularization.

Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

1. Introduction

Feature selection has been one of the most significant issues in machine learning and data mining. Following the success of Lasso (Tibshirani, 1994), learning algorithms with sparse regularization (a.k.a. sparse learning) have recently received significant attention. A classical problem is to estimate a signal $\beta^* \in \mathbb{R}^d$ from a feature matrix $X \in \mathbb{R}^{n\times d}$ and an observation $y = X\beta^* + \text{noise} \in \mathbb{R}^n$, under the assumption that β^* is sparse (i.e., β^* has $\bar{k} \ll d$ nonzero elements). Previous studies have proposed many powerful tools to estimate β^*. In addition, in certain applications, reducing the number of features has a significantly practical value (e.g., sensor selection in our case).

The general sparse learning problems can be formulated as follows (Jalali et al., 2011):

$$\bar{\beta} := arg\min_{\beta} : Q(\beta; X, y) \qquad s.t. : \quad \|\beta\|_0 \leq \bar{k}. \tag{1}$$

where $Q(\beta; X, y)$ is a convex smooth function in terms of β such as the least square loss (Tropp, 2004) (regression), the Gaussian MLE (or log-determinant divergence) (Ravikumar et al., 2011) (covariance selection), and the logistic loss (Kleinbaum & Klein, 2010) (classification). $\|\beta\|_0$ denotes ℓ_0-norm, that is, the number of nonzero entries of $\beta \in \mathbb{R}^d$. Hereinafter, we denote $Q(\beta; X, y)$ simply as $Q(\beta)$.

From an algorithmic viewpoint, we are mainly interested in three aspects for the estimator $\hat{\beta}$: (i) estimation error $\|\hat{\beta} - \bar{\beta}\|$; (ii) objective error $Q(\hat{\beta}) - Q(\bar{\beta})$; and (iii) feature selection error, that is, the difference between $\text{supp}(\hat{\beta})$ and $\bar{F} := \text{supp}(\bar{\beta})$, where $\text{supp}(\beta)$ is a feature index set corresponding to nonzero elements in β. Since the constraint

defines a non-convex feasible region, the problem is non-convex and generally NP-hard.

There are two types of approaches to solve this problem in the literature. Convex-relaxation approaches replace ℓ_0-norm by ℓ_1-norm as a sparsity penalty. Such approaches include Lasso (Tibshirani, 1994), Danzig selector (Candès & Tao, 2007), and L1-regularized logistic regression (Kleinbaum & Klein, 2010). Alternative greedy-optimization approaches include orthogonal matching pursuit (OMP) (Tropp, 2004; Zhang, 2009), backward elimination, and forward-backward greedy method (FoBa) (Zhang, 2011a), which use greedy heuristic procedure to estimate sparse signals. Both types of algorithms have been well studied from both theoretical and empirical perspectives.

FoBa has been shown to give better theoretical properties than LASSO and Dantzig selector for the least squared loss function: $Q(\beta) = \frac{1}{2}\|X\beta - y\|^2$ (Zhang, 2011a). Jalali et al. (2011) has recently extended it to general convex smooth functions. Their method and analysis, however, pose computational and theoretical issues. First, since FoBa solves a large number of single variable optimization problems in every forward selection step, it is computationally expensive for general convex functions if the sub-problems have no closed form solution. Second, though they have empirically shown that FoBa performs well for general smooth convex functions, their theoretical results are weaker than those for the least square case (Zhang, 2011a). More precisely, their upper bound for estimation error is looser and their analysis requires more restricted conditions for feature selection and signal recovery consistency. The question of whether or not FoBa can achieve the same theoretical bound in the general case as in the least square case motivates this work.

This paper addresses the theoretical and computational issues associated with the standard FoBa algorithm (hereinafter referred to as FoBa-obj because it solves single variable problems to minimize the objective in each forward selection step). We study a new algorithm referred to as "gradient" FoBa (FoBa-gdt) which significantly improves the computational efficiency of FoBa-obj. The key difference is that FoBa-gdt only evaluates gradient information in individual forward selection steps rather than solving a large number of single variable optimization problems. Our contributions are summarized as follows.

Theoretical Analysis of FoBa-obj and FoBa-gdt This paper presents three main theoretical contributions. First, we derive better theoretical bounds for estimation error, objective error, and feature selection error than existing analyses for FoBa-obj for general smooth convex functions (Jalali et al., 2011) under the same condition: restricted strong convexity condition. Second, we show that FoBa-gdt achieves the same theoretical performance as FoBa-obj. Our new bounds are consistent with the bounds of a special case, i.e., the least square case, and fills in the theoretical gap between the general loss (Jalali et al., 2011) and the least squares loss case (Zhang, 2011a). Our result also implies an interesting result: when the signal noise ratio is big enough, the NP hard problem (1) can be solved by using FoBa-obj or FoBa-gdt. Third, we show that the restricted strong convexity condition is satisfied for a class of commonly used machine learning objectives, e.g., logistic loss and least square loss, if the number of independent samples is greater than $\bar{k}\log d$ where $\bar{k}$ is the sparsity number and d is the dimension of the variable.

Application to Sensor Selection We have applied FoBa-gdt with the CRF loss function (referred to as FoBa-gdt-CRF) to sensor selection from time-series binary location signals (captured by pyroelectric sensors) for human activity recognition at homes, which is a fundamental problem in smart home systems and home energy management systems. In comparison with forward greedy and L1-regularized CRFs (referred to as L1-CRF), FoBa-gdt-CRF requires the smallest number of sensors for achieving comparable recognition accuracy. Although this paper mainly focuses on the theoretical analysis for FoBa-obj and FoBa-gdt, we conduct additional experiments to study the behaviors of FoBa-obj and FoBa-gdt in the long version of this paper (Liu et al., 2013).

