Forward Basis Selection for Sparse Approximation over Dictionary

Xiao-Tong Yuan
Department of Statistics
Rutgers University
xyuan@stat.rutgers.edu

Shuicheng Yan
ECE Department
National University of Singapore
eleyans@nus.edu.sg

Abstract

Recently, forward greedy selection method has been successfully applied to approximately solve sparse learning problems, characterized by a trade-off between sparsity and accuracy. In this paper, we generalize this method to the setup of sparse approximation over a pre-fixed dictionary. A fully corrective forward selection algorithm is proposed along with convergence analysis. The per-iteration computational overhead of the proposed algorithm is dominated by a subproblem of linear optimization over the dictionary and a subproblem to optimally adjust the aggregation weights. The former is cheaper in several applications than the Euclidean projection while the latter is typically an unconstrained optimization problem which is relatively easy to solve. Furthermore, we extend the proposed algorithm to the setting of non-negative/convex sparse approximation over a dictionary. Applications of our algorithms to several concrete learning problems are explored with efficiency validated on benchmark data sets.

1 Introduction

We consider in this paper the sparse learning problem where the target solution can potentially be approximated by a solution that admits a sparse representation in a give dictionary. Among others, several examples falling inside this model include: 1) Coordinatewise sparse learning where the optimal solution is expected to be a sparse combination of canonical basis vectors, 2) low rank matrix approximation where the target solution is expected to be the weighted sum of a few rank-1 matrices in the form of outer product of unit-norm vectors, and 3) boosting classification where strong classifier is a linear combination of several weak learners. Formally, this class of problems can be unified inside the following framework of sparse approximation over a dictionary V in a Euclidean space $\mathcal{E}$:

$$\min_{x\in\mathcal{E}} f(x), \quad \text{s.t. } x \in \mathcal{L}_K(V), \tag{1}$$

where f is assumed a real valued differentiable convex function and

$$\mathcal{L}_K(V) := \bigcup_{U\subseteq V, |U|\leq K} \left\{ \sum_{u\in U} \alpha_u u : \alpha_u \in \mathbb{R} \right\} \tag{2}$$

is the union of the linear hulls spanned by those subsets $U \subseteq V$ with cardinality $|U| \leq K$. Here we allow the dictionary V to be finite or infinite. In the aforementioned examples, V is the canonical basis vectors in coordinatewise sparse learning (finite), a certain family of rank-1 matrices in low-rank matrix approximation (infinite), and a set of weak classifiers in boosting (finite or infinite).

Due to the cardinality constraint, problem (1) is non-convex and thus we resort to approximation algorithms for solution. Recently, a sparse approximation algorithm known as Fully Corrective Forward Greedy Selection (FCFGS) (Shalev-Shwartz et al., 2010) has been proposed for coordinatewise sparse learning, then extended to low rank matrix learning (Shalev-Shwartz et al., 2011) and decision tree boosting (Johnson & Zhang, 2011). Theoretical analysis (Shalev-Shwartz et al., 2010; Zhang, 2011) and strong numerical evidences (Shalev-Shwartz et al., 2011; Johnson & Zhang, 2011) show that FCFGS is more appealing, both in sparsity and accuracy, than traditional forward selection algorithms such as sequential greedy approximation (Zhang, 2003) and gradient boosting (Friedman, 2001).

In this paper, we propose a generic forward selection algorithm, namely forward basis selection (FBS),

Appearing in Proceedings of the 15^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2012, La Palma, Canary Islands. Volume 22 of JMLR: W&CP 22.

which generalizes FCFGS to approximately solve problem (1). One important property of FBS is that it will automatically select out a group of bases in the dictionary for sparse representation. The $O(\frac{1}{\epsilon})$ rate of convergence is established for FBS under mild assumptions on dictionary and objective function. When dictionary is finite, a better $O(\ln \frac{1}{\epsilon})$ geometric rate bound can be obtained under proper conditions. We then extend the FBS to non-negative sparse approximation and convex sparse approximation which to our knowledge has not been explicitly addressed in the existing literatures on fully-corrective-type forward selection methods. Such extensions facilitate the applications of FBS to positive semi-definite matrix learning and more general convex constrained sparse learning problems. The convergence properties are analyzed for both extensions. On iterate complexity, the per-iteration computational overhead of FBS is dominated by a linear gradient projection and a subproblem to optimally adjust the aggregation weights of bases. The former is significantly cheaper in several applications than the Euclidean projection used in projected-gradient-type methods. The latter is typically of limited size and thus is relatively easy to solve. We study the applications of the proposed method and its variants in several concrete sparse learning problems, and evaluate the performances on several benchmarks. Before proceeding, we establish the notation to be used in the rest of this paper.

1.1 Notation

We denote $\langle \cdot, \cdot \rangle$ the linear product and $\| \cdot \| = \sqrt{\langle \cdot, \cdot \rangle}$ the Euclidean norm. For a vector x, we denote $\|x\|_1$ its ℓ_1-norm, $\|x\|_0$ the number of non-zero components, and $\mathrm{supp}(x)$ the indices of non-zero components. The *linear hull* of dictionary V is given by

$$\mathcal{L}(V) := \left\{ \sum_{v \in V} \alpha_v v : \alpha_v \in \mathbb{R} \right\}.$$

We say f has *restricted strong convexity* and *restricted strong smoothness* over $\mathcal{L}(V)$ at sparsity level k, if there exists positive constants $\rho_+(k)$ and $\rho_-(k)$ such that for any $x, x' \in \mathcal{L}(V)$ and $x - x' \in \mathcal{L}_k(V)$,

$$f(x') - f(x) - \langle \nabla f(x), x' - x \rangle \leq \frac{\rho_+(k)}{2} \|x - x'\|^2, \quad (3)$$

and

$$f(x') - f(x) - \langle \nabla f(x), x' - x \rangle \geq \frac{\rho_-(k)}{2} \|x - x'\|^2. \quad (4)$$

We say dictionary V is bounded with radius A if $\forall v \in V, \|v\| \leq A$. We say V is symmetric if $v \in V$ implies $-v \in V$.

In the next subsection, we briefly review the FCFGS algorithm for coordinatewise sparse learning which motivates our study.

1.2 Fully Corrective Forward Greedy Selection

The FCFGS (Shalev-Shwartz et al., 2010) as described in Algorithm 1 was originally proposed to solve the following coordinatewise sparse learning problem,

$$\min_{x \in \mathbb{R}^d} f(x), \quad \text{s.t. } \|x\|_0 \leq K. \quad (5)$$

At each iterate, FCFGS first selects a coordinate at which the gradient has the largest absolute value, and then adjust the coefficients on the coordinates selected so far to minimize f. The algorithm is demonstrated to be a fast and accurate sparse approximation method in both theory (Shalev-Shwartz et al., 2010; Zhang, 2011) and practice (Shalev-Shwartz et al., 2011; Johnson & Zhang, 2011). More precisely, there are two appealing aspects of FCFGS:

- Orthogonal coordinates selection: Provided that the gradient $\nabla f(x^{(k-1)})$ is nonzero, the algorithm always selects a new coordinate $j^{(k)}$ at iteration.
- Geometric rate of convergence: It is shown in (Shalev-Shwartz et al., 2010, Theorem 2.8) that FCFBS can achieve geometric rate of convergence under restricted strongly convex/smooth assumptions on f.

FCFGS is in spirit identical to the well recognized Orthogonal Matching Pursuit algorithm (Pati et al., 1993; Tropp & Gilbert, 2007) in signal processing society.

