Efficient k-Support-Norm Regularized Minimization via Fully Corrective Frank-Wolfe Method*

Bo Liu[†], Xiao-Tong Yuan[‡], Shaoting Zhang[§], Qingshan Liu[‡], Dimitris N. Metaxas[†]
[†]Department of Computer Science, Rutgers, The State University of New Jersey
[‡]Jiangsu Province Key Laboratory of Big Data Analysis Technology,
Nanjing University of Information Science and Technology
[§]Department of Computer Science, University of North Carolina at Charlotte
lb507@cs.rutgers.edu

Abstract

The k-support-norm regularized minimization has recently been applied with success to sparse prediction problems. The proximal gradient method is conventionally used to minimize this composite model. However it tends to suffer from expensive iteration cost as the proximity operator associated with k-support-norm needs exhaustive searching operations and thus could be time consuming in large scale settings. In this paper, we reformulate the k-support-norm regularized formulation into an identical constrained formulation and propose a fully corrective Frank-Wolfe algorithm to minimize the constrained model. Our method is inspired by an interesting observation that the convex hull structure of the k-support-norm ball allows the application of Frank-Wolfe-type algorithms with low iteration complexity. The convergence behavior of the proposed algorithm is analyzed. Extensive numerical results in learning tasks including logistic regression and matrix pursuit demonstrate the substantially improved computational efficiency of our algorithm over the state-of-the-art proximal gradient algorithms.

1 Introduction

In many sparsity inducing machine learning problems, a common practice is to use the ℓ_1 norm as a convex relaxation of the model parameter cardinality constraint. The ℓ_1 norm, however, tends to shrink excessive number of variables to zeros, regardless the potential correlation among the variables. In order to alleviate such an over-shrinkage issue of ℓ_1-norm, numerous methods including elastic net [Zou and Hastie, 2005] and trace Lasso [Grave *et al.*, 2011] have been proposed in literature. All of these methods tend to smooth the output parameters by averaging similar features rather than selecting out a single one. More recently, k-support-norm $\|\cdot\|_k^{sp}$ is proposed as a new alternative that provides the tightest convex relaxation of cardinality on the Euclidean norm unit ball [Argyriou *et al.*, 2012]. Formally, the k-support-norm of a vector $w \in \mathbb{R}^p$ is defined as

$$\|w\|_k^{sp} := \min\left\{\sum_{g\in\mathcal{G}_k} \|v_g\|_2 : \mathrm{supp}(v_g) \subseteq g, w = \sum_{g\in\mathcal{G}_k} v_g\right\},$$

where $\mathcal{G}_k$ denotes the set of all subsets of $\{1, 2, ..., p\}$ of cardinality at most k. More properties of k-support-norm are analyzed in [Argyriou *et al.*, 2012].

As a regularizer, the k-support-norm is characterized by simultaneously selecting a few relevant groups and penalizing the ℓ_2-norm of the selected individual groups. The following k-support-norm regularized model was considered in [Argyriou *et al.*, 2012] for sparse prediction tasks:

$$\min_w f(w) + \lambda(\|w\|_k^{sp})^2, \tag{1}$$

where $f(w)$ is a convex and differentiable objective function. The parameter k is regarded as an upper bound estimation of the number of non-zero elements in w. It has been shown that this model leads to improved learning guarantees as well as better algorithmic stability [Argyriou *et al.*, 2012].

Motivation One issue that hinders the applicability of the k-support-norm regularized model (1) is its high computational complexity in large scale settings. Indeed, proximal gradient methods are conventionally used for optimizing the composite minimization problem in (1) [Argyriou *et al.*, 2012; Lai *et al.*, 2014; Eriksson *et al.*, 2015]. Given the gradient vector, the per-iteration computational cost of proximal gradient methods is dominated by an proximity operator of the following form:

$$w^* = \arg\min_w \frac{1}{2}\|w - v\|_2^2 + \lambda(\|w\|_k^{sp})^2. \tag{2}$$

In the work of [Argyriou *et al.*, 2012], a closed form solution of (2) was derived based on an exhaustive search strategy. However, the complexity of such a strategy is $O(p(k+\log p))$ which is computationally expensive when p is large. Despite significant speed-ups have been reported in [Lai *et al.*, 2014; Eriksson *et al.*, 2015] by using binary search, those methods are still expensive for large scale problems.

It has been known that the k-support-norm ball $\mathcal{B}_\varpi^{k,sp} := \{w \in \mathbb{R}^p : \|w\|_k^{sp} \le \varpi\}$ is equivalent to the *convex hull*

*These match the formatting instructions of IJCAI-07. The support of IJCAI, Inc. is acknowledged.

of $\mathcal{C}^{(2)}_{k,\varpi} = \{w \in \mathbb{R}^p : \|w\|_0 \leq k, \|w\|_2 \leq \varpi\}$. In this sense, k-support-norm ball $\mathcal{B}^{k,sp}_{\varpi}$ provides a convex envelope of the nonconvex set $\mathcal{C}^{(2)}_{k,\varpi}$. Recently, Frank-Wolfe-type methods for minimizing a convex objective over a convex hull have received wide interests in optimization and machine learning [Zhang, 2003; Shalev-Shwartz *et al.*, 2010; Yuan and Yan, 2013]. These algorithms have been shown to achieve satisfying trade-off between convergence rate and iteration complexity. In this paper, inspired by the success of these algorithms, we consider applying the Frank-Wolfe method to solve the composite optimization problem (1).

Challenge and Contribution In this paper we propose to convert the k-support-norm regularized formulation into an identical constrained formulation by introducing an augmented variable as the bound of the regularizer $(\|w\|^{sp}_k)^2$. We then develop a fully corrective variant of the Frank-Wolfe algorithm to optimize the augmented model. The proposed algorithm is called k-FCFW in the following context. The convergence rate of the proposed algorithm is analyzed. Particularly, we prove that the proposed algorithm converges linearly under proper conditions. Our this result is stronger than the sublinear rate of convergence obtained in [Harchaoui *et al.*, 2015] for norm regularized minimization with Frank-Wolfe methods. The advantage of the proposed algorithm over prior algorithms is demonstrated by empirical results in various learning tasks.

