OPTIMIZATION IN SPARSE LEARNING: FROM CONVEXITY TO NON-CONVEXITY

by

BO LIU

A dissertation submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Computer Science

Written under the direction of

Professor Dimitris N. Metaxas

And approved by

_____

_____

_____

_____

New Brunswick, New Jersey

Janurary, 2019

ABSTRACT OF THE DISSERTATION

Optimization in Sparse Learning: from Convexity to Non-Convexity

by Bo Liu

Dissertation Director: Professor Dimitris N. Metaxas

Nowadays, the explosive data scale increase provides an unprecedented opportunity to apply machine learning methods in various application domains. The high-dimension data representation proposes curse of dimension challenge to machine learning models. Sparse model offers a tool that can alleviate this challenge by learning a low-dimension feature and model representation. Traditionally, the sparse model is learned by penalizing the ℓ_1-norm of the model parameter in optimization. Recent sparse model learning research is studying more accurate way of modeling the sparsity degree as prior knowledge, representative work includes k-support norm regularized minimization and ℓ_0-constrained minimization.

If the training loss is convex, minimizing the training loss with model parameter k-support-norm regularizer is still a convex optimization problem. In chapter 2 we introduce the proposed fully corrective Frank-Wolfe type algorithm, called k-FCFW, for k-support-norm regularized sparse model learning. We reformulate the regularized minimization into a constrained minimization task, then the the proposed algorithm is applied to solve the reformulated problem. In this work we compare the per-iteration complexity of the proposed k-FCFW algorithm with proximal gradient algorithm, which is conventionally used to solve the original problem. One theoretical contribution is

that we establish a linear convergence for the proposed algorithm under some standard assumptions.

The model parameter ℓ_0-norm can be directly used as constraint in the learning objective. However, in this condition even if training loss is convex, the problem is non-convex and NP-hard because of the ℓ_0-norm constraint. To obtain a tradeoff between model solving accuracy and efficiency, several primal domain greedy algorithms have been proposed. The algorithm properties such as model estimation error upper bound and support recovery are analyzed in literature. In chapter 3 we introduce the proposed dual space algorithm for the sparsity-constrained ℓ_2-norm regularized finite sum loss minimization problem. The sparse duality theory established in this work sets up the sufficient and necessary conditions under which the original non-convex problem can be equivalently solved in a concave dual formulation. The dual iterative hardthresholding (DIHT) algorithm and its stochastic variant are proved to be able to recover the parameter support without the Restricted Isometry Property (RIP) condition.

Distributed optimization algorithm is used for learning a global optimal model when training samples locate on different machines. Communication efficiency is one important concern in distributed model training algorithm design. Chapter 4 elaborates the proposed Newton-type inexact pursuit algorithm for the ℓ_0-constrained empirical risk minimization problem. The proposed algorithm iterates between inexactly solving a local sparse learning problem with existing single machine algorithms, and communicating gradient and model parameter between master machine and worker machines. Algorithm analysis shows the model estimation error upper bound has linear convergence rate.

In the last part, we conclude this dissertation and present the future direction of sparse model learning research.

Acknowledgements

First of all, I would like to thank my Ph.D advisor, Prof. Dimitris N. Metaxas, for his support and supervision of my Ph.D study. I learn a lot from his guidance, such as the broad vision in research problem definition, the creative thinking to figure out the research challenges and the effective communication in collaborative work. Those skills are very useful for me to pursue my Ph.D degree. I believe in my future career I will benefit from those skills. I witness his hardwork as the director of CBIM and the advisor who has graduated tens of Ph.D students. His passion on academic work inspires me to keep on improving my research.

Besides my advisor, I sincerely thank the rest of my Ph.D dissertation defense committee members, Prof. Vladimir Pavlovic, Prof. Ahmed Elgammal and Prof. Michael Katehakis, for their comments on my work. Prof. Kostas Bekris, Prof. Konstantinos Michmizos and Prof. Vinod Ganapathy are committee members of my Ph.D qualification exam. I also greatly appreciate their comments on my research and the questions they proposed. Their encouragement and feedback provide helpful guidance for me to move on my research after the qualification exam.

I am very grateful to all collaborators who work with me on the publications related to my dissertation research. They are Prof. Xiao-Tong Yuan and Prof. Qingshan Liu, Mr. Lezi Wang, Dr. Shaoting Zhang and Prof. Junzhou Huang. The generous help from the collaborators is one important reason to achieve the results.

During my Ph.D study, I also benefit a lot from the collaborative work with other professors, researchers and latmates through project, internship and other forms of joint research. They include Prof. Carol Neidle, Dr. Sam Zheng, Dr. Dongrui Wu, Dr. Jingjing Liu, Dr. Peng Yang, Dr. Fei Yang, Dr. Lin Zhong, Dr. Xiang Yu and Prof. Xi Peng and Mr. Dong Yang. The unforgettable joint work with them is of great help

to improve my research skill.

My gratitude also goes to all other faculty members, staffs of our department and the labmates of CBIM.

Last but not the least, I will never forget the continuous love and support from my parents. I deeply love them forever.

The major body of Chapter 2 was presented in the paper [62]. Chapter 3 of this dissertation is mainly from publication [61] and the under review paper [95]. Chapter 4 will be shown in the recently accepted paper [60].