1.1. Notation

Denote $e_j \in \mathbb{R}^d$ as the j^{th} natural basis in the space $\mathbb{R}^d$. The set difference $A - B$ returns the elements that are in A but outside of B. Given any integer $s > 0$, the restricted strong convexity constants (RSCC) $\rho_-(s)$ and $\rho_+(s)$ are defined as follows: for any $\|t\|_0 \leq s$ and $t = \beta' - \beta$, we require

$$\frac{\rho_-(s)}{2}\|t\|^2 \leq Q(\beta') - Q(\beta) - \langle \nabla Q(\beta), t\rangle \leq \frac{\rho_+(s)}{2}\|t\|^2.$$

Similar definitions can be found in (Bahmani et al., 2011; Jalali et al., 2011; Negahban et al., 2010; Zhang, 2009). If the objective function takes the quadratic form $Q(\beta) = \frac{1}{2}\|X\beta - y\|^2$, then the above definition is equivalent to the restricted isometric property (RIP) (Candès & Tao, 2005):

$$\rho_-(s)\|t\|^2 \leq \|Xt\|^2 \leq \rho_+(s)\|t\|^2,$$

where the well known RIP constant can be defined as $\delta = \max\{1 - \rho_-(s),\ \rho_+(s) - 1\}$. To give tighter values for $\rho_+(.)$ and $\rho_-(.)$, we only require RSCC to hold for all $\beta \in \mathcal{D}_s := \{\|\beta\|_0 \leq s \mid Q(\beta) \leq Q(0)\}$ throughout this paper. Finally we define $\hat{\beta}(F)$ as $\hat{\beta}(F) := \arg\min_{\mathrm{supp}(\beta)\subset F} : Q(\beta)$. Note that the problem is convex as long as $Q(\beta)$ is a convex function. Denote $\bar{F} := \mathrm{supp}(\bar{\beta})$ and $\bar{k} := |\bar{F}|$.

We make use of order notation throughout this paper. If a and b are both positive quantities that depend on n or p, we write $a = O(b)$ if a can be bounded by a fixed multiple of b for all sufficiently large dimensions. We write $a = o(b)$ if for any positive constant $\phi > 0$, we have $a \leq \phi b$ for all sufficiently large dimensions. We write $a = \Omega(b)$ if both $a = O(b)$ and $b = O(a)$ hold.

Algorithm 1 FoBa (FoBa-obj | FoBa-gdt)

Require: $\delta > 0$ | $\epsilon > 0$
Ensure: $\beta^{(k)}$

Let $F^{(0)} = \emptyset$, $\beta^{(0)} = 0$, $k = 0$,
while TRUE **do**
  %% stopping determination
  if $Q(\beta^{(k)}) - \min_{\alpha, j \notin F^{(k)}} Q(\beta^{(k)} + \alpha e_j) < \delta$ | $\|\nabla Q(\beta^{(k)})\|_\infty < \epsilon$ **then**
    break
  end if
  %% forward step
  $i^{(k)} = \arg\min_{i \notin F^{(k)}} \{\min_\alpha Q(\beta^{(k)} + \alpha e_i)\}$ | $i^{(k)} = \arg\max_{i \notin F^{(k)}} : |\nabla Q(\beta^{(k)})_i|$
  $F^{(k+1)} = F^{(k)} \cup \{i^{(k)}\}$
  $\beta^{(k+1)} = \hat{\beta}(F^{(k+1)})$
  $\delta^{(k+1)} = Q(\beta^{(k)}) - Q(\beta^{(k+1)})$
  $k = k + 1$
  %% backward step
  while TRUE **do**
    if $\min_{i \in F^{(k+1)}} Q(\beta^{(k)} - \beta_i^{(k)} e_i) - Q(\beta^{(k)}) \geq \delta^{(k)}/2$ **then**
      break
    end if
    $i^{(k)} = \arg\min_i Q(\beta^{(k)} - \beta_i^{(k)} e_i)$
    $k = k - 1$
    $F^{(k)} = F^{(k+1)} - \{i^{(k+1)}\}$
    $\beta^{(k)} = \hat{\beta}(F^{(k)})$
  end while
end while

2. Related Work

Tropp (2004) investigated the behavior of the orthogonal matching pursuit (OMP) algorithm for the least square case, and proposed a sufficient condition (an ℓ_∞ type condition) for guaranteed feature selection consistency. Zhang (2009) generalized this analysis to the case of measurement noise. In statistics, OMP is known as boosting (Buhlmann, 2006) and similar ideas have been explored in Bayesian network learning (Chickering & Boutilier, 2002). Shalev-Shwartz et al. (2010) extended OMP to the general convex smooth function and studied the relationship between objective value reduction and output sparsity. Other greedy methods such as ROMP (Needell & Vershynin, 2009) and CoSaMp (Needell & Tropp, 2008) were studied and shown to have theoretical properties similar to those of OMP. Zhang (2011a) proposed a Forward-backward (FoBa) greedy algorithm for the least square case, which is an extension of OMP but has stronger theoretical guarantees as well as better empirical performance: feature selection consistency is guaranteed under the sparse eigenvalue condition, which is an ℓ_2 type condition weaker than the ℓ_∞ type condition. Note that if the data matrix is a Gaussian random matrix, the ℓ_2 type condition requires the measurements n to be of the order of $O(s \log d)$ where s is the sparsity of the true solution and d is the number of features, while the ℓ_∞ type condition requires $n = O(s^2 \log d)$; see (Zhang & Zhang, 2012; Liu et al., 2012). Jalali et al. (2011) and Johnson et al. (2012) extended the FoBa algorithm to general convex functions and applied it to sparse inverse covariance estimation problems.

Convex methods, such as LASSO (Zhao & Yu, 2006) and Dantzig selector (Candès & Tao, 2007), were proposed for sparse learning. The basic idea behind these methods is to use the ℓ_1-norm to approximate the ℓ_0-norm in order to transform problem (1) into a convex optimization problem. They usually require restricted conditions referred to as irrepresentable conditions (stronger than the RIP condition) for guaranteed feature selection consistency (Zhang, 2011a). A multi-stage procedure on LASSO and Dantzig selector (Liu et al., 2012) relaxes such condition, but it is still stronger than RIP.