2 Related Work

2.1 k-Support-Norm Regularized Learning

The k-support-norm regularized learning models have been extensively studied in machine learning and computer vision. Multiple k-support-norm regularized convex models were investigated in [Blaschko, 2013]. The applications of k-support-norm to various computer vision problems have been explored in [Lai *et al.*, 2014]. The k-support-norm was applied to generalized dantzig selector for linear models [Chatterjee *et al.*, 2014]. In this paper the proposed algorithm applies to first-order k-support-norm regularized minimization problem, which is different from the squared k-support-norm that we consider in this work and [Argyriou *et al.*, 2012; Lai *et al.*, 2014; Eriksson *et al.*, 2015]. The authors of [Belilovsky *et al.*, 2015b] showed that it is helpful to use k-support-norm regularization in fMRI data analysis including classification, regression and data visualization. In [Belilovsky *et al.*, 2015a], total variation penalty is incorporated in the k-support framework and applied in image and neuroscience data analysis.

2.2 Frank-Wolfe Method

The history of Frank-Wolfe method dates back to [Frank and Wolfe, 1956] for constrained optimization problem

$$\min_{w} f(w) \quad s.t. \quad w \in D,$$

where D is a polytope convex set. It is also known as conditional gradient method. Due to the potential of efficiency improvement when applied in solving minimization problem with some forms of constraint, recently there is an increasing trend to revisit and restudy this method. The detail of original Frank-Wolfe method, its variants and algorithm analysis can be found in [Jaggi, 2013; Garber and Hazan, 2014]. Frank-Wolfe-type methods have been developed to solve numerous optimization problems including structural SVM [Lacoste-Julien *et al.*, 2013], trace norm regularization problem [Dudik *et al.*, 2012] and atomic norm constrained problem [Rao *et al.*, 2015]. In sparse learning, a number of Frank-Wolfe-type methods, e.g., forward greedy selection [Shalev-Shwartz *et al.*, 2010] and gradient Lasso [Kim and Kim, 2004], have been developed to solve sparsity constrained problems. In the context of boosting classification, the restricted gradient projection algorithms stated in [Grubb and Bagnell, 2011] is essentially a forward greedy selection method over ℓ_2-functional space.

Algorithm 1: k-FCFW Algorithm for k-support-norm regularized problem.

Input : $f(x)$, k, λ.
Initialization: $w^{(0)} = 0$, $\theta^{(0)} = 0$, $U = w^{(0)}$, $V = \theta^{(0)}$.
for $t = 1, 2, \ldots$ **do**
(S1) Compute $\nabla f(w^{(t-1)})$.
(S2) Solve the linear projection problem

$$\{u^{(t)}, v^{(t)}\} = \arg\min_{u,v} \langle \nabla f(w^{(t-1)}), u\rangle + \lambda v \quad \text{s.t.} \quad (\|u\|^{sp}_k)^2 \leq v. \tag{3}$$

(S3) Update

$$U^{(t)} = [U^{(t-1)}, u^{(t)}], V^{(t)} = [V^{(t-1)}, v^{(t)}].$$

(S4) Compute

$$\alpha^{(t)} = \min_{\alpha \in \triangle_t} f(U^{(t)}\alpha) + \lambda V^{(t)}\alpha, \tag{4}$$

where $\triangle_t = \{\alpha \in \mathbb{R}^{t+1} : \alpha \geq 0, \|\alpha\|_1 = 1\}$.
(S5) Update $w^{(t)} = U^{(t)}\alpha^{(t)}$, $\theta^{(t)} = V^{(t)}\alpha^{(t)}$.
end
Output: $w^{(t)}$.

3 The k-FCFW method for k-Support Norm Regularized Minimization

To apply the Frank-Wolfe method, we reformulate (1) into a constrained optimization problem:

$$\min_{w,\theta} G(w, \theta; \lambda) := f(w) + \lambda\theta \quad \text{s.t.} \quad (\|w\|^{sp}_k)^2 \leq \theta.$$

Here θ is an augmented variable bounding the term $(\|w\|^{sp}_k)^2$. We introduce in §3.1 the k-FCFW algorithm for solving the above constrained formulation. The convergence rate of k-FCFW will be analyzed in §3.2.

3.1 Algorithm Description

The k-FCFW algorithm for k-support-norm regularized minimization is outlined in Algorithm 1. As is known that the k-support-norm ball $\mathcal{B}^{k,sp}_{\varpi} := \{w \in \mathbb{R}^p : \|w\|^{sp}_k \leq \varpi\}$ is

equivalent to the convex hull of $\mathcal{C}^{(2)}_{k,\varpi} = \{w \in \mathbb{R}^p : \|w\|_0 \leq k, \|w\|_2 \leq \varpi\}$. That is,

$$\mathcal{B}^{k,sp}_{\varpi} = \text{conv}(\mathcal{C}^{(2)}_{k,\varpi}) = \left\{ \sum_{w \in \mathcal{C}^{(2)}_{k,\varpi}} \alpha_w w : \alpha_w \geq 0, \sum_w \alpha_w = 1 \right\}.$$