Dedication

Dedicated to my families.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Powerful machine learning models and large-scale training data motivate the rapid popularization of AI method in various areas such as data science, computer vision and natural language processing. The explosive model complexity and training data scale increase propose an urgent requirement for efficient model training algorithms. Optimization algorithm design for model training, as one of the fundamental issue in machine learning research, keeps on getting extensive attention from academia and industry.

Many machine learning problems can be generalized into solving the following minimization problem:

$$\min_{w} F(w; X, Y) \tag{1.1.1}$$

where $X = \{x_i\}_{i=1}^N, x_i \in \mathbb{R}^p$ denotes the feature representation of training samples; $Y = \{y_i\}_{i=1}^N, y_i \in \mathbb{R}$ denotes the corresponding labels; w denotes the model parameters to be learned and $F(\cdot)$ is proper training loss, such as square loss for regression

$$F(w; x_i, y_i) = \frac{1}{2}(y_i - w^\top x_i)^2,$$

or logistic loss for binary classification, when $y_i \in \{-1, 1\}$, that is

$$F(w; x_i, y_i) = \log\left(1 + \exp(-y_i w^\top x_i)\right).$$

The real world data, such as text and image, usually have high dimension feature representation. If the feature dimensionality significantly outnumbers training sample, the trained model is very likely to be overfitting. However, this high dimension feature representation is usually redundant. The feature redundancy gives us the chance to

find a low dimension feature and model representation. Sparse model learning, which aims to discover such a low dimension representation, has been demonstrated as an effective way to alleviate the model overfitting challenge. Apart from alleviating the model overfitting, sparse model also has other desirable properties, such as better model interpretability, running time efficiency in testing and model storage benefits.

The popular method that learns a sparse model used to be modeling ℓ_1-norm of model parameter w as a regularizer or constraint in the objective optimization, that is,

$$\min_w F(w; X, Y) + \lambda \|w\|_1 \tag{1.1.2}$$

or

$$\min_w F(w; X, Y) \;\; \text{subject to} \;\; \|w\|_1 \leq c \tag{1.1.3}$$

where λ and c are pre-defined hyper-parameters. These convex relaxation based methods are prone to introduce parameter estimation bias [97]. More recently, several alternative methods are proposed. Those methods can more accurately model the parameter sparsity degree as prior knowledge. Taking the linear model (e.g. linear regression) parameterized by $w \in R^d$ as example, we hope that the learned sparse model has only $k \leq p$ non-zero entries. In [2], k-support norm $\|w\|_k^{sp}$ is proposed and used as a regularizer to encourage the solution of w has k non-zero entries, the objective is

$$\min_w F(w; X, Y) + \lambda (\|w\|_k^{sp})^2. \tag{1.1.4}$$

The k-support norm is shown to be a tighter convex relaxation of ℓ_0-norm compared to ℓ_1-norm. Alternatively, we can directly solve the ℓ_0-constrained minimization problem [93, 41], that is

$$\min_w F(w; X, Y) \;\; \text{subject to} \;\; \|w\|_0 \leq k. \tag{1.1.5}$$

In this dissertation, I will introduce my research on optimization algorithm design and analysis for sparse model learning problems. The sparse model learning objective includes optimizing convex model with $\|w\|_k^{sp}$ regularizer as well as the model learning with parameter cardinality constraint. In addition to the proposed single machine algorithms, I will also introduce our recent research progress in communication efficient

distributed sparse model learning. The designed algorithm targeted for each specific problem significantly improves the model training efficiency compared to baseline algorithms.

1.2 Thesis Organization

The following of this dissertation is organized as follows:

In chapter 2, I will introduce the proposed Frank-Wolfe type algorithm for the k-support-norm regularized model learning task. The k-support-norm is proposed in [2]. It is used as a tighter convex relaxation of model parameter ℓ_0-norm than ℓ_1-norm in regularized sparse model learning problems [49]. The traditional method that solves the k-support-norm regularized minimization problem

$$\min_w F(w, X, Y) + \lambda(\|w\|_k^{sp})^2$$

is based on the proximal gradient algorithm, which has large computational cost, especially for high dimension model learning task. We propose a Frank-Wolfe type algorithm which has shown to have cheaper per-iteration computational cost and linear convergence rate. Both of the algorithm analysis and experiment results verify the superior efficiency of the proposed algorithm.

Chapter 3 is about our research on developing dual method for ℓ_0-constrained minimization problem. With the model sparsity degree as prior knowledge, we consider minimizing the ℓ_0-constrained empirical risk plus the ℓ_2-norm model parameter regularizer, which is defined as:

$$\min_w \frac{1}{N}\sum_{i=1}^{N} f(w; x_i, y_i) + \frac{\lambda}{2}\|w\|^2 \quad \text{subject to} \quad \|w\|_0 \leq k \tag{1.2.1}$$

where $f(\cdot)$ is convex training loss, $\{x_i, y_i\}_{i=1}^N, x_i \in \mathbb{R}^p, y_i \in \mathbb{R}$ are training samples. Iterative Hard Thresholding (IHT) is a popular class of first-order greedy selection methods among existing methods [93, 41]. Different from existing IHT-type algorithms that solves the problem in primal domain, our research explores solving (1.2.1) in dual space. We first establish a duality theory for ℓ_2-regularized sparse finite-sum minimization. Based on the theory, Dual Iterative Hard Thresholding (DIHT) algorithm and

its stochastic variant are proposed. Numerical experiments verify the sparse duality theory and higher efficiency of the proposed dual algorithms in sparse SVM-type model learning tasks.