Algorithm 1: Fully Corrective Forward Greedy Selection (FCFGS) (Shalev-Shwartz et al., 2010).

Initialization: $x^{(0)} = 0$, $F^{(0)} = \emptyset$.
Output: $x^{(K)}$.
for $k = 1, ..., K$ **do**
  Calculate

$$j^{(k)} = \arg\max_{j \in \{1,...,d\}} |[\nabla f(x^{(k-1)})]_j|. \quad (6)$$

  Set $F^{(k)} = F^{(k-1)} \cup \{j^{(k)}\}$ and update

$$x^{(k)} = \arg\min_{\mathrm{supp}(x) \subseteq F^{(k)}} f(x). \quad (7)$$

end

This paper proceeds as follows: We present in Section 2 the FBS algorithm along with convergence analysis. Two extensions of FBS are given and analyzed in Section 3. Several applications of FBS are studied in Section 4 and the related work is reviewed in Section 5. Experiments on real data are reported in Section 6. We conclude this work in Section 7.

2 Forward Basis Selection

The Forward Basis Selection (FBS) method is formally given in Algorithm 2. The working procedure is as follows: At each time instance k, we first search for a steepest descent direction $u^{(k)} \in V$ which solves the linear projection subproblem (8). Then the current iterate $x^{(k)}$ is updated via optimizing the subproblem (9) over the linear hull of the descent directions selected so far. Essentially, the subproblem (9) is an unconstrained convex optimization problem which can be efficiently optimized via some off-the-shelf approaches, e.g., quasi-Newton and conjugate gradient, provided that k is only moderately large. Specially, by choosing $V = \{\pm e^i, i = 1, ..., d\}$ where $\{e^i\}_{i=1}^d$ are the canonical basis vectors in $\mathbb{R}^d$, FBS reduces to FCFGS.

Algorithm 2: Forward Basis Selection (FBS).

Initialization: $x^{(0)} = 0$, $U^{(0)} = \emptyset$.
Output: $x^{(K)}$.
for $k = 1, ..., K$ **do**
  Calculate $u^{(k)}$ by solving

$$u^{(k)} = \arg\min_{u \in V} \langle \nabla f(x^{(k-1)}), u \rangle. \tag{8}$$

  Set $U^{(k)} = U^{(k-1)} \cup \{u^{(k)}\}$ and update

$$x^{(k)} = \arg\min_{x \in \mathcal{L}(U^{(k)})} f(x). \tag{9}$$

end

Since FBS is a generalization of FCFGS, one natural question is whether appealing aspects of FCFGS such as orthogonal coordinate selection and fast convergence can be similarly established for FBS. We will answer this question in the following analysis. The Lemma below shows that before reaching the optimality, Algorithm 2 always introduces at each iteration a new basis atom as the descent direction.

Lemma 1. *Assume that V is symmetric and we run Algorithm 2 until time instance k.*

(a) If $\langle \nabla f(x^{(k-1)}), u^{(k)} \rangle \neq 0$, then the elements in $U^{(k)} = \{u^{(1)}, \ldots, u^{(k)}\}$ are linearly independent.

(b) If $\langle \nabla f(x^{(k-1)}), u^{(k)} \rangle = 0$, then $x^{(k-1)}$ is the optimal solution over the linear hull $\mathcal{L}(V)$.

The proof is given in Appendix A.1. This lemma indicates that if we run Algorithm 2 until time instance k with $\langle \nabla f(x^{(k-1)}), u^{(k)} \rangle \neq 0$, then the atom set $U^{(k)}$ forms k bases in V. This corresponds to the orthogonal coordinate selection property of FCFBS, which justifies why we call Algorithm 2 as forward basis selection.

On convergence performance of FBS, we are interested in the approximation accuracy of the output $x^{(K)}$ towards a fixed S-sparse competitor $\bar{x} \in \mathcal{L}_S(V)$. We first discuss the special case where V is finite and then address the general case where V is bounded and symmetric.

2.1 A Special Case: V is Finite

Let us consider the special case that dictionary $V = \{v_1, ..., v_N\}$ is finite with cardinality N. Without loss of generality, we assume that the elements in V are linearly independent (otherwise we can replace V with its bases without affecting the feasible set). Thus, for any $x \in \mathcal{L}(V)$ the representation $x = \sum_{i=1}^N \alpha_i v_i$ is unique. Let $g(\alpha) := f\left(\sum_{i=1}^N \alpha_i v_i\right)$. Since $f(x)$ is convex, it is easy to verify that $g(\alpha)$ is convex in $\mathbb{R}^N$. We may convert problem (1) to the following standard coordinatewise sparse learning problem

$$\min_{\alpha \in \mathbb{R}^N} g(\alpha), \quad \text{s.t. } \|\alpha\|_0 \leq K. \tag{10}$$

In light of this conversion, we can straightforwardly apply the FCFGS (Algorithm 1) to solve problem (10). By making restricted strong convexity assumptions on $g(\alpha)$, it is known from (Shalev-Shwartz et al., 2010, Theorem 2.8) that the rate of convergence of FCFGS towards any sparse competitive solution is geometric.

2.2 General Cases

Given $x \in \mathcal{L}_K(V)$, it is known from the definition (2) that there exists a set $U \subseteq V$ with cardinality K such that $x = \sum_{u \in U} \alpha_u(x) u$. Typically, such a representation is not unique. In the following discussion, we are interested in the representation of x on $\mathcal{L}_K(V)$ with the smallest sum of absolute weights $\sum_{u \in U} |\alpha_u(x)|$.

Definition 1 (Minimal Representation Length)**.** *For any $x \in \mathcal{L}_K(V)$, the minimal representation length of x is defined as*

$$C_K(x) := \min_{U \subseteq V, |U| \leq K} \left\{ \sum_{v \in U} |\alpha_u(x)| : x = \sum_{u \in U} \alpha_u(x) u, \right\}.$$

The following theorem is our main result on approximation performance of FBS over a bounded and symmetric dictionary V.

Theorem 1. *Let us run FBS (Algorithm 2) with K iterations. Assume that V is symmetric and bounded with radius A. Assume that f is $\rho_+(1)$-restricted-strongly smooth over V. Given $\epsilon > 0$ and $\bar{x} \in \mathcal{L}_S(V)$, if $\forall k \leq K$, $f(x^{(k)}) > f(\bar{x})$ and*

$$K \geq \frac{2\rho_+(1)A^2 C_S(\bar{x})^2}{\epsilon} - 1, \tag{11}$$

then FBS will output $x^{(K)}$ satisfying $f(x^{(K)}) \leq f(\bar{x}) + \epsilon$.

The proof is given in Appendix A.2. Notice that the bound in the right hand side of (11) is proportional to the minimal representation length $C_K(\bar{x})$ which reflects the sparsity of $\bar{x}$ over the dictionary V.

3 Extensions

In this section, we extend FBS to the setup of non-negative and convex sparse approximation over a given dictionary. These extensions enhance the applicability of FBS to a wider range of sparse learning problems.