Therefore, given $v > 0$, the optimal u of (3) admits the following close-form solution:

$$u = -\frac{\sqrt{v}\nabla_k f(w^{(t-1)})}{\|\nabla_k f(w^{(t-1)})\|}, \tag{5}$$

where $\nabla_k f(w^{(t-1)})$ denotes a truncated version of $\nabla f(w^{(t-1)})$ with its top k (in magnitude) entries preserved. By substituting this back to (3) we get $v^{(t)}$ through the following quadratic program:

$$v^{(t)} = \arg\min_{v>0} -\|\nabla_k f(w^{(t-1)})\|\sqrt{v} + \lambda v.$$

Obviously $v^{(t)} = \left(\frac{\|\nabla_k f(w^{(t-1)})\|}{2\lambda}\right)^2$, and thus,

$$u^{(t)} = -\frac{\sqrt{v^{(t)}}\nabla_k f(w^{(t-1)})}{\|\nabla_k f(w^{(t-1)})\|} = -\frac{\nabla_k f(w^{(t-1)})}{2\lambda}.$$

At each time instance t, we record all previous updates in $U^{(t)} = \{w^{(0)}, u^{(1)}, ..., u^{(t)}\}$ and $V^{(t)} = \{\theta^{(0)}, v^{(1)}, v^{(2)}, ..., v^{(t)}\}$. At the t-th iteration, the optimal value of $w^{(t)}$ and $\theta^{(t)}$ are jointly searched on the convex hull define by $U^{(t)}$ and $V^{(t)}$. The subproblem (4) of estimating $\alpha^{(t)}$ is a simplex constrained smooth minimization problem. The scale of such a problem is dominated by the value of t. This subproblem can be solved via off-the-shelf algorithms such as the projected quasi-Newton (PQN) method [Schmidt *et al.*, 2009] as used in our implementation.

It is noteworthy that the subproblem (3) is equivalent to the following k-support-norm regularized linear program:

$$u^{(t)} = \arg\min_u \langle \nabla f(w^{(t-1)}), u\rangle + \lambda(\|u\|^{sp}_k)^2.$$

This is in contrast to the proximal gradient method which solves the quadratic proximity operator (2) at each iteration. Apparently the former is less expensive to solve than the latter which involves exhaustive search. In addition, when t is of moderate value and warm start is adopted to initialize the parameters, the subproblem (4) can often be accurately solved with very few iterations. In sparse learning problem the k value is often much smaller than model parameter dimension hence the overall computational cost of k-FCFW is expected to be lower than the proximal gradient algorithms.

Given a constant radius $\varpi_c > 0$, the k-support-norm constrained minimization problem

$$\min_w f(w) \quad \text{s.t.} \quad \|w\|^{sp}_k \leq \varpi_c \tag{6}$$

is essentially a special case of the regularized form by directly assigning $\theta = \varpi_c^2$. The proposed algorithm can be easily modified to solve the constrained model (6).

3.2 Convergence Analysis

To analyze the model convergence, we need the following key technical conditions imposed on the curvature of the objective function f restricted on sparse subspaces.

Definition 1 (Smoothness and restricted strong convexity)**.** *We say f is L-*smooth *if there exists a positive constant L such that for any w and w',*

$$f(w') - f(w) - \langle \nabla f(w), w' - w\rangle \leq \frac{L}{2}\|w - w'\|^2. \tag{7}$$

*We say f is $\rho(s)$-*strongly convex *at sparsity level s, if there exists positive constants $\rho(s)$ such that for any $\|w - w'\|_0 \leq s$,*

$$f(w') - f(w) - \langle \nabla f(w), w' - w\rangle \geq \frac{\rho(s)}{2}\|w - w'\|^2. \tag{8}$$

In the analysis to follow, we define

$$F(w; \lambda) := f(w) + \lambda(\|w\|^{sp}_k)^2,$$

and

$$\bar{w} = \arg\min_w F(w; \lambda).$$

Let $\bar{s} = \|\bar{w}\|_0$ and $\bar{\theta} = (\|\bar{w}\|^{sp}_k)^2$. Consider the radius r given by

$$r = \max\left\{ \frac{\|\nabla_k f(w)\|}{2\lambda} : F(w; \lambda) \leq F(w^{(0)}; \lambda) \right\}.$$

We now analyze the convergence of Algorithm 1. Before presenting the main result, we need some preliminaries.

Lemma 1. *There exist $\bar{U} = [\bar{u}_1, ..., \bar{u}_{\bar{l}}] \in \mathbb{R}^{p\times\bar{l}}$ with $\|\bar{u}_i\|_0 \leq k$, $\|\bar{u}_i\|_2 = \sqrt{\bar{\theta}}$, and $\bar{\alpha} \in \triangle_{\bar{l}}$ such that*

$$\bar{w} = \bar{U}\bar{\alpha}.$$

Please refer to Appendix A.1 for the proof of Lemma 1. In the following analysis, we will consider such a decomposition $\bar{w} = \bar{U}\bar{\alpha}$ as guaranteed by Lemma 1. Given a matrix M, we write its mathcal version $\mathcal{M}$ as a vector set consisting of the columns of M. Similarly, given a set $\mathcal{M}$ of vectors of the same size, we denote M be a matrix whose columns are the elements of $\mathcal{M}$. Let $\sigma_{\min}(M)$ be the smallest singular value of matrix M. We establish in the following theorem a linear rate of convergence for Algorithm 1.

Theorem 1. *Let $s = \max_t \|w^{(t)}\|_0$. Let $\mathcal{M}^{(t)} = \bar{\mathcal{U}} \cup \mathcal{U}^{(t)}$. Assume that there exists a $\bar{\beta} > 0$ such that $\sigma_{\min}(M^{(t)}) \geq \bar{\beta}$ for all t. Assume that f is L-smooth and $\rho(s + \bar{s})$-strongly convex. Given $\epsilon > 0$, let us run Algorithm 1 with $t \geq \frac{1}{\zeta}\ln\left[\frac{F(w^{(0)};\lambda) - F(\bar{w};\lambda)}{\epsilon}\right]$ where $\zeta := \min\left\{\frac{\rho(s+\bar{s})\bar{\beta}}{4lLr^2}, \frac{1}{2}\right\}$. Then Algorithm 1 will output $w^{(t)}$ satisfying*

$$F(w^{(t)}; \lambda) \leq F(\bar{w}; \lambda) + \epsilon.$$

The proof of Theorem 1 is given in Appendix A.2.