In chapter 4, I will introduce the proposed distributed optimization algorithm for the ℓ_0-constrained empirical risk minimization problem. In this problem setting, the training samples are distributed on multiple machines. Although each machine can learn a local optimal model parameter based on its local training samples, the task is to learn a global model parameter based on all training samples. The proposed method alternates between local inexact optimization of a Newton-type approximation and centralized global results aggregation. Theoretical analysis shows that for a general class of convex functions with Lipschitze continuous Hessian, the method converges linearly with contraction factor scaling inversely with data size; whilst the communication complexity required to reach desirable statistical accuracy scales logarithmically with the number of machines for some popular statistical learning models. For non-convex objective functions, our method can still be shown to converge globally. Numerical results on convex and non-convex model training tasks confirm the high efficiency of our method.

In chapter 5, I conclude this dissertation and discuss some future work of sparse learning research.

Chapter 2

k-Support Norm Regularized Minimization via Frank-Wolfe Method

2.1 Introduction

In many machine learning problems, it is common that the number of collected samples is substantially smaller than the dimensionality of the feature, implying that consistent estimators cannot be used unless additional assumptions are imposed on the model. One of the widely acknowledged prior assumptions is that the data exhibit low-dimensional structure, which can often be captured by imposing sparsity constraint on the model parameter space. Sparsity is typically obtained by regularizing the goodness of fit with sparsity inducing regularizer. However, as this typically leads to nonconvex optimization problems with high computational demands, the standard approach is to replace and relax these regularizers with convex surrogates for cardinality.

The most widely used regularizer is the ℓ_1-norm which is justified as the convex envelope of the cardinality ℓ_0-norm (i.e., number of non zero elements in a vector) on the ℓ_∞-norm unit ball [2]. 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 [101], pairwise elastic net [63], trace Lasso [29] and OSCAR [12] 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 [2]. The intuition is that in certain instances it might be reasonable to expect that not only the number of non-zero variable entry is bounded, but also the variable has bound on its Euclidean norm as well. Formally,

the k-support-norm of a vector $w \in \mathbb{R}^p$ is defined as

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

where $\mathcal{G}_k$ denotes the set of all subsets of $\{1, 2, ..., p\}$ of cardinality at most k. It is shown in [2] that the value of $\|w\|_k^{sp}$ can be computed through

$$\|w\|_k^{sp} = \left(\sum_{i=1}^{k-j-1} (|w|_i^{\downarrow})^2 + \frac{1}{r+1} (\sum_{i=k-j}^{p} |w|_i^{\downarrow})^2 \right)^{\frac{1}{2}},$$

where $|w|_i^{\downarrow}$ denotes the i-th largest element in $|w|$ and $|w|_0^{\downarrow}$ is assumed to be $+\infty$. $j \in \{0, 1, ..., k-1\}$ satisfies $|w|_{k-j-1}^{\downarrow} > \frac{1}{j+1} \left(\sum_{i=k-j}^{p} |w|_i^{\downarrow} \right) \geq |w|_{k-j}^{\downarrow}$. More properties of k-support-norm are analyzed in [2].

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 is considered in [2] for sparse prediction tasks:

$$\min_w F(w) + \lambda (\|w\|_k^{sp})^2, \tag{2.1.1}$$

where $F(w)$ is a convex and differentiable objective function parameterized by w. 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. As an extreme case with $k = 1$, the problem (2.1.1) reduces to the squared ℓ_1-norm regularized minimization problem known as exclusive Lasso [100]. In another extreme case with $k = p$, the problem (2.1.1) reduces to ridge regression.

2.1.1 Challenge and Motivation

One challenge that hinders the applicability of the k-support-norm regularized model (2.1.1) is its high computational complexity in large scale settings. Indeed, proximal gradient methods are conventionally used for optimizing the composite minimization problem in (2.1.1) [2, 49, 24]. Given the gradient vector, the per-iteration computational cost of proximal gradient methods is dominated by an proximity operator of the

following form:

$$w^* = \arg\min_w \frac{1}{2}\|w - v\|_2^2 + \lambda(\|w\|_k^{sp})^2. \tag{2.1.2}$$

In the pioneering work of [2], an exhaustive search strategy with $O(p(k + \log p))$ complexity is proposed to compute the above proximal operator, which is computationally expensive. Despite significant speed-ups have been reported in [49, 24], those methods are still exhaustive in nature and could be expensive for huge scale problems with relatively large sparsity parameter k.

It has been known that the k-support-norm ball $\mathcal{B}_r^{k,sp} := \{w \in \mathbb{R}^p : \|w\|_k^{sp} \leq r\}$ is equivalent to the convex hull of the following set of cardinality and ℓ_2-norm constraint:

$$\mathcal{C}_{k,r}^{(2)} = \{w \in \mathbb{R}^p : \|w\|_0 \leq k, \|w\|_2 \leq r\}.$$

That is,

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

In this sense, k-support-norm ball $\mathcal{B}_r^{k,sp}$ provides a convex envelope of the nonconvex set $\mathcal{C}_{k,r}^{(2)}$, which is referred to as the *coreset* of $\mathcal{B}_r^{k,sp}$.