3.1 Non-Negative Sparse Approximation

In certain sparse learning problems, e.g., non-negative sparse regression and positive semi-definite matrix learning, the target solution is expected to stay in a non-negative hull of a dictionary V given by

$$\mathcal{L}^+(V) := \left\{ \sum_{v \in V} \alpha_v v : \alpha_v \in \mathbb{R}^+ \right\}.$$

Let us consider the following problem of non-negative sparse approximation over V:

$$\min_{x \in \mathcal{E}} f(x), \quad \text{s.t. } x \in \mathcal{L}_K^+(V), \tag{12}$$

where

$$\mathcal{L}_K^+(V) := \bigcup_{U \subseteq V, |U| \leq K} \left\{ \sum_{u \in U} \alpha_u u : \alpha_u \in \mathbb{R}^+ \right\}.$$

To apply FBS to this problem, we have to modify the update (9) to adapt the non-negative constraint:

$$x^{(k)} = \operatorname*{arg\,min}_{x \in \mathcal{L}^+(U^{(k)})} f(x). \tag{13}$$

The proceeding subproblem is essentially a smooth optimization over half-space with scale dominated by the time instance k. It can be efficiently solved via quasi-Newton methods such as PQN (Schmidt et al., 2009).

Following the similar argument as in the Section 2.2, it can be proved that Theorem 1 is still valid for this extension when V is bounded and symmetric.

Specially, when V is finite with cardinality N, as discussed in Section 2.1 that we may convert problem (12) to the following non-negative coordinatewise sparse learning problem

$$\min_{\alpha \in \mathbb{R}^N} g(\alpha), \quad \text{s.t. } \|\alpha\|_0 \leq K, \alpha \geq 0. \tag{14}$$

To apply the FCFGS (Algorithm 1) to solve problem (14), we have to make the following slight modifications of (6) and (7) to adapt the non-negative constraint:

$$j^{(k)} = \operatorname*{arg\,min}_{i \in \{1,\ldots,N\}} [\nabla g(\alpha^{(k-1)})]_i, \tag{15}$$

$$\alpha^{(k)} = \operatorname*{arg\,min}_{\mathrm{supp}(\alpha) \subseteq F^{(k)}, \alpha \geq 0} g(\alpha). \tag{16}$$

By making restricted strong convexity assumptions on $g(\alpha)$, with the similar arguments as in (Shalev-Shwartz et al., 2010, Theorem 2.8), it can be proved that the geometric rate of convergence of FCFGS still holds with the preceding modifications (15) and (16).

3.2 Convex Sparse Approximation

In many sparse learning problems, the feasible set is a convex hull $\mathcal{L}^\triangle(V)$ of a dictionary V given by

$$\mathcal{L}^\triangle(V) := \left\{ \sum_{v \in V} \alpha_v v : \alpha_v \in \mathbb{R}^+, \sum_v \alpha_v = 1 \right\}.$$

For example, in Lasso (Tibshirani, 1996), the solution is restricted in the ℓ_1-norm ball which is a convex hull of the canonical basis and their negative counterparts. Generally, for any convex dictionary V we have $V = \mathcal{L}^\triangle(V)$. Therefore the feasible set of any convex optimization problem is the convex hull of itself.

Let us consider the following problem of convex sparse approximation over V:

$$\min_{x \in \mathcal{E}} f(x), \quad \text{s.t. } x \in \mathcal{L}_K^\triangle(V), \tag{17}$$

where

$$\mathcal{L}_K^\triangle(V) := \bigcup_{U \subseteq V, |U| \leq K} \left\{ \sum_{u \in U} \alpha_u u : \alpha_u \in \mathbb{R}^+, \sum_u \alpha_u = 1 \right\}.$$

To apply FBS to solve problem (17), we modify the update (9) to adapt the convex constraint:

$$x^{(k)} = \operatorname*{arg\,min}_{x \in \mathcal{L}^\triangle(U^{(k)})} f(x). \tag{18}$$

The preceding subproblem is essentially a smooth optimization over simplex with scale k. Again, it can be efficiently solved via off-the-shelf methods such as PQN.

We next establish convergence rates of FBS (with modification (18)) for convex sparse approximation.

3.2.1 V is Finite

When V is finite with cardinality N, based on the discussion in Section 2.1 we may convert problem (17) to the following convex sparse learning problem

$$\min_{\alpha\in\mathbb{R}^N} g(\alpha), \quad \text{s.t. } \|\alpha\|_0 \leq K, \alpha \in \triangle_N. \tag{19}$$

where $\triangle_N := \{\alpha \in \mathbb{R}^N : \alpha \in \mathbb{R}^+, \|\alpha\|_1 = 1\}$ is the N-dimensional simplex. To apply the FCFGS (Algorithm 1) to solve problem (19), we have to make the following modifications of (6) and (7) to adapt the convexity constraint:

$$j^{(k)} = \arg\min_{i\in\{1,\ldots,N\}} [\nabla g(\alpha^{(k-1)})]_i, \tag{20}$$

$$\alpha^{(k)} = \arg\min_{\text{supp}(\alpha)\subseteq F^{(k)}, \alpha\in\triangle_k} g(\alpha). \tag{21}$$

The following result shows that by making restricted strong convexity/smoothness assumptions on g, the geometric rate of convergence of FCFGS (Shalev-Shwartz et al., 2010, Theorem 2.8) is still valid with the preceding modifications. This result is a non-trivial extension of the result (Shalev-Shwartz et al., 2010, Theorem 2.8) to the setting of convex sparse approximation. In the rest of this subsection, the restricted strong smoothness and restricted strong convexity are both defined over canonical bases.

Theorem 2. *Let $g(\alpha)$ be a differentiable convex function with domain $\mathbb{R}^N$ and $\bar{\alpha}$ a S-sparse vector in a simplex. Let us run K iterations of FCFGS (Algorithm 1) to solve problem* (19) *with update* (20) *and* (21). *Assume that g is $\rho_+(K+1)$-strongly smooth and is $\rho_-(K+S)$-strongly convex. Assume that g is L-Lipschitz continuous, i.e., $|g(\alpha) - g(\alpha')| \leq L\|\alpha - \alpha'\|$. Given $\epsilon > 0$, if $\forall k \leq K$, $g(\alpha^{(k)}) > g(\bar{\alpha})$ and*

$$K \geq \frac{1}{s(K,S)} \ln \frac{g(\alpha^{(0)}) - g(\bar{\alpha})}{\epsilon}, \tag{22}$$

where

$$s(K,S) := \min\left\{\frac{\rho_+(K+1)}{L}, \frac{\rho_-(K+S)}{4S\rho_+(K+1)}\right\},$$

then FCFGS (Algorithm 1) will output $\alpha^{(K)}$ satisfying $g(\alpha^{(K)}) \leq g(\bar{\alpha}) + \epsilon$.

The proof is given in Appendix A.3. To the best of our knowledge, Theorem 2 for the first time establishes a geometric rate of convergence for fully-corrective-type convex sparse approximation approaches.

3.2.2 V is Bounded

We now turn to the general case where V is a bounded set. The following theorem is our main result.

Theorem 3. *Let us run K iterations of FBS (Algorithm 2) with $x^{(k)}$ updated by* (18). *Assume that V is bounded with radius A. Assume that f is $\rho_+(K+1)$-strongly smooth over V. Given $\epsilon > 0$ and $\bar{x} \in \mathcal{L}^{\triangle}(V)$, if $\forall k \leq K$, $f(x^{(k)}) > f(\bar{x})$ and*

$$K \geq \log_2\left[\frac{f(x^{(0)}) - f(\bar{x})}{4\rho_+(K+1)A^2}\right] + \frac{8\rho_+(K+1)A^2}{\epsilon} - 1,$$

then FBS will output $x^{(K)}$ satisfying $f(x^{(K)}) \leq f(\bar{x}) + \epsilon$.

The proof is given in A.4. Note that in this result, we do not require V to be symmetric.