Remark 1. *In general constrained minimization problems, the Frank-Wolfe method is known to have $O(\frac{1}{t})$ convergence rate [Jaggi, 2013] and $O(\frac{1}{t^2})$ if the constraint set is*

strongly-convex [Garber and Hazan, 2014]. Several linear convergence guarantees are established by adding various specific assumptions to either constraint set or loss function in literatures such as [Beck and Teboulle, 2004; Ñanculef et al.*, 2014], but they are not directly applicable to our problem. In a recent work of [Lacoste-Julien and Jaggi, 2015], a global linear convergence rate is proved for the constrained optimization on polytope. Their analysis, however, does not fit for our algorithm as the constraint* $(\|w\|_k^{sp})^2 \leq v$ *is a* k*-support-norm cone, rather than a polytope. This imposes extra challenges in analysis. Another related work is [Harchaoui* et al.*, 2015] which also applies Frank-Wolfe method to regularized minimization problems. However it is restrictive as it requires an estimation of the bound of the regularizer which is hard to know in practice. Also, the sublinear convergence rate established in that paper is weaker than the linear rate proved in Theorem 1.*

Remark 2. *In Algorithm 1 we have required the subproblem* (4) *in Step (S4) to be solved exactly. This could be be computationally demanding if the objective function* f *is highly nonlinear and* t *is relatively large. Instead of solving the subproblem* (4) *exactly, a more realistic option in practice is to find a suboptimal solution up to a precision* $\varepsilon > 0$ *w.r.t. the first-order optimality condition. That is,* $\{w^{(t)}, \theta^{(t)}\}$ *satisfy for any* $w = U^{(t)}\alpha$ *and* $\theta = V^{(t)}\alpha$*,*

$$\langle \nabla f(w^{(t-1)}), w - w^{(t-1)} \rangle + \lambda(\theta - v^{(t-1)}) \geq -\varepsilon.$$

Following the similar arguments in the proof of Theorem 1 we can be prove that $F(w^{(t)}; \lambda) \leq F(\bar{w}; \lambda) + \epsilon + O(\varepsilon)$ *after* $t = O\left(\ln(\frac{1}{\epsilon})\right)$ *steps of iteration. In other words, the optimization error of the subproblem* (4) *does not accumulate during iteration.*

4 Experiments

This section is devoted to showing the empirical performance of k-FCFW and comparing it to the state-of-the-art algorithms for k-support-norm regularized minimization. All the considered algorithms are implemented in Matlab and tested on a computer equipped with 3.0GHz CPU and 32GB RAM.

4.1 k-Support-Norm L_2-Logistic Regression

Given a binary training set $\{x_m, y_m\}_{m=1}^M, x_m \in \mathbb{R}^p, y_m \in \{-1, 1\}$, the k-support-norm regularized logistic regression problem is formulated as

$$\min_w F(w) = \sum_{m=1}^{M} \ell(w; x_m, y_m) + \frac{\tau}{2}\|w\|^2 + \lambda(\|w\|_k^{sp})^2, \quad (9)$$

where $\ell(w; x_m, y_m) = \log\left(1 + \exp(-y_m w^\top x_m)\right)$ is the logistic loss on a training pair (x_m, y_m). The parameter τ controls the strong convexity of the loss function.

We test the algorithm efficiency on a synthetic dataset. The model parameter ground truth w_{gt} is designed to be a p-dimension vector as follows:

$$w_{gt} = [\underbrace{1, 1, \cdots, 1}_{p'}, \underbrace{0, 0, \cdots, 0}_{p-p'}].$$

Within the supporting set we partition the feature dimension into g groups. Each group of $\{x_{m,1}, x_{m,2}, ..., x_{m,p'/g}\}$, ..., $\{x_{m,(g-1)p'/g+1}, ..., x_{m,p'}\}$ follows normal distribution that the mean value of each group is drawn from $\mathcal{N}(0, 1)$. Other dimensions are drawn from $\mathcal{N}(0, 1)$ as noise. Finally x_m is normalized by $x_m = x_m/\|x_m\|$. The data label $\{y_m\}_{m=1}^M$ follows Bernoulli distribution with probability $\mathbb{P}(y_m = 1|x_m) = \frac{\exp(w_{gt}^\top x_m)}{1+\exp(w_{gt}^\top x_m)}$. The task is designed as selecting the top k most discriminative features for classification using logistic regression model through solving (9).

We produce the training data by setting $M = 500$, $p = 10^6$, $g = 20$, $p' = 10000$, $k = 2000, 4000, 6000, 8000$ and 10000, respectively. λ is selected by grid search according to the testing result on a validation set of size 100. We set the termination criterion as $\frac{|F(w^{(t)}) - F(w^{(t-1)})|}{F(w^{(t-1)})} \leq 10^{-4}$. We replicate the experiment 10 times and report the mean of the numerical results.

We compare the efficiency of k-FCFW with three state-of-the-art proximal gradient methods: (1) the Box Norm solver (denoted by BN) proposed in [McDonald *et al.*, 2014]; (2) the binary search based solver proposed in [Lai *et al.*, 2014] (denoted by BS); and (3) the solver proposed in [Eriksson *et al.*, 2015] which tries to find the active set (AS) by a two-step binary searching strategy. All of these proximal gradient solvers are implemented in the framework of FISTA [Beck and Teboulle, 2009].

The running time of the considered algorithms is shown in Figure 1(a). It can be observed that our method is significantly faster than all the three comparing solvers.

We also compare the efficiency of k-FCFW with ADMM [Parikh and Boyd, 2013] which is another popular framework for regularized minimization problems. Since AS has been observed to be superior to the other considered proximity operator solvers, we equip ADMM with AS as its proximity operator solver. The running time curves of ADMM-AS are drawn in Figure 1(a). Clearly, ADMM-AS is inferior to k-FCFW and the proximal gradient algorithms as well. Actually, we observe that ADMM-AS fails to converge to the desired accuracy given maximum number of iterations. In Figure 1(b), we plot the convergence curves of the considered algorithms under $k = 2000$ and 10000. It can be observed that our method needs significant less number of iterations to reach comparable optimization accuracy.