2.1.2 Contributions

In this chapter we will show that, compared to the existing methods that solves the regularized formulation (2.1.1) though proximal gradient based methods, the proposed fully corrective Frank-Wolfe based method is more convenient and efficient in sense of optimization. One important reason is that the constrained formulation the convex hull structure of the k-support-norm ball suggests extremely simple gradient projection operation. We name the proposed algorithm as k-FCFW in the following context. The original Frank-Wolfe method is designed to solve the constrained minimization problem. In our method, by introducing an augmented variable as the bound of k-support regularization $\|w\|_k^{sp}$ to (2.1.1), the regularized objective is converted into a constrained minimization problem. The two model parameters are iteratively optimized by a fully corrective variant of the Frank-Wolfe algorithm. Theoretical analysis is conducted to establish the converge rate and model estimation error of the proposed

algorithm. Particularly, we prove that the proposed algorithm converges linearly under proper conditions, which is stronger than the sublinear convergence rate proved in [32] for norm regularized minimization problem. We further discuss the extension of the proposed algorithm in low-rank introducing problem and show that it has obvious computational cost advantage. Experimental results on various forms of loss function demonstrate the superior efficiency of the proposed algorithm.

2.1.3 Organization

The remaining of this chapter is organized as follows: In §2.2 we briefly review some properties and related work of k-support-norm, k-support-norm regularized proximal operator and the forward greedy selection algorithms for sparsity constrained optimization. In §2.3 we present our k-support-norm constrained minimization problem and two forward greedy selection algorithms for optimization. In §2.4 we access the performance of the proposed model and algorithms on a number of benchmark datasets. Finally, in §2.5 we conclude this paper and prospect the future work.

2.2 Related Work

In this section we briefly review some previous literatures on k-support-norm and the Frank-Wolfe method for optimization.

2.2.1 k-Support Norm Regularized Minimization

Using k-support-norm enables us with the flexibility of controlling the cardinality of solution by setting the value of k. In [67], box norm is proposed based as an extension of k-support-norm. Multiple k-support-norm regularized convex models are investigated in [9]. The applications of k-support-norm in various computer vision problems have been explored in [49]. In [17], the k-support norm is applied in generalized dantzig selector for linear model. The authors of [8] show that it is helpful to use k-support-norm regularization in fMRI data analysis including classification, regression and data visualization task. In [7], total variation penalty is incorporated in the k-support framework

and applied in image and neuroscience data analysis.

Most of the k-support norm related applications require to solve the regularized minimization task (2.1.1). Proximal gradient algorithm [71] is the acknowledged state of the art algorithm designed for solving (2.1.1). The proximal operator (2.1.2) has been proved to have the following closed form solution [2]:

$$w^*_{s_i} = \text{sign}(v_{s_i})q_i, \ \ i = 1, ..., p,$$

where v_{s_i} is the i-th largest (in magnitude) entry of v and

$$q_i = \begin{cases} \frac{\beta}{1+\beta} v_{s_i} & i = 1, 2, ..., k-r-1 \\ v_{s_i} - \frac{\sum_{i=k-r}^{l} v_{s_i}}{l-k+(\beta+1)r+\beta+1} & i = k-r, ..., l \\ 0 & i = l+1, ..., p \end{cases},$$

in which $\beta = 2\lambda$, r and l are integers satisfying

$$\begin{cases} \frac{1}{\beta+1} v_{s_{k-r-1}} > \frac{\sum_{i=k-r}^{l} v_{s_i}}{l-k+(\beta+1)r+\beta+1} \geq \frac{1}{\beta+1} v_{s_{k-r}} \\ v_{s_l} > \frac{\sum_{i=k-r}^{l} v_{s_i}}{l-k+(\beta+1)r+L+1} \geq v_{s_{l+1}} \end{cases}.$$

The major computational cost of the above proximal operator lies in how to find integers r and l. In [2], an exhaustive search method with $O(p(k + \log p))$ complexity was used to find r and l. In [49], a binary search strategy with $O((p + k) \log p)$ complexity was proposed to find l, which has been shown to be much more efficient and scalable than exhaustive search in large scale problems. In [24], the authors explored the dual form of k-support-norm and proposed to use a more efficient iterative binary searching method to find the optimal integers r and l.

2.2.2 Frank-Wolfe Method

The history of Frank-Wolfe method dates back to [26] for polytope constrained optimization. It is also known as conditional gradient algorithm. Consider a compact convex set S. The Frank-Wolfe method applies to solve the following constrained optimization problem:

$$w^* = \arg\min_{w} f(w) \quad \text{subject to } w \in S, \tag{2.2.1}$$

where f is assumed to be a real valued convex function.

Due to the potential of efficiency improvement when applied in solving minimization problem with certain form of constraint, recently there is an increasing trend to revisit and restudy this method [36, 27, 28]. Frank-Wolfe based methods have been developed to solve some optimization problems such as structural SVM [48], trace norm regularization problem [22] and atomic norm constrained problem [73]. In [32], the conditional gradient methods are used to solve regularized convex optimization problems with specific forms of regularization including nuclear norm and total variation. However, this method replies on an estimation of the bound of regularization, which is difficult to obtain.