3. The Gradient FoBa Algorithm

This section introduces the standard FoBa algorithm, that is, FoBa-obj, and its variant FoBa-gdt. Both algorithms start from an empty feature pool F and follow the same procedure in every iteration consisting of two steps: a forward step and a backward step. The forward step evaluates the "goodness" of all features outside of the current feature set F, selects the best feature to add to the current feature pool F, and then optimizes the corresponding coefficients of all features in the current feature pool F to obtain a new β. The elements of β in F are nonzero and the rest are zeros. The backward step *iteratively* evaluates the "badness" of all features outside of the current feature set F, removes "bad" features from the current feature pool F, and recomputes the optimal β over the current feature set F. Both algorithms use the same definition of "badness" for a feature: the increment of the objective after removing this feature. Specifically, for any features i in the current feature pool F, the "badness" is defined as $Q(\beta - \beta_i e_i) - Q(\beta)$, which is a positive number. It is worth to note that the forward step selects one and only one feature while the backward step may remove zero, one, or more features. Finally, both algo-

rithms terminate when no "good" feature can be identified in the forward step, that is, the "goodness" of all features outside of F is smaller than a threshold.

The main difference between FoBa-obj and FoBa-gdt lies in the definition of "goodness" in the forward step and their respective stopping criterion. FoBa-obj evaluates the goodness of a feature by its maximal reduction of the objective function. Specifically, the "goodness" of feature i is defined as $Q(\beta) - \min_\alpha Q(\beta + \alpha e_i)$ (a larger value indicates a better feature). This is a direct way to evaluate the "goodness" since our goal is to decrease the objective as much as possible under the cardinality condition. However, it may be computationally expensive since it requires solving a large number of one-dimensional optimization problems, which may or may not be solved in a closed form. To improve computational efficiency in such situations, FoBa-gdt uses the partial derivative of Q with respect to individual coordinates (features) as its "goodness' measure: specifically, the "goodness" of feature i is defined as $|\nabla Q(\beta)_i|$. Note that the two measures of "goodness" are always non-negative. If feature i is already in the current feature set F, its "goodness" score is always zero, no matter which measure to use. We summarize the details of FoBa-obj and FoBa-gdt in Algorithm 1: the plain texts correspond to the common part of both algorithms, and the ones with solid boxes and dash boxes correspond to their individual parts. The superscript (k) denotes the k^{th} iteration incremented/decremented in the forward/backward steps.

Gradient-based feature selection has been used in a forward greedy method (Zhang, 2011b). FoBa-gdt extends it to a Forward-backward procedure (we present a detailed theoretical analysis of it in the next section). The main workload in the forward step for FoBa-obj is on Step 4, whose complexity is $O(TD)$, where T represents the iterations needed to solve $\min_\alpha : \ Q(\beta^{(k)} + \alpha e_j)$ and D is the number of features outside of the current feature pool set $F^{(k)}$. In comparison, the complexity of Step 4 in FoBa is just $O(D)$. When T is large, we expect FoBa-gdt to be much more computationally efficient. The backward steps of both algorithms are identical. The computational costs of the backward step and the forward step are comparable in FoBa-gdt (but not FoBa-obj), because their main work loads are on Step 10 and Step 21 (both are solving $\hat{\beta}(.)$) respectively and the times of running Step 21 is always less than that of Step 10.

4. Theoretical Analysis

This section first gives the termination condition of Algorithms 1 with FoBa-obj and FoBa-gdt because the number of iterations directly affect the values of RSCC ($\rho_+(.)$, $\rho_-(.)$, and their ratio), which are the key factors in our main results. Then we discuss the values of RSCC in a class of commonly used machine learning objectives. Next we present the main results of this paper, including upper bounds on objective, estimation, and feature selection errors for both FoBa-obj and FoBa-gdt. We compare our results to those of existing analyses of FoBa-obj and show that our results fill the theoretical gap between the least square loss case and the general case.

4.1. Upper Bounds on Objective, Estimation, and Feature Selection Errors

We first study the termination conditions of FoBa-obj and FoBa-gdt, as summarized in Theorems 1 and 2 respectively.

Theorem 1. *Take* $\delta > \frac{4\rho_+(1)}{\rho_-(s)^2}\|\nabla Q(\bar{\beta})\|_\infty^2$ *in Algorithm 1 with FoBa-obj where* s *can be any positive integer satisfying* $s \leq n$ *and*

$$(s - \bar{k}) > (\bar{k} + 1)\left[\left(\sqrt{\frac{\rho_+(s)}{\rho_-(s)}} + 1\right)\frac{2\rho_+(1)}{\rho_-(s)}\right]^2. \quad (2)$$

Then the algorithm terminates at some $k \leq s - \bar{k}$.

Theorem 2. *Take* $\epsilon > \frac{2\sqrt{2}\rho_+(1)}{\rho_-(s)}\|\nabla Q(\bar{\beta})\|_\infty$ *in Algorithm 1 with FoBa-gdt, where* s *can be any positive integer satisfying* $s \leq n$ *and Eq.(2). Then the algorithm terminates at some* $k \leq s - \bar{k}$.

To simply the results, we first assume that the condition number $\kappa(s) := \rho_+(s)/\rho_-(s)$ is bounded (so is $\rho_+(1)/\rho_-(s)$ because of $\rho_+(s) \geq \rho_+(1)$). Then both FoBa-obj and FoBa-gdt terminate at some k proportional to the sparsity $\bar{k}$, similar to OMP (Zhang, 2011b) and FoBa-obj (Jalali et al., 2011; Zhang, 2011a). Note that the value of k in Algorithm 1 is exactly the cardinality of $F^{(k)}$ and the sparsity of $\beta^{(k)}$. Therefore, Theorems 1 and 2 imply that if $\kappa(s)$ is bounded, FoBa-obj and FoBa-gdt will output a solution with sparsity proportional to that of the true solution $\bar{\beta}$.