4.2 k-Support-Norm Matrix Pursuit

In this group of experiments, we apply the proposed method to the k-support-norm regularized matrix pursuit problem. In many graph based machine learning algorithms such as semi-supervised classification and subspace segmentation, a key step is learning the sample affinity matrix [Liu *et al.*, 2010; Yuan and Li, 2014]. Matrix pursuit has been proved to be an effective approach. The results of [Lai *et al.*, 2014] indicate that the k-support-norm regularized matrix pursuit method achieves superior performance in various applications. The k-support-norm regularized matrix pursuit is formulated as:

$$\min_{W \in \mathbb{R}^{n \times n}} \frac{1}{2}\|X - XW\|_F^2 + \lambda(\|vec(W)\|_k^{sp})^2, \quad (10)$$

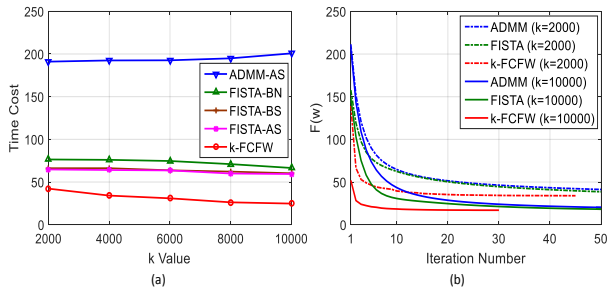

Figure 1: Results on synthetic dataset: (a) Running time (in second) curves of the considered comparing methods under different values of k. (b) Convergence curves of the considered methods under $k = 2000$ and 10000. All the curves are drawn from the starting point $F(w^{(1)})$.

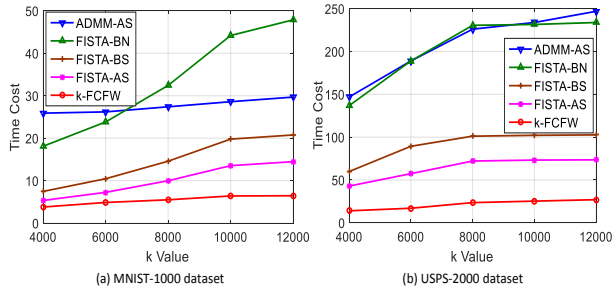

Figure 2: Running time (in second) curves of the considered methods on (a) MNIST-1000 dataset and (b) USPS-2000 dataset.

where $X \in \mathbb{R}^{d\times n}$ is the data matrix with n samples in d-dimension space and $vec(W)$ denotes the vectorization of W.

The MNIST [LeCun *et al.*, 1998] and USPS [Hull, 1994] datasets are adopted for testing. For MNIST dataset, we resize each image into 14×14 then normalize the gray value into $[0, 1]$. The pixel values are then vectorized as image feature. The USPS dataset is preprocessed by [Cai *et al.*, 2011]. Each image is represented by a 256-dimension feature vector and the feature values are normalized into $[-1, 1]$. Each image of both datasets are further normalized to be a unit vector. We select 100 images per digit from MNIST and 200 images per digit from USPS hence the size of datasets are 1000 and 2000, respectively. We use the same optimization termination criterion as in the previous experiment. The algorithms are tested under $k = 4000, 6000, 8000, 10000$ and 12000 in the k-support-norm. The regularization parameter is set to be $\lambda = 20$.

We first compare k-FCFW with three FISTA algorithms that respectively employ proximity operator solver BN, BS and AS. The comparison of average time cost over 10 replications is illustrated in Figure 2. The time cost curves of ADMM-AS are also shown in Figure 2. The convergence curves of the considered methods evaluated by $F(W)$ when $k = 4000$ and 12000 are shown in Figure 3. From the results we can see that in these cases, k-FCFW is the most efficient one for optimization. As the value of k increases, the efficiency advantage of k-FCFW gets more significant.

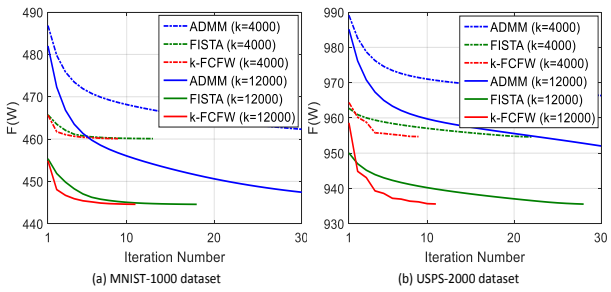

Figure 3: The convergence curves of considered methods on (a) MNIST-1000 and (b) USPS-2000 datasets under $k = 4000$ and 12000. The starting point of each curve is $F(W^{(1)})$.

5 Conclusion

In this paper, we proposed k-FCFW as a fully corrective Frank-Wolfe algorithm for optimizing the k-support-norm regularized loss minimization problem. We have established a linear rate of convergence for the proposed algorithm, which to our knowledge is new for Frank-Wolfe-type algorithms when applied to composite formulation. Comparing to the conventionally adopted proximal gradient algorithms and ADMM, k-FCFW has superior rate of convergence in theory and practice. Numerical results in logistic regression and matrix pursuit applications confirmed that k-FCFW is significantly more efficient than several state-of-the-art proximal gradient descent methods, especially in large scale settings. To conclude, both theoretical analysis and empirical observations suggest that k-FCFW is a computationally attractive alternative to the proximal gradient algorithms for solving the k-support-norm regularized minimization problems. As a future study in this line, we will investigate the generalization of k-FCFW to generic group-sparsity-norm regularized minimization problems of which the model considered in this paper is a special case.