Many sparse model learning algorithm can be viewed as variants of Frank-Wolfe method. Specifically, when the constraint set in (2.2.1) is a convex hull in S, i.e., $w \in \mathrm{co}(S)$, where

$$\mathrm{co}(S) = \left\{ \sum_{i=1}^{|S|} \alpha_i u_i : \alpha_i \geq 0, \sum_{i=1}^{|S|} \alpha_i = 1, u_i \in S \right\},$$

the sequential greedy approximation (SGA) method was proposed to find a sparse approximate solution of (2.2.1). Recently, several variants of SGA, e.g., forward greedy selection [80] and gradient Lasso [44] have been proposed in sparse learning. It has also received significant interests in semi-definite program [33] and low-rank matrix completion/approximation [38, 79, 96]. In the context of boosting classification, the restricted gradient projection algorithms stated in [30] is essentially a forward greedy selection method over ℓ_2-functional space.

2.3 The Fully Corrective Frank-Wolfe Method for k-Support Norm Regularized Minimization

We consider using the class of Frank-Wolfe algorithm to solve the regularized minimization problem (2.1.1). To this end, we first reformulate (2.1.1) into a constrained optimization problem:

$$\min_{w,\theta} G(w, \theta; \lambda) := f(w) + \lambda\theta \quad \text{s.t.} \quad \left(\|w\|_k^{sp}\right)^2 \leq \theta, \tag{2.3.1}$$

Here θ is an augmented variable bounding the term $\left(\|w\|_k^{sp}\right)^2$. We next introduce the detail of k-FCFW algorithm. After that algorithm convergence and parameter estimation error analysis is provided.

2.3.1 Algorithm Description

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

Initialization: *set* $w^{(0)}$ *to be a* k*-sparse vector,* $\theta^{(0)} = \|w^{(0)}\|^2$, $U = w^{(0)}$, $V = \theta^{(0)}$.

for $t = 1, 2, \ldots$ **do**

(S1) Compute $\nabla f(w^{(t-1)})$.

(S2) Solve the constrained linear problem

$$\{u^{(t)}, v^{(t)}\} = \arg\min_{u,v} \langle \nabla f(w^{(t-1)}), u\rangle + \lambda v \qquad \text{subject to} \qquad \left(\|u\|_k^{sp}\right)^2 \leq v. \tag{2.3.2}$$

(S3) Update $U^{(t)} = [U^{(t-1)}, u^{(t)}], V^{(t)} = [V^{(t-1)}, v^{(t)}]$.

Compute

$$\alpha^{(t)} = \min_{\alpha \in \triangle_t} f(U^{(t)}\alpha) + \lambda V^{(t)}\alpha, \tag{2.3.3}$$

where $\triangle_t = \{\alpha \in \mathbb{R}^{t+1} : \alpha \geq 0, \|\alpha\|_1 = 1\}$;

(S4) Update

$$w^{(t)} = U^{(t)}\alpha^{(t)}, \qquad \theta^{(t)} = V^{(t)}\alpha^{(t)}.$$

end

Output: $w^{(t)}$.

The k-FCFW algorithm for k-support norm regularized minimization is outlined in Algorithm 1. Since the k-support-norm ball $\|w\|_k^{sp} \leq v$ is a convex hull of the set $\mathcal{C}_{k,v}^{(2)} = \{w \in \mathbb{R}^p : \|w\|_0 \leq k, \|w\|_2 \leq v\}$, it is trivial to derive that, given the optimal $v > 0$, the optimal u of (2.3.2) admits the following close-form solution:

$$u = -\frac{\sqrt{v}\nabla_k f(w^{(t-1)})}{\|\nabla_k f(w^{(t-1)})\|}, \tag{2.3.4}$$

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 (2.3.2) we get that $v^{(t)}$ solves the

following quadratic program:

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

Obviously,

$$v^{(t)} = \left(\frac{\|\nabla_k f(w^{(t-1)})\|}{2\lambda}\right)^2, \tag{2.3.5}$$

and thus,

$$u^{(t)} = -\frac{\sqrt{v^{(t)}}\nabla_k f(w^{(t-1)})}{\|\nabla_k f(w^{(t-1)})\|} = -\frac{\nabla_k f(w^{(t-1)})}{2\lambda}. \tag{2.3.6}$$

At each time instance t, we keep a memory of all previous updates, that is

$$U^{(t)} = \{w^{(0)}, u^{(1)}, ..., u^{(t)}\},$$
$$V^{(t)} = \{\theta^{(0)}, v^{(1)}, v^{(2)}, ..., v^{(t)}\}.$$

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 (2.3.3) 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 some off-the-shelf algorithms such as the projected quasi-Newton (PQN) method [77].

It is noteworthy that the subproblem (2.3.2) is equivalent to the following k-support-norm regularized linear program:

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

This is different from the proximal gradient method which involves solving the quadratic proximity operator (2.1.2) at each iteration. Apparently solving (2.3.2) is more efficient than solving (2.1.2). When t is of moderate value and warm start is adopted to initialize the parameters, the subproblem (2.3.3) can be efficiently solved with few iterations.