Most existing works simply assume that $\kappa(s)$ is bounded or have similar assumptions. We make our analysis more complete by discussing the values of $\rho_+(s)$, $\rho_-(s)$, and their ratio $\kappa(s)$. Apparently, if $Q(\beta)$ is strongly convex and Lipschitzian, then $\rho_-(s)$ is bounded from below and $\rho_+(s)$ is bounded from above, thus restricting the ratio $\kappa(s)$. To see that $\rho^+(s)$, $\rho^-(s)$, and $\kappa(s)$ may still be bounded under milder conditions, we consider a common structure for $Q(\beta)$ used in many machine learning formulations:

$$Q(\beta) = \frac{1}{n}\sum_{i=1}^{n} l_i(X_{i.}\beta, y_i) + R(\beta) \quad (3)$$

where $(X_{i.}, y_i)$ is the i^{th} training sample with $X_{i.} \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$, $l_i(.,.)$ is convex with respect to the first argument and could be different for different i, and both $l_i(.,.)$

and $R(.)$ are twice differentiable functions. $l_i(.,.)$ is typically the loss function, e.g., the quadratic loss $l_i(u,v) = (u-v)^2$ in regression problems and the logistic loss $l_i(u,v) = \log(1+\exp\{-uv\})$ in classification problems. $R(\beta)$ is typically the regularization, e.g., $R(\beta) = \frac{\mu}{2}\|\beta\|^2$.

Theorem 3. *Let s be a positive integer less than n, and λ^-, λ^+, λ_R^-, and λ_R^+ be positive numbers satisfying*

$$\lambda^- \leq \nabla_1^2 l_i(X_{i.}\beta, y_i) \leq \lambda^+, \quad \lambda_R^- I \preceq \nabla^2 R(\beta) \preceq \lambda_R^+ I$$

($\nabla_1^2 l_i(.,.)$ is the second derivative with respect to the first argument) for any i and $\beta \in \mathcal{D}_s$. Assume that $\lambda_R^- + 0.5\lambda^- > 0$ and the sample matrix $X \in \mathbb{R}^{n\times d}$ has independent sub-Gaussian isotropic random rows or columns (in the case of columns, all columns should also satisfy $\|X_{.j}\| = \sqrt{n}$). If the number of samples satisfies $n \geq Cs\log d$, then

$$\rho_+(s) \leq \lambda_R^+ + 1.5\lambda^+ \tag{4a}$$

$$\rho_-(s) \geq \lambda_R^- + 0.5\lambda^- \tag{4b}$$

$$\kappa(s) \leq \frac{\lambda_R^+ + 1.5\lambda^+}{\lambda_R^- + 0.5\lambda^-} =: \kappa \tag{4c}$$

hold with high probability[1], where C is a fixed constant. Furthermore, define $\bar{k}$, $\bar{\beta}$, and δ (or ϵ) in Algorithm 1 with FoBa-obj (or FoBa-gdt) as in Theorem 1 (or Theorem 2). Let

$$s = \bar{k} + 4\kappa^2(\sqrt{\kappa}+1)^2(\bar{k}+1) \tag{5}$$

and $n \geq Cs\log d$. We have that s satisfies (2) and Algorithm 1 with FoBa-obj (or FoBa-gdt) terminates within at most $4\kappa^2(\sqrt{\kappa}+1)^2(\bar{k}+1)$ iterations with high probability.

Roughly speaking, if the number of training samples is large enough, i.e., $n \geq \Omega(\bar{k}\log d)$ (actually it could be much smaller than the dimension d of the train data), we have the following with high probability: Algorithm 1 with FoBa-obj or FoBa-gdt outputs a solution with sparsity at most $\Omega(\bar{k})$ (this result will be improved when the nonzero elements of $\bar{\beta}$ are strong enough, as shown in Theorems 4 and 5); s is bounded by $\Omega(\bar{k})$; and $\rho_+(s)$, $\rho_-(s)$, and $\kappa(s)$ are bounded by constants. One important assumption is that the sample matrix X has independent sub-Gaussian isotropic random rows or columns. In fact, this assumption is satisfied by many natural examples, including Gaussian and Bernoulli matrices, general bounded random matrices whose entries are independent bounded random variables with zero mean and unit variances. Note that from the definition of "sub-Gaussian isotropic random vectors" (Vershynin, 2011, Definitions 19 and 22), it even allows the dependence within rows or columns but not both. Another important assumption is $\lambda_R^- + 0.5\lambda^- > 0$, which means that either λ_R^- or λ^- is positive (both of them are nonnegative from the convexity assumption). We can simply verify that (i) for the quadratic case $Q(\beta) = \frac{1}{n}\sum_{i=1}^n (X_{i.}\beta - y_i)^2$, we have $\lambda^- = 1$ and $\lambda_R^- = 0$; (ii) for the logistic case with bounded data matrix X, that is $Q(\beta) = \frac{1}{n}\sum_{i=1}^n \log(1+\exp\{-X_{i.}\beta y_i\}) + \frac{\mu}{2}\|\beta\|^2$, we have $\lambda_R^- = \mu > 0$ and $\lambda^- > 0$ because $\mathcal{D}_s$ is bounded in this case.

[1] "With high probability" means that the probability converges to 1 with the problem size approaching to infinity.

Now we are ready to present the main results: the upper bounds of estimation error, objective error, and feature selection error for both algorithms. $\rho_+(s)$, $\rho_+(1)$, and $\rho_-(s)$ are involved in all bounds below. One can simply treat them as constants in understanding the following results, since we are mainly interested in the scenario when the number of training samples is large enough. We omit proofs due to space limitations (the proofs are provided in the long version of this paper (Liu et al., 2013)). The main results for FoBa-obj and FoBa-gdt are presented in Theorems 4 and 5 respectively.