A Appendix

A.1 Proof of Lemma 1

Proof. Consider

$$\tilde{\mathcal{U}} = \arg\min_{\mathcal{U}} \left\{ \sum_{u\in\mathcal{U}} \|u\|_2 : \|u\|_0 \le k, \bar{w} = \sum_{u\in\mathcal{U}} u \right\}.$$

Let $\bar{l} = |\tilde{\mathcal{U}}|$ and $\tilde{\mathcal{U}} = \{\tilde{u}_i\}_{i=1}^{\bar{l}}$. Based on the definition of k-support-norm we have that $\bar{w} = \sum_i \tilde{u}_i$ and $\sum_i \|\tilde{u}_i\|_2 = \|\bar{w}\|_k^{sp} = \sqrt{\bar{\theta}}$. We construct $\bar{U} = [\bar{u}_1, ..., \bar{u}_{\bar{l}}]$ with $\bar{u}_i$ defined by $\bar{u}_i = \sqrt{\bar{\theta}}\tilde{u}_i/\|\tilde{u}_i\|$. Then we can show that $\bar{w}$ admits a decomposition of $\bar{w} = \bar{U}\bar{\alpha}$ with some $\bar{\alpha}$ lies in a $\bar{l}$-dimensional simplex $\triangle_{\bar{l}}$. Indeed, since $\bar{w} = \sum_i \tilde{u}_i = \sum_i \bar{u}_i(\|\tilde{u}_i\|/\sqrt{\bar{\theta}})$, we may define $\bar{\alpha}_i = \|\tilde{u}_i\|/\sqrt{\bar{\theta}}$ such that $\sum_i \bar{\alpha}_i = 1$. □

A.2 Proof of Theorem 1

Proof. From the definition of $G(w, \theta; \lambda)$, the step (S4) of Algorithm 1 and the assumption that $f(w)$ is the L-smooth function, we have

$$
\begin{aligned}
&G(w^{(t)},\theta^{(t)};\lambda)\\
=&f(U^{(t)}\alpha^{(t)})+\lambda V^{(t)}\alpha^{(t)}\\
\le&f((1-\eta)U^{(t-1)}\alpha^{(t-1)}+\eta u^{(t)})+\\
&\lambda((1-\eta)V^{(t-1)}\alpha^{(t-1)}+\eta v^{(t)})\\
=&f((1-\eta)w^{(t-1)}+\eta u^{(t)})+\lambda((1-\eta)\theta^{(t-1)}+\eta v^{(t)})\\
\le&f(w^{(t-1)})+\eta\Gamma(u^{(t)})+2\eta^2r^2L+\lambda[(1-\eta)\theta^{(t-1)}+\eta v^{(t)}]\\
=&f(w^{(t-1)})+\lambda\theta^{(t-1)}+\eta\Phi(u^{(t)},v^{(t)})+2\eta^2r^2L\\
=&G(w^{(t-1)},\theta^{(t-1)};\lambda)+\eta\Phi(u^{(t)},v^{(t)})+2\eta^2r^2L.
\end{aligned}
$$

where

$$
\begin{aligned}
\Gamma(u^{(t)})&=\langle\nabla f(w^{(t-1)}),u^{(t)}-w^{(t-1)}\rangle,\\
\Phi(u^{(t)},v^{(t)})&=\Gamma(u^{(t)})+\lambda(v^{(t)}-\theta^{(t-1)}).
\end{aligned}
$$

For simplicity, let us now denote $x^{(t)}=[w^{(t)};\theta^{(t)}]\in\mathbb{R}^{d+1}$ as the concatenation of $w^{(t)}$ and $\theta^{(t)}$. We define $\bar{V}=[\bar{\theta},...,\bar{\theta}]\in\mathbb{R}^{\bar{l}}$. Similarly, we denote $\bar{X}=[\bar{U};\bar{V}]\in\mathbb{R}^{(p+1)\times\bar{l}}$ and $X^{(t)}=[U^{(t)};V^{(t)}]\in\mathbb{R}^{(p+1)\times t}$. By writing $G(x^{(t)};\lambda)=G(w^{(t)},\theta^{(t)};\lambda)$, the preceding inequality can be equivalently written as

$$
\begin{aligned}
G(x^{(t)};\lambda)\le&G(x^{(t-1)};\lambda)+\eta\langle\nabla G(x^{(t-1)}),x^{(t)}-x^{(t-1)}\rangle\\
&+2\eta^2r^2L.
\end{aligned}
$$

Let $\mathcal{X}^c:=\bar{\mathcal{X}}\setminus\mathcal{X}^{(t-1)}$. Assume $\mathcal{X}^c\neq\emptyset$. From the update rule of $\{\theta^{(t)},v^{(t)}\}$ in (S2) we know the following inequality holds for any $x\in\mathcal{X}^c$:

$$
\langle\nabla G(x^{(t-1)}),x^{(t)}-x^{(t-1)}\rangle\le\langle\nabla G(x^{(t-1)},\lambda),x-x^{(t-1)}\rangle.
$$

Let $\xi=\sum_{x\in\mathcal{X}^c}\bar{\alpha}_x$. From the above two inequalities we get

$$
\begin{aligned}
\xi G(x^{(t)};\lambda)\le&\xi G(x^{(t-1)};\lambda)+\eta[\sum_{x\in\mathcal{X}^c}\bar{\alpha}_x\langle\nabla G(x^{(t-1)};\lambda),x\rangle\\
&-\xi\langle\nabla G(x^{(t-1)};\lambda),x^{(t-1)}\rangle]+2\eta^2r^2\xi L.
\end{aligned}
\tag{11}
$$

Since $\sum_{x\in\mathcal{X}^{(t-1)}}\bar{\alpha}_x/(1-\xi)=1$, from the optimality of $\alpha^{(t-1)}$ (see the step (S4)) we can derive that

$$
\langle\nabla G(x^{(t-1)};\lambda),\sum_{x\in\mathcal{X}^{(t-1)}}\bar{\alpha}_x x/(1-\xi)-x^{(t-1)}\rangle\ge0. \tag{12}
$$