2.3.2 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 (Restricted strong smoothness and strong convexity)**.** *We say f is* L-smooth *if there exists a positive constant L such that for any w and w',*

$$f(w') - f(w) - \langle \nabla f(w), w' - w \rangle \leq \frac{L}{2}\|w - w'\|^2. \tag{2.3.7}$$

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

$$f(w') - f(w) - \langle \nabla f(w), w' - w \rangle \geq \frac{\rho_s}{2}\|w - w'\|^2. \tag{2.3.8}$$

The conditions of restricted strong convexity/smoothness are key to the analysis of several previous greedy selection methods [80, 96, 3]. In the following, we define

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

and

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

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

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

Proof. A proof of this result is given in §2.6.1. □

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 $\mathcal{M}$.

We show in the following theorem that a stronger geometric rate of convergence can be established 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*-strongly smooth and* $\rho_{s+\bar{s}}$*-strongly convex. Given* $\epsilon > 0$, *let us run* t *iterations of Algorithm 1 with*

$$t \geq \frac{1}{\zeta} \ln \left[\frac{F(w^{(0)}; \lambda) - F(\bar{w}; \lambda)}{\epsilon} \right]$$

where

$$\zeta := \min \left\{ \frac{\rho_{s+\bar{s}} \bar{\beta}}{4\bar{l} L r^2}, \frac{1}{2} \right\}.$$

then Algorithm 1 will output $w^{(t)}$ *satisfying*

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

Proof. A proof of this result is given in §2.6.2. □

Remark 1. *There have been several literatures that dedicate in exploring the convergence rate of FW methods in constrained minimization problem. In general case, the Frank-Wolfe method is known to have* $\mathcal{O}(\frac{1}{t})$ *convergence rate [36]. An* $\mathcal{O}(\frac{1}{t^2})$ *convergence rate is proved in [28] for applying Frank-Wolfe method in constrained minimization over strongly-convex sets. Several linear convergence guarantees are established [56, 5, 45, 50, 46, 69] by adding various specific assumptions to either constraint set or loss function, which are not directly applicable to our problem. In recent work of [47], a global linear convergence rate is proved for a number of Frank-Wolfe algorithm variants given the polytope constraint set. This analysis doesn't perfectly fit for our algorithm that applies FW method in solving the regularized optimization objective (2.1.1). In each iteration of our algorithm, we adaptively update the value of* v. *The constraint* $(\|w\|_k^{sp})^2 \leq v$ *is a* k*-support norm cone, rather than a polytope as a result. This imposes extra challenges in analysis.*

Remark 2. *In Algorithm 1 we have required the subproblem (2.3.3) in Step (S3) 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 (2.3.3) 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 $F(w^{(t)}; \lambda) \leq F(\bar{w}; \lambda) + \epsilon + \mathcal{O}(\varepsilon)$ *after* $t = \mathcal{O}\left(\ln(\frac{1}{\epsilon})\right)$ *steps of iteration. In other words, the optimization error of the subproblem (2.3.3) does not accumulate during the iteration.*

2.3.3 Parameter Estimation Error Analysis

The parameter estimation error can be analyzed based on the convergence results established in the previous subsection.

Lemma 2. *Let* w *be an* s*-sparse vector. Assume that* f *is* $\rho_{s+\bar{s}}$*-strongly convex. It holds that*

$$\|w - \bar{w}\| \leq \sqrt{\frac{2(F(w;\lambda) - F(\bar{w};\lambda))}{\rho_{s+\bar{s}}}}.$$

Proof. A proof of this Lemma is provided in §2.6.3. □

Based on Theorem 1 and Lemma 2, we directly obtain the following corollary on the estimation error of k-FCFW.

Corollary 1. *Given* $\epsilon > 0$ *and the conditions in Theorem 1 are satisfied, after* $t = \mathcal{O}\left(\ln(\frac{1}{\epsilon})\right)$ *running, Algorithm 1 will output* $w^{(t)}$ *satisfying* $\|w^{(t)} - \bar{w}\| = \mathcal{O}(\sqrt{\epsilon})$.

2.4 Experiments

We conduct experiments to verify the high efficiency of k-FCFW by testing its empirical performance given numerous forms of k-support norm regularized optimization task with different forms of loss functions including logistic loss and matrix pursuit. All the considered algorithms are implemented in Matlab and tested on a computer equipped with 3.0GHz CPU and 32GB RAM.

2.4.1 k-Support-Norm Regularized ℓ_2-Logistic Regression

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

$$\min_w F(w) = \frac{1}{N}\sum_{i=1}^{N} \log\left(1 + \exp(-y_i w^\top x_i)\right) + \frac{\tau}{2}\|w\|^2 + \lambda(\|w\|_k^{sp})^2, \tag{2.4.1}$$

The parameter τ controls the strong convexity of the loss function.