Remark 1. *When dictionary V is convex, the FBS with modification* (18) *can be regarded as a generic first-order method to minimize f over V. The first-order optimization approaches have been extensively studied and applied in machine learning. On one hand, compared to the optimal first-order methods (Tseng, 2008; Nesterov, 2004) which converge with rate $\mathcal{O}(1/\sqrt{\epsilon})$ and the quasi-Newton methods (Schmidt et al., 2009) with near super-linear convergence rate, a moderately increased number of $\mathcal{O}(1/\epsilon)$ steps are needed in total by the FBS for arbitrary convex objectives. On the other hand, as demonstrated shortly in Section 4 that for some relatively complex constraints, e.g., ℓ_1-norm and nuclear-norm constraints, the linear projection operator* (8) *used in FBS is significantly cheaper than Euclidean projection operator used in most projected gradient methods. Therefore, the $\mathcal{O}(1/\epsilon)$ rate in Theorem 3 represents the price for the severe simplification in each individual step, as well as the inherent sparsity over V.*

4 Applications

In this section, we apply FBS and its extensions to several statistical learning problems which can be formulated as (1) with particular choices of dictionary V. Here we focus on three applications: low-rank matrix learning, positive semi-definite matrix learning and ℓ_1-ball constrained sparse learning.

4.1 Low-Rank Matrix Learning

Let us consider the following low-rank constrained matrix learning problem which is widely applied in matrix

completion and approximation:

$$\min_{X\in\mathbb{R}^{m\times n}} f(X), \quad \text{s.t. } \text{rank}(X)\leq K. \tag{23}$$

The motivation of applying FBS to this problem is the observation that the feasible set $\{X\in\mathbb{R}^{m\times n}: \text{rank}(X)\leq K\}=\mathcal{L}_K(V_{lr})$ where

$$V_{lr}:=\{uv^T, u\in\mathbb{R}^m, v\in\mathbb{R}^n, \|u\|=\|v\|=1\}.$$

This is because, by the SVD theory any $X\in\mathbb{R}^{m\times n}$ of rank no more than K can be written as $X=\sum_{i=1}^K \sigma_i u_i v_i^T$. Based on this equivalence, we may solve the following problem

$$\min_{X\in\mathbb{R}^{m\times n}} f(X), \quad \text{s.t. } X\in\mathcal{L}_K(V_{lr}).$$

We now specify FBS for sparse approximation in this special case. The linear projection (8) at time instance k becomes

$$Y^{(k)}=\arg\max_{Y\in V_{lr}}\langle -\nabla f(X^{(k-1)}), Y\rangle. \tag{24}$$

The following result establishes a closed-form solution for the preceding linear projection.

Proposition 1. *For any $X\in\mathbb{R}^{m\times n}$, one solution of $\bar{Y}=\arg\max_{Y\in V_{lr}}\langle X,Y\rangle$ is given by $\bar{Y}=uv^T$ where u and v are the left and right singular vectors corresponding to the largest singular value of X.*

The proof is given in Appendix A.5. By invoking the preceding proposition to (24) we immediately get that $Y^{(k)}=uv^T$ where u and v are the leading left and right singular vectors of $-\nabla f(X^{(k-1)})$. On the problem of leading singular vector computation, some efficient procedures using the Lanczos algorithm can be found in (Hazan, 2008; Arora et al., 2005).

4.2 Positive Semidefinite & Low Rank Matrix Learning

In this subsection, we consider the following problem of convex optimization over the cone of Positive Semi-Definite (PSD) matrices with low rank constraint.

$$\min_{X\in\mathbb{R}^{n\times n}} f(X), \quad \text{s.t. } X\succeq 0, \text{rank}(X)\leq K. \tag{25}$$

To solve this problem, we consider applying FBS to perform sparse approximation over $\mathcal{L}_K^+(V_{sdp})$ where V_{psd} is given by

$$V_{psd}:=\{uu^T, u\in\mathbb{R}^n, \|u\|=1\}.$$

This is motivated from the SVD theory that any PSD matrix $X\in\mathbb{R}^{n\times n}$ with rank at most K can be written as $X=\sum_{i=1}^K \sigma_i u_i u_i^T$ with $\sigma_i\geq 0$. In this case, at time instance k, the subproblem (8) in FBS becomes

$$Y^{(k)}=\arg\max_{Y\in V_{psd}}\langle -\nabla f(X^{(k-1)}), Y\rangle. \tag{26}$$

The following result shows that we can find a closed-form solution for the preceding linear projection.

Proposition 2. *For any matrix $X\in\mathbb{R}^{n\times n}$, one solution of $\bar{Y}=\arg\max_{Y\in V_{psd}}\langle X,Y\rangle$ is given by $\bar{Y}=uu^T$ where u is the leading eigenvector of X.*

The proof is given in Appendix A.6. By invoking the preceding proposition to (26) we immediately get that $Y^{(k)}=uu^T$ where u is the leading eigenvector of $-\nabla f(X^{(k-1)})$. The Lanczos algorithm can be utilized for leading eigenvector calculation.

We conclude this example by pointing out that FBS is also directly applicable to solve Semi-Definite Program (SDP), i.e., problem (25) without the rank constraint. Indeed, SDP is a special case of problem (25) when $K\to\infty$.

4.3 Sparse Learning over ℓ_1-norm Ball

Consider the problem of convex minimization over ℓ_1-norm ball which is widely applied in signal processing and machine learning:

$$\min_{x\in\mathbb{R}^d} f(x), \quad \text{s.t. } \|x\|_1\leq\tau. \tag{27}$$

The inspiration of using FBS to solve this problem is from the observation that the ℓ_1-norm ball $\|x\|_1\leq\tau$ is the a convex hull of the set $V_{\ell_1,\tau}=\{\pm\tau e^i, i=1,...,d\}$. In order to do convex sparse approximation, we may apply the variant of FBS as stated in Section 3.2 to solve the following problem:

$$\min_{x\in\mathbb{R}^d} f(x), \quad \text{s.t. } x\in\mathcal{L}_K^{\triangle}(V_{\ell_1,\tau}). \tag{28}$$

We now specify FBS for this case. The gradient linear projection (20) is given by

$$u^{(k)}=-\tau\text{sign}([\nabla f(x^{(k)})]_j)e^j,$$

where $j=\arg\max_i|[\nabla f(x^{(k)})]_i|$. Such a linear projection only involves simple max-operation of a vector and thus is more efficient especially in high dimensional data set than Euclidean ℓ_1-norm ball projection (Duchi et al., 2008) which requires relatively more sophisticated vector operations.

5 Related Work

Recently, forward greedy selection algorithms have received wide interests in machine learning. A category of algorithms called *coreset* (Clarkson, 2008)

have been successfully applied in functional approximation (Zhang, 2003) and coordinatewise sparse learning (Kim & Kim, 2004). This body of work dates back to the Frank-Wolfe algorithm (Frank & Wolfe, 1956) for polytope constrained optimization. Some variants of coreset method are proposed in the scenarios of SDP (Hazan, 2008) and low-rank matrix completion/approximation (Jaggi & Sulovský, 2010; Shalev-Shwartz et al., 2011) which only requires partial SVD for leading singular value at individual iteration step. In the context of boosting classification, the restricted gradient projection algorithms stated in (Grubb & Bagnell, 2011) is essentially a forward greedy selection method over $\mathcal{L}^2$-functional space. Recently, Tewari et al. (2011) proposed a Frank-Wolf-type method to minimize convex objective over the (scaled) convex hull of a a collection of atoms. Different from their method, FBS always introduces a new basis (atom) into the active set and thus leads to sharper convergence rate under proper assumptions.