Note that $\alpha_x^{(t-1)}=0$ for $x\notin\mathcal{X}^{(t-1)}$ and $\bar{\alpha}_x=0$ for $x\notin\bar{\mathcal{X}}$. Therefore

$$
\begin{aligned}
&\sum_{x\in\mathcal{X}^c}\bar{\alpha}_x\langle\nabla G(x^{(t-1)};\lambda),x\rangle\\
=&\sum_{x\in\mathcal{X}^c}\langle\nabla G(x^{(t-1)};\lambda),\bar{\alpha}_x x-(1-\xi)\alpha_x^{(t-1)}x\rangle\\
\le&\sum_{x\in\mathcal{X}^{(t-1)}\cup\bar{\mathcal{X}}}\langle\nabla G(x^{(t-1)};\lambda),\bar{\alpha}_x x-(1-\xi)\alpha_x^{(t-1)}x\rangle\\
=&\langle\nabla G(x^{(t-1)};\lambda),\bar{x}-(1-\xi)x^{(t-1)}\rangle\\
=&\langle\nabla G(x^{(t-1)};\lambda),\bar{x}-x^{(t-1)}\rangle+\xi\langle\nabla G(x^{(t-1)};\lambda),x^{(t-1)}\rangle,
\end{aligned}
$$

where the inequality follows (12). Combining the preceding inequality with (8) we obtain

$$
\begin{aligned}
&\sum_{x\in\mathcal{X}^c}\bar{\alpha}_x\langle\nabla G(x^{(t-1)};\lambda),x\rangle-\xi\langle\nabla G(x^{(t-1)};\lambda),x^{(t-1)}\rangle\\
\le&\langle\nabla G(x^{(t-1)};\lambda),\bar{x}-x^{(t-1)}\rangle\\
=&\langle\nabla f(w^{(t-1)}),\bar{w}-w^{(t-1)}\rangle+\lambda(\bar{\theta}-\theta^{(t-1)})\\
\le&f(\bar{w})-f(w^{(t-1)})-\frac{\rho(s+\bar{s})}{2}\|w^{(t-1)}-\bar{w}\|^2\\
&+\lambda(\bar{\theta}-\theta^{(t-1)})\\
\le&G(\bar{x};\lambda)-G(x^{(t-1)};\lambda)\\
&-\frac{\rho(s+\bar{s})}{2}\|\sum_{u\in\mathcal{U}^{(t-1)}}\alpha_u u-\sum_{u\in\bar{\mathcal{U}}}\bar{\alpha}_u u\|^2\\
\le&G(\bar{x})-G(x^{(t-1)})-\frac{\rho(s+\bar{s})\bar{\beta}}{2}\sum_{u\in\bar{\mathcal{U}}\setminus\mathcal{U}^{(t-1)}}\bar{\alpha}_u^2\\
\le&G(\bar{x})-G(x^{(t-1)})-\frac{\rho(s+\bar{s})\bar{\beta}\xi^2}{2\bar{l}},
\end{aligned}
$$

where the last inequality follows $\sum_{u\in\bar{\mathcal{U}}\setminus\mathcal{U}^{(t-1)}}\bar{\alpha}_u^2\ge(\sum_{u\in\bar{\mathcal{U}}\setminus\mathcal{U}^{(t-1)}}\bar{\alpha}_u)^2/\bar{l}$. By combining the above with (11) we get

$$
\begin{aligned}
\xi G(x^{(t)};\lambda)\le&\xi G(x^{(t-1)};\lambda)-\eta[G(x^{(t-1)};\lambda)-G(\bar{x};\lambda)+\\
&\frac{\rho(s+\bar{s})\bar{\beta}\xi^2}{2\bar{l}}]+2\eta^2r^2\xi L.
\end{aligned}
\tag{13}
$$

Let us choose $\eta=\xi\zeta\le1$ in the above inequality with $\zeta:=\min\left\{\frac{\rho(s+\bar{s})\bar{\beta}}{4\bar{l}Lr^2},\frac{1}{2}\right\}$. Then we have

$$
G(x^{(t)};\lambda)\le G(x^{(t-1)};\lambda)-\zeta(G(x^{(t-1)};\lambda)-G(\bar{w};\lambda)).
$$

Let us denote $\epsilon_t:=G(x^{(t)};\lambda)-G(\bar{x};\lambda)$. Applying this inequality recursively we obtain $\epsilon_t\le\epsilon_0(1-\zeta)^t$. Using the inequality $1-x\le\exp(-x)$ and rearranging we get that $\epsilon_t\le\epsilon_0\exp(-\zeta t)$. When $t\ge\frac{1}{\zeta}\ln\frac{\epsilon_0}{\epsilon}$, it can be guaranteed that $\epsilon_t\le\epsilon$. The desired result follows directly from

$$
F(w^{(t)};\lambda)-F(\bar{w};\lambda)\le G(w^{(t)},\theta^{(t)};\lambda)-G(\bar{w};\lambda)=\epsilon_t.
$$

□

Acknowledgments

Xiao-Tong Yuan was supported partially by the National Natural Science Foundation of China under Grant 61402232, 61522308, and partially by the Natural Science Foundation of Jiangsu Province of China under Grant BK20141003. Qingshan Liu was supported partially by National Natural Science Foundation of China under Grant 61532009.

References

[Argyriou *et al.*, 2012] Andreas Argyriou, Rina Foygel, and Nathan Srebro. Sparse prediction with the k-support norm. In *Advances in Neural Information Processing Systems*, 2012.

[Beck and Teboulle, 2004] Amir Beck and Marc Teboulle. A conditional gradient method with linear rate of convergence for solving convex linear systems. *Mathematical Methods of Operations Research*, 59(2):235–247, 2004.