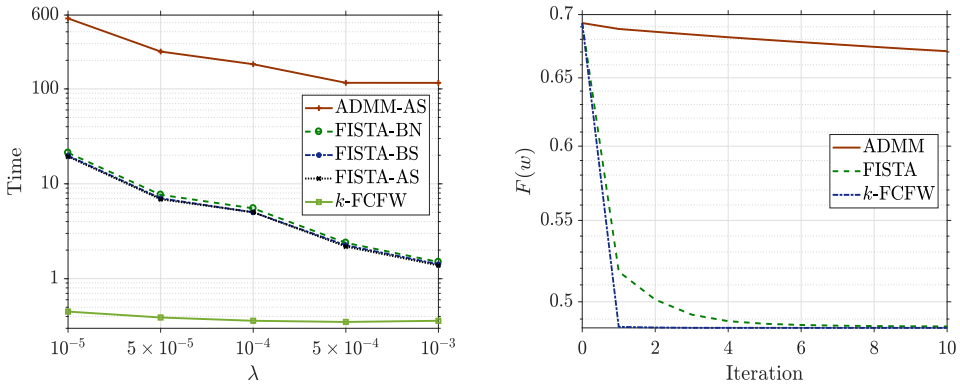
We test the algorithm efficiency on a synthetic dataset. The model parameter ground truth $\bar{w}$ is designed to be a p-dimension vector as follows:

$$\bar{w} = [\underbrace{10, 10, \cdots, 10}_{p'}, \underbrace{0, 0, \cdots, 0}_{p-p'}].$$

Each training sample is designed to have two components. the first p'-dimension is drawn from Gaussian distribution $\mathcal{N}(0, \Sigma)$, $\Sigma_{i,j} = \begin{cases} 1 & \text{if } i == j \\ 0.5^{\frac{|i-j|}{2}} & \text{if } i \neq j \end{cases}$. Other $p - p'$ dimensions are drawn from Gaussian $\mathcal{N}(0, 1)$ as noise. The label y_i follows Bernoulli distribution with probability $\mathbb{P}(y_i = 1|x_i) = \frac{\exp(\bar{w}^\top x_i)}{1+\exp(\bar{w}^\top x_i)}$. The task is designed as selecting the top $k = p'$ most discriminative features for classification using logistic regression model through solving (2.4.1).

We produce the training data by setting $N = 500$, $p = 10^5$, $p' = 5 \times 10^3$, respectively. 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 [67]; (2) the binary search based solver (denoted by BS) proposed in [49]; and (3) the solver proposed in [24] 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 [6]. We also compare the efficiency of k-FCFW with ADMM [71] which is another popular framework for regularized minimization problems.

The running time of the considered algorithms is shown in Figure 2.4.1(a). The value of λ is varied to be $\{10^{-5}, 5 \times 10^{-5}, 10^{-4}, 5 \times 10^{-4}, 10^{-3}\}$. We first run FISTA algorithm to reach the convergence state determined by $\frac{|F(w^{(t)}) - F(w^{(t-1)})|}{F(w^{(t)})} \leq 10^{-4}$, then we run other algorithms to the same training loss or maximum number of iteration is



(a) Time Cost, $k = 5K$

(b) Convergence, $\lambda = 10^{-3}$

Figure 2.4.1: Results on synthetic dataset: (a) Running time (in second) curves of the considered comparing methods under different values of λ. (b) Convergence curves of the considered methods under $k = 5K, \lambda = 10^{-4}$.

reached. It can be observed that our method is significantly faster than all the three comparing solvers.

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 curve of ADMM-AS is drawn in Figure 2.4.1(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 2.4.1(b), we plot the convergence curves of the considered algorithms under $\lambda = 10^{-4}$. It can be observed that our method needs significant less number of iterations to reach comparable optimization accuracy.

2.4.2 k-Support-Norm Matrix Pursuit

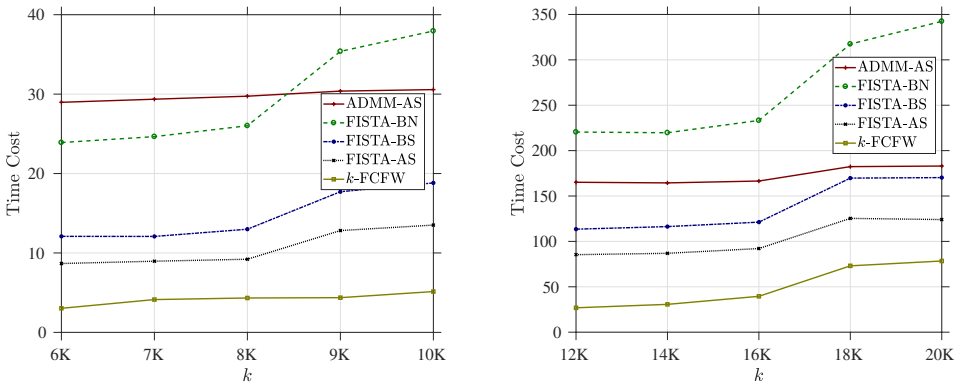
In this group of experiments, we apply the proposed method to the k-support-norm regularized matrix pursuit problem. Matrix pursuit has extensive applications such as subspace segmentation, semi-supervised learning and sparse coding. The results of [49] indicate that the k-support-norm regularized matrix pursuit method achieves superior performance in various applications. The k-support-norm regularized matrix pursuit is

formulated as:

$$\min_{W\in\mathbb{R}^{n\times n}} \frac{1}{2}\|X - XW\|_F^2 + \lambda(\|vec(W)\|_k^{sp})^2, \tag{2.4.2}$$