The modified FBS with update (18) can be taken as a generic first-order method for convex optimization. In many existing projected gradient algorithms, e.g. proximal gradient methods (Tseng, 2008) and quasi-Newton methods (Schmidt et al., 2009), Euclidean projection is utilized at each iteration to guarantee the feasibility of solution. Differently, our method utilizes the linear projection operator (8) which is cheaper than Euclidean projection in problems such as SDP. Recently, a forward-selection-type of algorithm has been studied in (Jaggi, 2011) for convex optimization, which can be regarded as a generalized steepest descent method. Our method differs from this method in the fully corrective adjustment at each iteration which improves the convergence.

6 Experiments

In this section, we demonstrate the numerical performances of FBS in two applications: low rank representation for subspace segmentation and sparse SVMs for document classification. Our algorithms are implemented in Matlab (Version 7.7, Vista). All runs are performed on a commodity desktop with Intel Core2/Quad 2.80GHz and 8G RAM.

6.1 FBS for Low Rank Representation

We test in this experiment the performance of FBS when applied to low rank and PSD constrained matrix learning. Specially, we focus on the following problem of low rank representation (Liu et al., 2010; Ni et al., 2010) for subspace segmentation:

$$\min_{D\in\mathbb{R}^{n\times n}} \mathrm{rank}(D), \quad \text{s.t. } X = XD, D \succeq 0, \qquad (29)$$

where $X = [x_1, x_2, ..., x_n] \in \mathbb{R}^{d\times n}$ are n observed data vectors drawn from p unknown linear subspaces $\{\mathcal{S}_i\}_{i=1}^p$. The analysis in (Liu et al., 2010) shows that the optimal representation D^* of problem (29) captures the global structure of data and thus naturally forms an affinity matrix for spectral clustering. Furthermore, it is justified in (Ni et al., 2010) that the PSD constraint is effective to enforce the representation D to be a valid kernel.

To apply FBS to solve the low rank representation problem, we alternatively solve a penalized version which fits the model (25):

$$\min_{D\in\mathbb{R}^{n\times n}} \|X - XD\|_F^2, \ \text{s.t. } \mathrm{rank}(D) \le K, D \succeq 0, \ (30)$$

where $\|\cdot\|_F$ is the Frobenius norm. We can apply the non-negative variant of FBS as stated in Section 3.1 & 4.2 to approximately solve problem (30).

We conduct the experiment on the Extended Yale Face Database B (EYD-B)[1]. The EYD-B contains 16,128 images of 38 human subjects under 9 poses and 64 illumination conditions. Following the experimental setup in (Liu et al., 2010), we use the first 10 individuals with 64 near frontal face images for each individual in our experiment. The size of each cropped gray scale image is 42×48 pixels and we use the raw pixel vales to form data vectors of dimension 2016. Each image vector is then normalized to unit length.

We compare FBS with the LRR (Liu et al., 2010)[2] which solves problem (29) via Augmented Lagrange Multiplier (ALM). For clustering, the respectively learnt representations D by FBS and LRR are fed into the same spectral clustering routine. In this experiment, we initialize $D^{(0)} = 0$ and set $K = 70$ in FBS. Table 1 lists the results on EYD-B. It can be observed that FBS and LRR achieve the comparative clustering accuracies while the former needs much less CPU time. Meanwhile, it can be seen from the row "Rank" that FBS outputs a representation matrix with lower rank than that of LRR. This experiment validates that FBS is an efficient and effective sparse approximation method for low rank representation problem.

Table 1: Results on the EYD-B dataset.

Algorithms	FBS	LRR
Rank	70	135
CPU time	31.9	114.8
Accuracy (%)	64.8	63.9

[1] http://vision.ucsd.edu/~leekc/ExtYaleDatabase/ExtYaleB.html

[2] Matlab code is available at http://sites.google.com/site/guangcanliu/

6.2 FBS for Sparse L_2-SVMs

Denote $\mathcal{D} = \{(x_i, y_i)\}_{1 \le i \le n}$ a set of observed data, $x_i \in \mathbb{R}^d$ is the feature vector, and $y_i \in \{+1, -1\}$ is the binary class label. Let us consider the following problem of L_2-SVMs constrained by ℓ_1-norm ball:

$$\min_{w \in \mathbb{R}^d} R(w) + \frac{\lambda}{2}\|w\|^2, \quad \text{s.t. } \|w\|_1 \le \tau, \tag{31}$$

where $R(w) := \frac{1}{2n}\sum_{i=1}^{n}(\max\{0, 1 - y_i\langle w, x_i\rangle\})^2$ is the empirical risk suffered from w. We apply the convex sparse approximation variant of FBS as discussed in Section 3.2 & 4.3 to solve problem (31). For this experiment, we use the rcv1.binary dataset ($d = 47,236$) which is a standard benchmark for binary classification on sparse data. A training subset of size $n = 20,242$ and a testing subset of size $20,000$ are used. In this experiment, we initialize $w^{(0)} = 0$ and set $\lambda = 10^{-5}$.

We compare FBS with two representative projected gradient methods, the APG (Tseng, 2008) and the PQN (Schmidt et al., 2009), both call the Euclidean projection to project the current iterate onto the feasible set. From Figure 1(a) we can observe that PQN converges the fastest, while FBS converges sharper than APG. Figure 1(b) plots the objective evolving curves of FBS under different radius τ, which show that FBS works well under a large range of τ. Table 2 lists the quantitative results by different algorithms. It can be observed from the row "Sparsity" that FBS outputs the sparsest solution at the cost of a slightly increased testing error. This can be interpreted by the sparse approximation nature of FBS. From the row "CPU Projection" we can see that the linear projection used in FBS is more efficient than the Euclidean projection used in APG and PQN. On overall computational efficiency, PQN performs the best.

Table 2: Results on the rcv1.binary dataset.

Algorithms	FBS	APG	PQN
Objective	0.22	0.22	0.22
Iteration	51	200	9
Sparsity	51	600	112
CPU Projection (sec.)	0.03	0.74	0.15
CPU over all (sec.)	4.52	6.56	0.74
Testing Error (%)	11.64	10.87	10.40

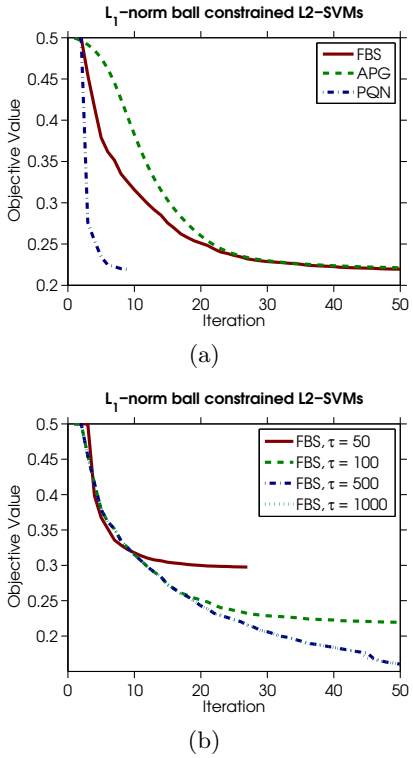

Figure 1: Objective value evolving curves on the rcv1.binary data set. For better viewing, please see the original pdf file.