Theorem 4. *Let s be any number that satisfies (2) and choose δ as in Theorem 1 for Algorithm 1 with FoBa-obj. Consider the output $\beta^{(k)}$ and its support set $F^{(k)}$. We have*

$$\|\beta^{(k)} - \bar{\beta}\|^2 \leq \frac{16\rho_+^2(1)\delta}{\rho_-^2(s)}\bar{\Delta},$$

$$Q(\beta^{(k)}) - Q(\bar{\beta}) \leq \frac{2\rho_+(1)\delta}{\rho_-(s)}\bar{\Delta},$$

$$\frac{\rho_-(s)^2}{8\rho_+(1)^2}|F^{(k)} - \bar{F}| \leq |\bar{F} - F^{(k)}| \leq 2\bar{\Delta},$$

where $\gamma = \frac{4\sqrt{\rho_+(1)\delta}}{\rho_-(s)}$ and $\bar{\Delta} := |\{j \in \bar{F} - F^{(k)} : |\bar{\beta}_j| < \gamma\}|$.

Theorem 5. *Let s be any number that satisfies (2) and choose ϵ as in Theorem 2 for Algorithm 1 with FoBa-gdt. Consider the output $\beta^{(k)}$ and its support set $F^{(k)}$. We have*

$$\|\beta^{(k)} - \bar{\beta}\|^2 \leq \frac{8\epsilon^2}{\rho_-^2(s)}\bar{\Delta},$$

$$Q(\beta^{(k)}) - Q(\bar{\beta}) \leq \frac{\epsilon^2}{\rho_-(s)}\bar{\Delta},$$

$$\frac{\rho_-(s)^2}{8\rho_+(1)^2}|F^{(k)} - \bar{F}| \leq |\bar{F} - F^{(k)}| \leq 2\bar{\Delta},$$

where $\gamma = \frac{2\sqrt{2}\epsilon}{\rho_-(s)}$ and $\bar{\Delta} := |\{j \in \bar{F} - F^{(k)} : |\bar{\beta}_j| < \gamma\}|$.

Although FoBa-obj and FoBa-gdt use different criteria to evaluate the "goodness" of each feature, they actually guarantee the same properties. Choose ϵ^2 and δ in the order of $\Omega(\|\nabla Q(\bar{\beta})\|_\infty^2)$. For both algorithms, we have that the estimation error $\|\beta^{(k)} - \bar{\beta}\|^2$ and the objective error $Q(\beta^{(k)}) - Q(\bar{\beta})$ are bounded by $\Omega(\bar{\Delta}\|\nabla Q(\bar{\beta})\|_\infty^2)$, and the feature selection errors $|F^{(k)} - \bar{F}|$ and $|\bar{F} - F^{(k)}|$ are bounded by $\Omega(\bar{\Delta})$. $\|\nabla Q(\bar{\beta})\|_\infty$ and $\bar{\Delta}$ are two key factors

in these bounds. $\|\nabla Q(\bar{\beta})\|_\infty$ roughly represents the noise level[2]. $\bar{\Delta}$ defines the number of weak channels of the true solution $\bar{\beta}$ in $\bar{F}$. One can see that if all channels of $\bar{\beta}$ on $\bar{F}$ are strong enough, that is, $|\bar{\beta}_j| > \Omega(\|\nabla Q(\bar{\beta})\|_\infty)\ \forall j \in \bar{F}$, $\bar{\Delta}$ turns out to be 0. In other words, all errors (estimation error, objective error, and feature selection error) become 0, when the signal noise ratio is big enough. Note that under this condition, the original NP hard problem (1) is solved exactly, which is summarized in the following corollary:

Corollary 1. *Let s be any number that satisfies (2) and choose δ (or ϵ) as in Theorem 1 (or 2) for Algorithm 1 with FoBa-gdt (or FoBa-obj). If*

$$\frac{|\bar{\beta}_j|}{\|\nabla Q(\bar{\beta})\|_\infty} \geq \frac{8\rho_+(1)}{\rho_-^2(s)} \quad \forall j \in \bar{F},$$

then problem (1) can be solved exactly.

One may argue that since it is difficult to set δ or ϵ, it is still hard to solve (1). In practice, one does not have to set δ or ϵ and only needs to run Algorithm 1 without checking the stopping condition until all features are selected. Then the most recent $\beta^{(\bar{k})}$ gives the solution to (1).

4.2. Comparison for the General Convex Case

Jalali et al. (2011) analyzed FoBa-obj for general convex smooth functions and here we compare our results to theirs. They chose the true model β^* as the target rather than the true solution $\bar{\beta}$. In order to simplify the comparison, we assume that the distance between the true solution and the true model is not too great[3], that is, we have $\beta^* \approx \bar{\beta}$, $\text{supp}(\beta^*) = \text{supp}(\bar{\beta})$, and $\|\nabla Q(\beta^*)\|_\infty \approx \|\nabla Q(\bar{\beta})\|_\infty$. We compare our results from Section 4.1 and the results in (Jalali et al., 2011). In the estimation error comparison, we have from our results:

$$\begin{aligned}&\|\beta^{(k)} - \beta^*\| \approx \|\beta^{(k)} - \bar{\beta}\| \\ \leq&\Omega(\bar{\Delta}^{1/2}\|\nabla Q(\bar{\beta})\|_\infty) \approx \Omega(\bar{\Delta}^{1/2}\|\nabla Q(\beta^*)\|_\infty)\end{aligned}$$

and from the results in (Jalali et al., 2011): $\|\beta^{(k)} - \beta^*\| \leq \Omega(\bar{k}\|\nabla Q(\beta^*)\|_\infty)$. Note that $\Delta^{1/2} \leq \bar{k}^{1/2} \ll \bar{k}$. Therefore, under our assumptions with respect to β^* and $\bar{\beta}$, our analysis gives a tighter bound. Notably, when there are a large number of strong channels in $\bar{\beta}$ (or approximately β^*), we will have $\bar{\Delta} \ll \bar{k}$.