[Beck and Teboulle, 2009] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[Belilovsky *et al.*, 2015a] Eugene Belilovsky, Andreas Argyriou, Gaël Varoquaux, and Matthew Blaschko. Convex relaxations of penalties for sparse correlated variables with bounded total variation. *Machine Learning*, 100(2-3):533–553, 2015.

[Belilovsky *et al.*, 2015b] Eugene Belilovsky, Katerina Gkirtzou, Michail Misyrlis, Anna B Konova, Jean Honorio, Nelly Alia-Klein, Rita Z Goldstein, Dimitris Samaras, and Matthew B Blaschko. Predictive sparse modeling of fmri data for improved classification, regression, and visualization using the k-support norm. *Computerized Medical Imaging and Graphics*, 46:40–46, 2015.

[Blaschko, 2013] Matthew Blaschko. A note on k-support norm regularized risk minimization. *arXiv preprint arXiv:1303.6390*, 2013.

[Cai *et al.*, 2011] Deng Cai, Xiaofei He, Jiawei Han, and Thomas S Huang. Graph regularized nonnegative matrix factorization for data representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(8):1548–1560, 2011.

[Chatterjee *et al.*, 2014] Soumyadeep Chatterjee, Sheng Chen, and Arindam Banerjee. Generalized dantzig selector: Application to the k-support norm. In *Advances in Neural Information Processing Systems*, 2014.

[Dudik *et al.*, 2012] Miro Dudik, Zaid Harchaoui, and Jérôme Malick. Lifted coordinate descent for learning with trace-norm regularization. In *International Conference on Artificial Intelligence and Statistics*, 2012.

[Eriksson *et al.*, 2015] Anders Eriksson, Trung Thanh Pham, Tat-Jun Chin, and Ian Reid. The k-support norm and convex envelopes of cardinality and rank. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[Frank and Wolfe, 1956] Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3(1-2):95–110, 1956.

[Garber and Hazan, 2014] Dan Garber and Elad Hazan. Faster rates for the frank-wolfe method over strongly-convex sets. In *Internatioal Conference on Machine Learning*, 2014.

[Grave *et al.*, 2011] Edouard Grave, Guillaume R Obozinski, and Francis R Bach. Trace lasso: a trace norm regularization for correlated designs. In *Advances in Neural Information Processing Systems*, 2011.

[Grubb and Bagnell, 2011] Alexander Grubb and J. Andrew Bagnell. Generalized boosting algorithms for convex optimization. In *International Conference on Machine Learning*, 2011.

[Harchaoui *et al.*, 2015] Zaid Harchaoui, Anatoli Juditsky, and Arkadi Nemirovski. Conditional gradient algorithms for norm-regularized smooth convex optimization. *Mathematical Programming*, 152:75–112, 2015.

[Hull, 1994] Jonathan J. Hull. A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence,*, 16(5):550–554, 1994.

[Jaggi, 2013] Martin Jaggi. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *International Conference on Machine Learning*, 2013.

[Kim and Kim, 2004] Y. Kim and J. Kim. Gradient lasso for feature selection. In *International Conference on Machine Learning*, 2004.

[Lacoste-Julien and Jaggi, 2015] Simon Lacoste-Julien and Martin Jaggi. On the global linear convergence of Frank-Wolfe optimization variants. In *Advances in Neural Information Processing Systems*, 2015.

[Lacoste-Julien *et al.*, 2013] Simon Lacoste-Julien, Martin Jaggi, Mark Schmidt, and Patrick Pletscher. Block-coordinate Frank-Wolfe optimization for structural svms. *International Conference on Machine Learning*, 2013.

[Lai *et al.*, 2014] Hanjiang Lai, Yan Pan, Canyi Lu, Yong Tang, and Shuicheng Yan. Efficient k-support matrix pursuit. In *European Conference on Computer Vision*, 2014.

[LeCun *et al.*, 1998] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[Liu *et al.*, 2010] Bo Liu, Meng Wang, Richang Hong, Zhengjun Zha, and Xian-Sheng Hua. Joint learning of labels and distance metric. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 40(3):973–978, 2010.

[McDonald *et al.*, 2014] Andrew M McDonald, Massimiliano Pontil, and Dimitris Stamos. Spectral k-support norm regularization. In *Advances in Neural Information Processing Systems*, 2014.

[Ñanculef *et al.*, 2014] Ricardo Ñanculef, Emanuele Frandi, Claudio Sartori, and Héctor Allende. A novel Frank-Wolfe algorithm. analysis and applications to large-scale svm training. *Information Sciences*, 285:66–99, 2014.

[Parikh and Boyd, 2013] Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in optimization*, 1(3):123–231, 2013.

[Rao *et al.*, 2015] Nikhil Rao, Parikshit Shah, and Stephen Wright. Forward-backward greedy algorithms for atomic norm regularization. *IEEE Transactions on Single Processing*, 63(21):5798–5811, 2015.

[Schmidt *et al.*, 2009] Mark W Schmidt, Ewout Berg, Michael P Friedlander, and Kevin P Murphy. Optimizing costly functions with simple constraints: A limited-memory projected quasi-newton algorithm. In *International Conference on Artificial Intelligence and Statistics*, 2009.

[Shalev-Shwartz *et al.*, 2010] Shai Shalev-Shwartz, Nathan Srebro, and Tong Zhang. Trading accuracy for sparsity in optimization problems with sparsity constraints. *SIAM Journal on Optimization*, 20(6):2807–2832, 2010.

[Yuan and Li, 2014] Xiao-Tong Yuan and Ping Li. Sparse additive subspace clustering. In *European Conference on Computer Vision*. 2014.

[Yuan and Yan, 2013] Xiao-Tong Yuan and Shuicheng Yan. Forward basis selection for pursuing sparse representations over a dictionary. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):3025–3036, 2013.

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

[Zou and Hastie, 2005] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.