7 Conclusion

The proposed FBS algorithm generalizes the FCFGS from coordinatewise sparse approximation to a relaxed setting of sparse approximation over a fixed dictionary. At each iteration, FBS automatically selects a new basis atom in the dictionary achieving the minimum inner product with the current gradient, and then optimally adjusting the combination weights of the bases selected so far. We then extend FBS to the setup of non-negative and convex sparse approximation. Convergence analysis shows that FBS and its extensions generally converge sublinearly, while geometric rate of convergence can be derived under stronger conditions. The per-iteration computational overhead of FBS is dominated by a linear projection which is more efficient than Euclidean projection in problems such as coordinatewise sparsity and low-rank constrained learning. The subproblem of combination weights optimization can be efficiently solved via off-the-shelf methods. The proposed methods are applicable to several sparse learning problems with efficiency validated by experiments on benchmarks. To conclude, FBS is a generic yet efficient method for sparse approximation over a fixed dictionary.

Acknowledgment

This work was mainly performed when Dr. Xiao-Tong Yuan was a postdoctoral fellow in National University of Singapore. We would like to acknowledge to support of NExT Research Center funded by MDA, Singapore, under the research grant: WBS:R-252-300-001-490.

Appendix

A Technical Proofs

The goal of this appendix section is to prove several results stated in the main body of this paper.

A.1 The Proof of Lemma 1

Proof. Part (a): We prove the claim with induction. Obviously, the claim holds for $k = 1$ (since $u^{(1)} \neq 0$). Given that the claim holds until time instance $k-1$. Assume that at time instance k, $\{u^{(1)}, \ldots, u^{(k)}\}$ are linearly dependent. Since $\{u^{(1)}, \ldots, u^{(k-1)}\}$ are linearly independent, we have that $u^{(k)}$ can be expressed as a linear combination of $\{u^{(1)}, \ldots, u^{(k-1)}\}$. Due to the optimality of $x^{(k-1)}$ for solving (9) at time instance $k-1$, we have $\langle \nabla f(x^{(k-1)}), u^{(i)} \rangle = 0$, $i \leq k-1$. Therefore, $\langle \nabla f(x^{(k-1)}), u^{(k)} \rangle = 0$, which leads to contradiction. Thus, the claim holds for k. This proves the desired result.

Part (b): Given that $\langle \nabla f(x^{(k-1)}), u^{(k)} \rangle = 0$, we have $\forall v \in V$, $\langle \nabla f(x^{(k-1)}), v \rangle \geq 0$, which implies $\langle \nabla f(x^{(k-1)}), v \rangle = 0$ since V is symmetric. Therefore $x^{(k-1)}$ is optimal over $\mathcal{L}(V)$. □

A.2 The Proof of Theorem 1

Proof. From the update of $x^{(k)}$ in (9) and the definition of restricted strong smoothness in (3) we get that $\forall \eta \geq 0$,

$$\begin{aligned}
& f(x^{(k)}) \\
\leq\ & f\left(x^{(k-1)} + \eta u^{(k)}\right) \\
\leq\ & f(x^{(k-1)}) + \eta \langle \nabla f(x^{(k-1)}), u^{(k)} \rangle + \frac{\rho_+(1) A^2 \eta^2}{2} \\
\leq\ & f(x^{(k-1)}) + \frac{\eta}{C_S(\bar{x})} \langle \nabla f(x^{(k-1)}), \bar{x} \rangle + \frac{\rho_+(1) A^2 \eta^2}{2} \\
=\ & f(x^{(k-1)}) + \frac{\eta}{C_S(\bar{x})} \langle \nabla f(x^{(k-1)}), \bar{x} - x^{(k-1)} \rangle \\
& + \frac{\rho_+(1) A^2 \eta^2}{2} \\
\leq\ & f(x^{(k-1)}) + \frac{\eta}{C_S(\bar{x})} (f(\bar{x}) - f(x^{(k-1)})) + \frac{\rho_+(1) A^2 \eta^2}{2},
\end{aligned}$$

where the second inequality follows the restricted strong smoothness and the boundness assumption of V, the third inequality follows (8) and the assumption that V is symmetric, the first equality follows the optimality condition $\langle \nabla f(x^{(k-1)}), x^{(k-1)} \rangle = 0$ of the iterate $x^{(k-1)}$, and the last inequality follows from the convexity of f. Particularly, the preceding inequality holds for

$$\eta = \frac{f(x^{(k-1)}) - f(\bar{x})}{\rho_+(1) A^2 C_S(\bar{x})} > 0,$$

and consequently

$$f(x^{(k)}) \leq f(x^{k-1}) - \frac{(f(x^{(k-1)}) - f(\bar{x}))^2}{2\rho_+(1) A^2 C_S(\bar{x})^2}.$$

Denote $\epsilon_k := f(x^{(k)}) - f(\bar{x})$. The preceding inequality implies

$$\epsilon_k \leq \epsilon_{k-1} - \frac{\epsilon_{k-1}^2}{2\rho_+(1) A^2 C_S(\bar{x})^2}.$$

Invoking the Lemma B.2 in (Shalev-Shwartz et al., 2010) shows that

$$\epsilon_k \leq \frac{2\rho_+(1) A^2 C_S(\bar{x})^2}{k+1}.$$

If K satisfies (11), then it is guaranteed $\epsilon_K \leq \epsilon$. □

A.3 Proof of Theorem 2

We first prove the following lemma which is key to our analysis.

Lemma 2. *Given $\tilde{\alpha} \in \triangle_N$ with $supp(\tilde{\alpha}) = \tilde{F}$, let F be an index set such that $\tilde{F} \setminus F \neq \emptyset$. Let*

$$\alpha = \mathop{\arg\min}_{supp(\alpha) \subseteq F, \alpha \in \triangle_N} g(\alpha).$$

Assume that g is L-Lipschitz continuous, $\rho_+(|F|+1)$-restricted-strongly smooth and $\rho_-(|F \cup \tilde{F}|)$-restricted-strongly convex. Assume that $g(\alpha) > g(\tilde{\alpha})$. Let $j = \arg\min_i [\nabla g(\alpha)]_i$. Then there exists $\eta \in [0,1]$ such that

$$g(\alpha) - g((1-\eta)\alpha + \eta e^j) \geq s(g(\alpha) - g(\tilde{\alpha})),$$

where constant s is given by

$$s := \min\left\{ \frac{\rho_+(|F|+1)}{L}, \frac{\rho_-(|F \cup \tilde{F}|)}{4\rho_+(|F|+1)|\tilde{F}|} \right\}. \quad \text{(A.1)}$$

Proof. Due to the strong smoothness of $g(\alpha)$ and $\alpha \in \triangle_N$, we have that for $\eta \in [0,1]$ the following inequality holds

$$\begin{aligned}
& g((1-\eta)\alpha + \eta e^j) \\
\leq\ & h_j(\eta) := g(\alpha) + \eta \langle \nabla g(\alpha), e^j - \alpha \rangle + 2\eta^2 \rho_+(|F|+1).
\end{aligned}$$

The definition of j implies $h_j(\eta) \leq h_i(\eta)$, $i = 1, ..., N$. The lemma is a direct consequence of the following stronger statement

$$g(\alpha) - h_j(\eta) \geq s(g(\alpha) - g(\tilde{\alpha})), \quad \text{(A.2)}$$

for an appropriate choice of $\eta \in [0,1]$ and s given by (A.1). We now turn to show the validity of the inequality (A.2).