Let us next consider the condition required for feature selection consistency, that is, $\text{supp}(\text{F}^{(\text{k})}) = \text{supp}(\bar{\text{F}}) = \text{supp}(\beta^*)$. We have from our results:

$$\|\bar{\beta}_j\| \geq \Omega(\|\nabla Q(\bar{\beta})\|_\infty)\ \forall j \in \text{supp}(\beta^*)$$

and from the results in (Jalali et al., 2011):

$$\|\beta^*_j\| \geq \Omega(\bar{k}\|\nabla Q(\beta^*)\|_\infty)\ \forall j \in \text{supp}(\beta^*).$$

When $\beta^* \approx \bar{\beta}$ and $\|\nabla Q(\beta^*)\|_\infty \approx \|\nabla Q(\bar{\beta})\|_\infty$, our results guarantee feature selection consistency under a weaker condition.

4.3. A Special Case: Least Square Loss

We next consider the least square case: $Q(\beta) = \frac{1}{2}\|X\beta - y\|^2$ and shows that our analysis for the two algorithms in Section 4.1 fills in a theoretical gap between this special case and the general convex smooth case.

Following previous studies (Candès & Tao, 2007; Zhang, 2011b; Zhao & Yu, 2006), we assume that $y = X\beta^* + \varepsilon$ where the entries in ε are independent random sub-gaussian variables, β^* is the true model with the support set $\bar{F}$ and the sparsity number $\bar{k} := |\bar{F}|$, and $X \in \mathbb{R}^{n\times d}$ is normalized as $\|X_{\cdot i}\|^2 = 1$ for all columns $i = 1, \cdots, d$. We then have following inequalities with high probability (Zhang, 2009):

$$\|\nabla Q(\beta^*)\|_\infty = \|X^T\varepsilon\|_\infty \leq \Omega(\sqrt{n^{-1}\log d}), \quad (6)$$

$$\|\nabla Q(\bar{\beta})\|_\infty \leq \Omega(\sqrt{n^{-1}\log d}), \quad (7)$$

$$\|\bar{\beta} - \beta^*\|_2 \leq \Omega(\sqrt{n^{-1}\bar{k}}), \quad (8)$$

$$\|\bar{\beta} - \beta^*\|_\infty \leq \Omega(\sqrt{n^{-1}\log \bar{k}}), \quad (9)$$

implying that $\bar{\beta}$ and β^* are quite close when the true model is really sparse, that is, when $\bar{k} \ll n$.

An analysis for FoBa-obj in the least square case (Zhang, 2011a) has indicated that the following estimation error bound holds with high probability:

$$\begin{aligned}&\|\beta^{(k)} - \beta^*\|^2 \leq \Omega(n^{-1}(\bar{k}+ \\ &\log d|\{j \in \bar{F} :\ |\beta^*_j| \leq \Omega(\sqrt{n^{-1}\log d})\}|)) \quad (10)\end{aligned}$$

as well as the following condition for feature selection consistency: if $|\beta^*_j| \geq \Omega(\sqrt{n^{-1}\log d})\ \forall j \in \bar{F}$, then

$$\text{supp}(\beta^{(k)}) = \text{supp}(\beta^*) \quad (11)$$

Applying the analysis for general convex smooth cases in (Jalali et al., 2011) to the least square case, one obtains the following estimation error bound from Eq. (6)

$$\|\beta^{(k)} - \beta^*\|^2 \leq \Omega(\bar{k}^2\|\nabla Q(\beta^*)\|^2_\infty) \leq \Omega(n^{-1}\bar{k}^2\log d)$$

and the following condition of feature selection consistency: if $|\beta^*_j| \geq \Omega(\sqrt{\bar{k}n^{-1}\log d})\ \forall\ j \in \bar{F}$, then

$$\text{supp}(\beta^{(\text{k})}) = \text{supp}(\beta^*).$$

[2] To see this, we can consider the least square case (with standard noise assumption and each column of the measurement matrix $X \in \mathbb{R}^{n\times d}$ is normalized to 1): $\|\nabla Q(\bar{\beta})\|_\infty \leq \Omega(\sqrt{n^{-1}\log d}\sigma)$ holds with high probability, where σ is the standard derivation.

[3] This assumption is not absolutely fair, but holds in many cases, such as in the least square case, which will be made clear in Section 4.3.

One can observe that the general analysis gives a looser bound for estimation error and requires a stronger condition for feature selection consistency than the analysis for the special case.

Our results in Theorems 4 and 5 bridge this gap when combined with Eqs. (8) and (9). The first inequalities in Theorems 4 and 5 indicate that

$$\begin{aligned}
&\|\beta^{(k)} - \beta^*\|^2 \le (\|\beta^{(k)} - \bar{\beta}\| + \|\bar{\beta} - \beta^*\|)^2 \\
\le &\Omega\Big(n^{-1}(\bar{k} + \log d \,|\, \{j \in \bar{F} - F^{(k)} : \\
&|\bar{\beta}_j| < \Omega(n^{-1/2}\sqrt{\log d})\}|)\Big) \quad \text{[from Eq. (8)]} \\
\le &\Omega\Big(n^{-1}(\bar{k} + \log d \,|\, \{j \in \bar{F} - F^{(k)} : \\
&|\beta^*_j| < \Omega(n^{-1/2}\sqrt{\log d})\}|)\Big) \quad \text{[from Eq. (9)]}
\end{aligned}$$

which is consistent with the results in Eq. (10). The last inequality in Theorem 5 also implies that feature selection consistency is guaranteed as well, as long as $|\bar{\beta}_j| > \Omega(\sqrt{n^{-1}\log d})$ (or $|\beta^*_j| > \Omega(\sqrt{n^{-1}\log d})$) for all $j \in \bar{F}$. This requirement agrees with the results in Eq. (11).

5. Application: Sensor Selection for Human Activity Recognition

Machine learning technologies for smart home systems and home energy management systems have recently attracted much attention. Among the many promising applications such as optimal energy control, emergency alerts for elderly persons living alone, and automatic life-logging, a fundamental challenge for these applications is to recognize human activity at homes, with the smallest number of sensors. The data mining task here is to minimize the number of sensors without significantly worsening recognition accuracy. We used pyroelectric sensors, which return binary signals in reaction to human motion.