Denote $F^c = \tilde{F} \setminus F$ and $\tau = \sum_{i\in F^c} \tilde{\alpha}_i$. It holds that (recall $\tilde{\alpha}_i \geq 0$)

$$\begin{aligned}\tau h_j(\eta) &\leq \sum_{i\in F^c} \tilde{\alpha}_i h_i(\eta) \\ &= \tau g(\alpha) + \eta\left(\sum_{i\in F^c} \tilde{\alpha}_i [\nabla g(\alpha)]_i - \tau\langle \nabla g(\alpha), \alpha\rangle\right) \\ &\quad +2\eta^2\tau\rho_+(|F|+1). \qquad \text{(A.3)}\end{aligned}$$

From the optimality of α we get that

$$\langle \nabla g(\alpha), \sum_{i\in F} \tilde{\alpha}_i e^i/(1-\tau) - \alpha\rangle \geq 0. \qquad \text{(A.4)}$$

Indeed, $\sum_{i\in F} \tilde{\alpha}_i e^i/(1-\tau) \in \triangle_N$ and is supported on F. Additionally, $\alpha_i = 0$ for $i \notin F$ and $\tilde{\alpha}_i = 0$ for $i \notin \tilde{F}$. Therefore

$$\begin{aligned}&\sum_{i\in F^c} \tilde{\alpha}_i [\nabla g(\alpha)]_i \\ &= \sum_{i\in F^c} (\tilde{\alpha}_i [\nabla g(\alpha)]_i - (1-\tau)\alpha_i) \\ &\leq \sum_{i\in F\cup\tilde{F}} (\tilde{\alpha}_i [\nabla g(\alpha)]_i - (1-\tau)\alpha_i) \\ &= \langle \nabla g(\alpha), \tilde{\alpha} - (1-\tau)\alpha\rangle \\ &= \langle \nabla g(\alpha), \tilde{\alpha} - \alpha\rangle + \tau\langle \nabla g(\alpha), \alpha\rangle,\end{aligned}$$

where the inequality follows (A.4). Combining the preceding inequality with (4) we obtain that

$$\begin{aligned}&\sum_{i\in F^c} \tilde{\alpha}_i [\nabla g(\alpha)]_i - \tau\langle \nabla g(\alpha), \alpha\rangle \\ &= \langle \nabla g(\alpha), \tilde{\alpha} - \alpha\rangle \\ &\leq g(\tilde{\alpha}) - g(\alpha) - \frac{\rho_-(|F\cup\tilde{F}|)}{2}\|\alpha - \tilde{\alpha}\|^2.\end{aligned}$$

Combining the above with (A.3)

$$\begin{aligned}&\tau h_j(\eta) \\ &\leq \tau g(\alpha) - \eta\left(g(\alpha) - g(\tilde{\alpha}) + \frac{\rho_-(|F\cup\tilde{F}|)}{2}\|\alpha - \tilde{\alpha}\|^2\right) \\ &\quad +2\eta^2\tau\rho_+(|F|+1).\end{aligned}$$

Invoking Lemma 3 on the right hand side of the preceding inequality we get that $\exists\hat{\eta} \in [0,1]$ such that

$$g(\alpha) - h_j(\hat{\eta}) \geq \frac{\delta}{2\tau}\min\left\{1, \frac{\delta}{4\tau\rho_+(|F|+1)}\right\},$$

where $\delta := g(\alpha) - g(\tilde{\alpha}) + \frac{\rho_-(|F\cup\tilde{F}|)}{2}\|\alpha - \tilde{\alpha}\|^2$. We next distinguish the following two cases:

(a) If $\delta \geq 4\tau\rho_+(|F|+1)$. In this case,

$$\begin{aligned}g(x) - h_j(\hat{\eta}) &\geq \frac{\delta}{2\tau} \\ &\geq 2\rho_+(|F|+1) \\ &\geq \frac{\rho_+(|F|+1)(g(\alpha) - g(\tilde{\alpha}))}{L},\end{aligned}$$

where the last inequality follows from the Lipschitz continuity $g(\alpha) - g(\tilde{\alpha}) \leq L\|\alpha - \tilde{\alpha}\| \leq 2L$.

(b) If $\delta < 4\tau\rho_+(|F|+1)$. In this case,

$$\begin{aligned}&g(\alpha) - h_j(\hat{\eta}) \\ &\geq \frac{\delta^2}{8\tau^2\rho_+(|F|+1)} \\ &\geq \frac{2\rho_-(|F\cup\tilde{F}|)(g(\alpha) - g(\tilde{\alpha}))\|\alpha - \tilde{\alpha}\|^2}{8\tau^2\rho_+(|F|+1)}, \\ &\geq \frac{\rho_-(|F\cup\tilde{F}|)(g(\alpha) - g(\tilde{\alpha}))\sum_{i\in F^c}\tilde{\alpha}_i^2}{4\tau^2\rho_+(|F|+1)} \\ &\geq \frac{\rho_-(|F\cup\tilde{F}|)(g(\alpha) - g(\tilde{\alpha}))}{4\rho_+(|F|+1)\|\tilde{\alpha}\|_0} \\ &\geq \frac{\rho_-(|F\cup\tilde{F}|)(g(\alpha) - g(\tilde{\alpha}))}{4\rho_+(|F|+1)|\tilde{F}|}.\end{aligned}$$

Combining both cases we prove the claim (A.2). □

Proof of Theorem 2. Denote $\epsilon_k := g(\alpha^{(k)}) - g(\bar{\alpha})$. The definition of update (21) implies that $g(\alpha^{(k)}) \leq \min_{\eta\in[0,1]} g((1-\eta)\alpha^{(k-1)} + \eta e^{j^{(k)}})$. The conditions of Lemma 2 are satisfied and therefore we obtain that (with $F = F^{(k)}$ and $\tilde{F} = \mathrm{supp}(\bar{\alpha})$)

$$\epsilon_{k-1} - \epsilon_k = g(\alpha^{(k-1)}) - g(\alpha^{(k)}) \geq s(K,S)\epsilon_{k-1},$$

where s is given by

$$s(K,S) := \min\left\{\frac{\rho_+(K+1)}{L}, \frac{\rho_-(K+S)}{4S\rho_+(K+1)}\right\}$$

Therefore, $\epsilon_k \leq \epsilon_{k-1}(1 - s(K,S))$. Applying this inequality recursively we obtain $\epsilon_k \leq \epsilon_0\,(1 - s(K,S))^k$. Using the inequality $1 - s \leq \exp(-s)$ and rearranging we get that $\epsilon_k \leq \epsilon_0 \exp(-ks(K,S))$. When K satisfies (22), it can be guaranteed that $\epsilon_K \leq \epsilon$. □

A.4 The Proof of Theorem 3

The following simple lemma is useful in our analysis.

Lemma 3. *Denote by $f : [0,1] \to \mathbb{R}$ a quadratic function $f(x) = ax^2 + bx + c$ with $a > 0$ and $b \leq 0$. Then we have $\min_{x\in[0,1]} f(x) \leq c + \frac{b}{2}\min\{1, -\frac{b}{2a}\}$.*

Proof of Theorem 3. By the definition of restricted strongly-smooth (3) and the definition of $x^{(k)}$ in (18) it holds that

$$\begin{aligned}
& f(x^{(k)}) \\
\leq & \min_{\eta\in[0,1]} f((1-\eta)x^{(k-1)}+\eta u^{(k)}) \\
\leq & \min_{\eta\in[0,1]} f(x^{(k-1)})+\eta\langle\nabla f(x^{(k-1)}),u^{(k)}-x^{(k-1)}\rangle \\
& +\frac{\eta^2\rho_+(K+1)}{2}\|u^{(k)}-x^{(k-1)}\|^2 \\
\leq & \min_{\eta\in[0,1]} f(x^{(k-1)})+\eta\langle\nabla f(x^{(k-1)}),u^{(k)}-x^{(k-1)}\rangle \\
& +2\rho_+(K+1)A^2\eta^2 \\
\leq & \min_{\eta\in[0,1]} f(x^{(k-1)})+\eta\langle\nabla f(x^{(k-1)}),\bar{x}-x^{(k-1)}\rangle \\
& +2\rho_+(K+1)A^2\eta^2 \\
\leq & \min_{\eta\in[0,1]} f(x^{(k-1)})+\eta(f(\bar{x})-f(x^{(k-1)})) \\
& +2\rho_+(K+1)A^2\eta^2,
\end{aligned}$$

where the third inequality follows the boundness of set V, the forth inequality follows the update rule (8), and the last inequality follows the convexity of f.