Fig. 1 shows our experimental room layout and sensor locations. The numbers represent sensors, and the ellipsoid around each represents the area covered by it. We used 40 sensors, i.e., we observe a 40-dimensional binary time series. A single person lives in the room for roughly one month, and data is collected on the basis of manually tagging his activities into the pre-determined 14 categories summarized in Table 5. For data preparation reasons, we use the first 20% (roughly one week) samples in the data, and divide it into 10% for training and 10% for testing. The numbers of training and test samples are given in Table 1.

Pyroelectric sensors are preferable over cameras for two practical reasons: cameras tend to create a psychological barrier and pyroelectric sensors are much cheaper and easier to implement at homes. Such sensors only observe

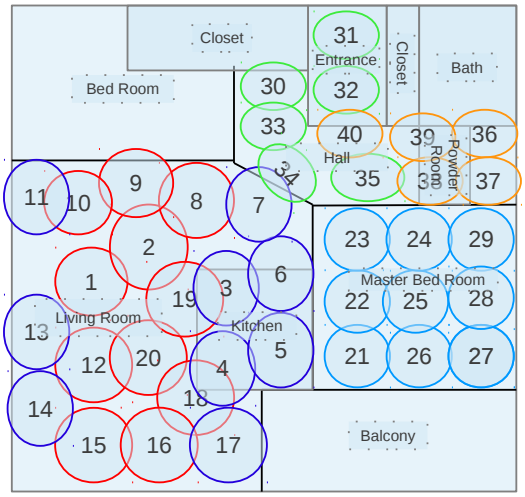

Figure 1. Room layout and sensor locations.

noisy binary location information. This means that, for high recognition accuracy, history (sequence) information must be taken into account. The binary time series data follows a linear-chain conditional random field (CRF) (Lafferty et al., 2001; Sutton & McCallum, 2006). Linear-chain CRF gives a smooth and convex loss function; see the long version of this paper (Liu et al., 2013) for more details of CRF.

Our task then is sensor selection on the basis of noisy binary time series data, and to do this we apply our FoBa-gdt-CRF (FoBa-gdt with CRF objective function). Since it is very expensive to evaluate the CRF objective value and its gradient, FoBa-obj becomes impractical in this case (a large number of optimization problems in the forward step make it computationally very expensive). Here, we consider a sensor to have been "used" if at least one feature related to it is used in the CRF. Note that we have 14 activity-signal binary features (i.e., indicators of sensor/activity simultaneous activations) for each single sensor, and therefore we have $40 \times 14 = 560$ such features in total. In addition, we have $14 \times 14 = 196$ activity-activity binary features (i.e., indicators of the activities at times $t-1$ and t). Here we only enforced sparsity on the first type of features.

First we compare FoBa-gdt-CRF with Forward-gdt-CRF (Forward-gdt with CRF loss function) and L1-CRF[4] in terms of test recognition error over the number of sensors selected (see the top of Fig. 2). We can observe that

- The performance for all methods improves when the um-

[4]L1-CRF solves the optimization problem with CRF loss + L1 regularization. Since it is difficult to search the whole L1 regularization parameter value space, we investigated a number of discrete values.

Table 1. Activities in the sensor data set

ID	Activity	train / test samples
1	Sleeping	81K / 87K
2	Out of Home (OH)	66K / 42K
3	Using Computer	64K / 46K
4	Relaxing	25K / 65K
5	Eating	6.4K / 6.0K
6	Cooking	5.2K / 4.6K
7	Showering (Bathing)	3.9K / 45.0K
8	No Event	3.4K / 3.5K
9	Using Toilet	2.5K / 2.6K
10	Hygiene (brushing teeth, etc.)	1.6K / 1.6K
11	Dishwashing	1.5K /1.8K
12	Beverage Preparation	1.4K / 1.4K
13	Bath Cleaning/Preparation	0.5K / 0.3K
14	Others	6.5K / 2.1K
Total	-	270K / 270K

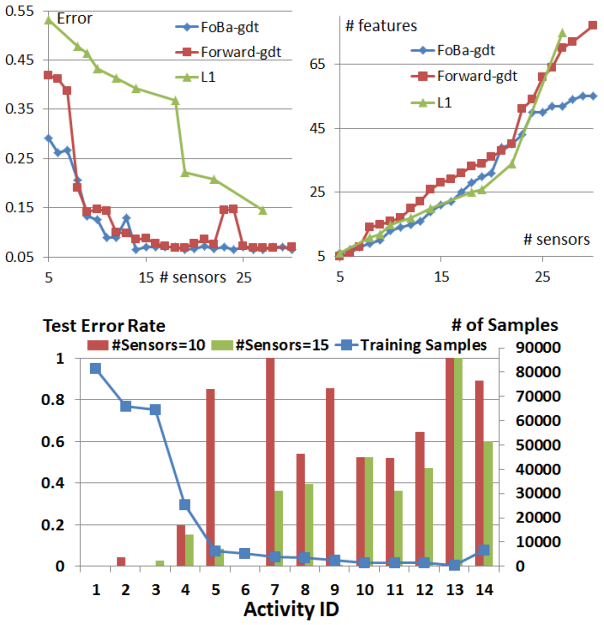

Figure 2. Top: comparisons of FoBa-gdt-CRF, Forward-gdt-CRF and L1-CRF. Bottom: test error rates (FoBa-gdt-CRF) for individual activities.

ber of sensors increases.

- FoBa-gdt-CRF and Forward-gdt-CRF achieve comparable performance. However, FoBa-gdt-CRF reduces the error rate slightly faster, in terms of the number of sensors.
- FoBa-gdt-CRF achieves its best performance with 14-15 sensors while Forward-gdt-CRF needs 17-18 sensors to achieve the same error level. We obtain sufficient accuracy by using fewer than 40 sensors.
- FoBa-gdt-CRF consistently requires fewer features than Forward-gdt-CRF to achieve the same error level when using the same number of sensors.

We also analyze the test error rates of FoBa-gdt-CRF for individual activities. We consider two cases with the number of sensors being 10 and 15, and report their test error rates for each individual activity in the bottom of Fig. 2. We observe that:

- The high frequency activities (e.g., activities $\{1,2,3,4\}$) are well recognized in both cases. In other words, FoBa-gdt-CRF is likely to select sensors (features) which contribute to the discrimination of high frequency activities.
- The error rates for activities $\{5, 7, 9\}$ significantly improve when the number of sensors increases from 10 to 15. Activities 7 and 9 are *Showering* and *Using Toilet*, and the use of additional sensors $\{36, 37, 40\}$ seems to have contributed to this improvement. Also, a dinner table was located near sensor 2, which is why the error rate w.r.t. activity 5 (Eating) significantly decreases from the case # of sensors=10 to # of sensors=15 by including sensor 2.

Table 2. Sensor IDs selected by FoBa-gdt-CRF.

# of sensors=10	$\{1, 4, 5, 9, 10, 13, 19, 28, 34, 38\}$
# of sensors=15	$\{$# of sensors=10$\}$ + $\{2, 7, 36, 37, 40\}$

6. Conclusion

This paper considers two forward-backward greedy methods, including a state-of-the-art greedy method FoBa-obj and its variant FoBa-gdt which is more efficient than FoBa-obj, for solving sparse feature selection problems with general convex smooth functions. We systematically analyze the theoretical properties of both algorithms. Our main contributions include: (i) We derive better theoretical bounds for FoBa-obj and FoBa-gdt than existing analyses (Jalali et al., 2011) for general smooth convex functions. Our result also suggests that the NP hard problem (1) can be solved by FoBa-obj and FoBa-gdt if the signal noise ratio is big enough; (ii) Our new bounds are consistent with the bounds of a special case (least squares) (Zhang, 2011a) and fill a previously existing theoretical gap for general convex smooth functions (Jalali et al., 2011); (iii) We provide the condition to satisfy the restricted strong convexity condition in commonly used machine learning problems; (iv) We apply FoBa-gdt (with the conditional random field objective) to the sensor selection problem for human indoor activity recognition and our results show that FoBa-gdt can successfully remove unnecessary sensors and is able to select more valuable sensors than other methods (including the ones based on forward greedy selection and L1-regularization). In the future work, we plan to extend FoBa algorithms to minimize a general convex smooth function over a low rank constraint.

7. Acknowledgements

We would like to sincerely thank Professor Masamichi Shimosaka of the University of Tokyo for providing sensor data collected in his research and Professor Stephen Wright of the University of Wisconsin-Madison for constructive comments and helpful advice. The majority of the work reported here was done during the internship of the first author at NEC Laboratories America, Cupertino, CA.

References

Bahmani, S., Boufounos, P., and Raj, B. Greedy sparsity-constrained optimization. *ASILOMAR*, pp. 1148–1152, 2011.

Buhlmann, P. Boosting for high-dimensional linear models. *Annals of Statistics*, 34:559–583, 2006.

Candès, E. J. and Tao, T. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51 (12):4203–4215, 2005.

Candès, E. J. and Tao, T. The Dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics*, 35(6):2313–2351, 2007.

Chickering, D. M. and Boutilier, C. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, 2002.

Jalali, A., Johnson, C. C., and Ravikumar, P. D. On learning discrete graphical models using greedy methods. *NIPS*, 2011.

Johnson, C. C., Jalali, A., and Ravikumar, P. D. High-dimensional sparse inverse covariance estimation using greedy methods. *Journal of Machine Learning Research - Proceedings Track*, 22:574–582, 2012.

Kleinbaum, D. G. and Klein, M. Logistic regression. *Statistics for Biology and Health*, pp. 103–127, 2010.

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *ICML*, pp. 282–289, 2001.

Liu, J., Wonka, P., and Ye, J. A multi-stage framework for dantzig selector and LASSO. *Journal of Machine Learning Research*, 13:1189–1219, 2012.

Liu, J., Fujimaki, R., and Ye, J. Forward-backward greedy algorithms for general convex smooth functions over a cardinality constraint. *arXiv:1401.0086*, 2013.

Needell, D. and Tropp, J. A. CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis*, 26:301–321, 2008.

Needell, D. and Vershynin, R. Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit. *Foundations of Computational Mathematics*, 9(3):317–334, 2009.

Negahban, S., Ravikumar, P. D., Wainwright, M. J., and Yu, B. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *CoRR*, abs/1010. 2731, 2010.

Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

Shalev-Shwartz, S., Srebro, N., and Zhang, T. Trading accuracy for sparsity in optimization problems with sparsity constraints. *SIAM Journal on Optimization*, 20(6): 2807–2832, 2010.

Sutton, C. and McCallum, A. An introduction to conditional random fields for relational learning. *Introduction to Statistical Relational Learning*, pp. 93–128, 2006.

Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1994.

Tropp, J. A. Greed is good: algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242, 2004.

Vershynin, R. Introduction to the non-asymptotic analysis of random matrices. *arXiv:1011.3027*, 2011.

Zhang, C.-H. and Zhang, T. A general theory of concave regularization for high dimensional sparse estimation problems. *Statistical Science*, 27(4), 2012.

Zhang, T. On the consistency of feature selection using greedy least squares regression. *Journal of Machine Learning Research*, 10:555–568, 2009.

Zhang, T. Adaptive forward-backward greedy algorithm for learning sparse representations. *IEEE Transactions on Information Theory*, 57(7):4689–4708, 2011a.

Zhang, T. Sparse recovery with orthogonal matching pursuit under RIP. *IEEE Transactions on Information Theory*, 57(9):6215–6221, 2011b.

Zhao, P. and Yu, B. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006.