Denote $\epsilon_k := f(x^{(k)})-f(\bar{x})$. Invoking Lemma 3 on the preceding inequality we get that

$$f(x^{(k)})\leq f(x^{(k-1)})+\frac{-\epsilon_{k-1}}{2}\min\left\{1,\frac{\epsilon_{k-1}}{4\rho_+(K+1)A^2}\right\},$$

which implies

$$\epsilon_k\leq\epsilon_{k-1}+\frac{-\epsilon_{k-1}}{2}\min\left\{1,\frac{\epsilon_{k-1}}{4\rho_+(K+1)A^2}\right\}.$$

When $\epsilon_{k-1}\geq 4\rho_+(K+1)A^2$, we obtain that $\epsilon_k\leq\frac{1}{2}\epsilon_{k-1}$, that is, ϵ_k converges towards $4\rho_+(K+1)A^2$ in geometric rate. Hence we need at most $\log_2\left[\frac{\epsilon_0}{4\rho_+(K+1)A^2}\right]$ to achieve this level of precision. Subsequently, we have

$$\epsilon_k\leq\epsilon_{k-1}-\frac{\epsilon_{k-1}^2}{8\rho_+(K+1)A^2}.$$

Invoking Lemma B.2 in (Shalev-Shwartz et al., 2010) we have

$$\epsilon_k\leq\frac{8\rho_+(K+1)A^2}{k+1},$$

and $\epsilon_k\leq\epsilon$ after at most $\frac{8\rho_+(K+1)A^2}{\epsilon}-1$ more steps. Altogether, FBS converges to the desired precision ϵ if

$$K\geq\log_2\left[\frac{\epsilon_1}{4\rho_+(K+1)A^2}\right]+\frac{8\rho_+(K+1)A^2}{\epsilon}-1.$$

This proves the validity of the Theorem. □

A.5 The Proof of Proposition 1

Proof. Let $\|X\|_2=\max\{\sigma_i,i=1,...,r\}$ be the spectral norm of matrix X. From the well known fact that spectral norm $\|\cdot\|_2$ and nuclear norm $\|\cdot\|_*$ are dual from one another, see, e.g., (Cai et al., 2010; Candès & Recht, 2009), we get that $\langle X,Y\rangle\leq\|X\|_2\|Y\|_*\leq\|X\|_2$. The equality holds for $Y=uv^T$ where u and v are the leading left singular vector and right singular vector of X, respectively. This proves the claim. □

A.6 The Proof of Proposition 2

Proof. Rewrite the matrices in terms of the eigen-decomposition of the positive semidefinite matrix $Y=U\Lambda U^T=\sum_{i=1}^n\lambda_i u_i u_i^T$, where λ a vector containing the diagonal entries of Λ. From the constraint of Y we have $\lambda\in\mathcal{S}:=\{\lambda_i\geq 0\text{ and }\sum_i\lambda_i\leq 1\}$. Insert this expression into the objective function

$$\max_{Y\in V_{psd}}\langle X,Y\rangle=\max_{\lambda\in\mathcal{S},U\in\mathcal{O}}\sum_{i=1}^n\lambda_i u_i^T X u_i, \tag{A.5}$$

where $\mathcal{O}$ is the set of orthonormal matrices also known as the Stiefel manifold. Let v be the leading eigenvector of X. Then $\sum_{i=1}^n\lambda_i u_i^T X u_i\leq\sum_{i=1}^n\lambda_i v^T X v\leq v^T X v$. Obviously the equality holds for $Y=vv^T$. This proves the result. □

References

Arora, S., Hazan, E., and Kale, S. Fast algorithms for approximate semidefinite programming using the multiplicative weights update method. In *FOCS*, pp. 339–348, 2005.

Cai, J., Candès, E., and Shen, Z. A singular value thresholding algorithm for matrix completion. *SIAM J. Optimiz.*, 20:1956–1982, 2010.

Candès, E. and Recht, B. Exact matrix completion via convex optimization. In *Foundations of Computational Mathematics*, pp. 717–772, 2009.

Clarkson, K. Coresets, sparse greedy approximation, and the frank-wolfe algorithm. In *SODA*, pp. 922–931, 2008.

Duchi, J., Shalev-Shwartz, S., Singer, Y., and Chandra, T. Efficient projections onto the ℓ_1-ball for learning in high dimensions. In *ICML*, pp. 272–279, 2008.

Frank, M. and Wolfe, P. An algorithm for quadratic programming. *Naval Res. Logist. Quart.*, 5:95–110, 1956.

Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29: 1189–1232, 2001.

Grubb, A. and Bagnell, J. Generalized boosting algorithms for convex optimization. In *ICML*, 2011.

Hazan, E. Sparse approximate solutions to semidefinite programs. In *LATIN*, pp. 306–316, 2008.

Jaggi, M. Convex optimization without projection steps. 2011. URL `http://arxiv.org/abs/1108.1170`.

Jaggi, M. and Sulovský, M. A simple algorithm for nuclear norm regularized problem. In *ICML*, 2010.

Johnson, R. and Zhang, T. Learning nonlinear functions using regularized greedy forest. 2011. URL `http://arxiv.org/abs/1109.0887`.

Kim, Y. and Kim, J. Gradient lasso for feature selection. In *ICML*, 2004.

Liu, G. C., Lin, Z. C., and Yu, Y. Robust subspace segmentation by low-rank representation. In *International Conference on Machine Learning (ICML)*, 2010.

Nesterov, Y. *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer, 2004.

Ni, Y. Z., Sun, J., Yuan, X.-T., Yan, S. C., and Cheong, L.-F. Robust low-rank subspace segmentation with semidefinite guarantees. 2010. URL `http://arxiv.org/abs/1009.3802`.

Pati, Y. C., Rezaiifar, R., and Krishnaprasad, P. S. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Proceedings of the 27th Annual Asilomar Conference on Signals, Systems and Computers*, 1993.

Schmidt, M., Berg, E., Friedlander, M., and Murphy, K. Optimizing costly functions with simple constraints: A limited-memory projected quasi-newton algorithm. In *AISTATS*, pp. 456–463, 2009.

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

Shalev-Shwartz, S., Gonen, A., and Shamir, O. Large-scale convex minimization with a low-rank constraint. In *ICML*, 2011.

Tewari, A., Ravikumar, P., and Dhillon, I. S. Greedy algorithms for structurally constrained high dimensional problems. In *Nueral Information Processing Systems*, 2011.

Tibshirani, R. Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B.*, 58(1):267–288, 1996.

Tropp, J. A. and Gilbert, A. C. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Trans. Info. Theory*, 53(12):4655–4666, 2007.

Tseng, P. On accelerated proximal gradient methods for convex-concave optimization. *submitted to SIAM Journal of Optimization*, 2008.

Zhang, T. Sequential greedy approximation for certain convex optimization problems. *IEEE Transactions on Information Theory*, 49(3):682–691, 2003.

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