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

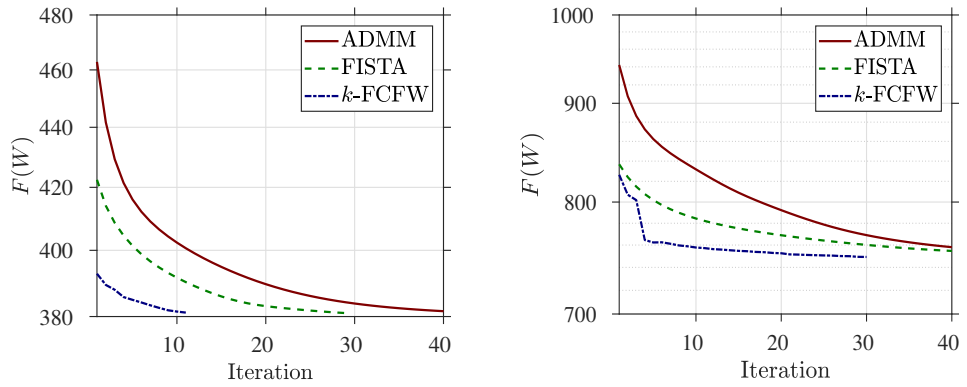
The MNIST [53] and USPS [35] datasets are adopted for testing. For MNIST dataset, we resize each image into 14×14 then normalize the gray value into $[0, 1]$. The pixel values are then vectorized as image feature. The USPS dataset is preprocessed by [14]. Each image is represented by a 256-dimension feature vector and the feature values are normalized into $[-1, 1]$. Each image of both datasets are 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 varying k values in the k-support-norm. The regularization parameter is set to be $\lambda = 10$.



(a) MNIST (b) USPS

Figure 2.4.2: Time cost comparison between k-FCFW and baseline algorithms on MNIST and USPS datasets.

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 2.4.2. The time cost curves of ADMM-AS are also shown in Figure 2.4.2. The convergence curves of the considered methods evaluated



(a) MNIST, $k = 8K$. (b) USPS, $k = 16K$.

Figure 2.4.3: The convergence curves of considered methods on MNIST-1000 and USPS-2000 datasets. The starting point of each curve is $F(W^{(1)})$.

by $F(W)$ when $k = 8K$ for MNIST dataset and $k = 16K$ for USPS dataset are shown in Figure 2.4.3. From these results we can see that in all cases, k-FCFW is the most efficient one for optimization.

2.5 Conclusion

In this chapter, 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.

2.6 Appendix

2.6.1 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 $\sum_i \bar{\alpha}_i = 1$. □

2.6.2 Proof of Theorem 1

Proof. From the definition of $G(w, \theta; \lambda)$ and the step (S4) of Algorithm 1 we have that

$$\begin{aligned}
G(w^{(t)}, \theta^{(t)}; \lambda) =& f(U^{(t)}\alpha^{(t)}) + \lambda V^{(t)}\alpha^{(t)} \\
\le& f((1-\eta)U^{(t-1)}\alpha^{(t-1)} + \eta u^{(t)}) + \lambda((1-\eta)V^{(t-1)}\alpha^{(t-1)} + \eta v^{(t)}) \\
=& f((1-\eta)w^{(t-1)} + \eta u^{(t)}) + \lambda((1-\eta)\theta^{(t-1)} + \eta v^{(t)}) \\
\le& f(w^{(t-1)}) + \eta\langle \nabla f(w^{(t-1)}), u^{(t)} - w^{(t-1)}\rangle + 2\eta^2 r^2 L + \lambda((1-\eta)\theta^{(t-1)} + \eta v^{(t)}) \\
=& f(w^{(t-1)}) + \lambda\theta^{(t-1)} + \eta\left(\langle \nabla f(w^{(t-1)}), u^{(t)} - w^{(t-1)}\rangle + \lambda(v^{(t)} - \theta^{(t-1)})\right) + 2\eta^2 r^2 L \\
=& G(w^{(t-1)}, \theta^{(t-1)}; \lambda) + \eta\left(\langle \nabla f(w^{(t-1)}), u^{(t)} - w^{(t-1)}\rangle + \lambda(v^{(t)} - \theta^{(t-1)})\right) + 2\eta^2 r^2 L.
\end{aligned}$$

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 equivalent written as

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

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 $x \in \mathcal{X}^c$:

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

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

$$\xi G(x^{(t)};\lambda) \le \xi G(x^{(t-1)};\lambda) + \eta\left(\sum_{x\in\mathcal{X}^c} \bar{\alpha}_x\langle\nabla G(x^{(t-1)};\lambda), x\rangle - \xi\langle\nabla G(x^{(t-1)};\lambda), x^{(t-1)}\rangle\right) + 2\eta^2 r^2 \xi L. \tag{2.6.1}$$

Since $\sum_{x\in\mathcal{X}^{(t-1)}} \bar{\alpha}_x/(1-\xi) = 1$, from the optimality of $\alpha^{(t-1)}$ (see the step (S3)) we can derive that

$$\langle\nabla G(x^{(t-1)};\lambda), \sum_{x\in\mathcal{X}^{(t-1)}} \bar{\alpha}_x x/(1-\xi) - x^{(t-1)}\rangle \ge 0. \tag{2.6.2}$$

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 (2.6.2). Combining the preceding inequality with (2.3.8) we obtain that

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

where the last inequality follows $\sum_{u\in\bar{\mathcal{U}}\setminus\mathcal{U}^{(t-1)}}\bar{\alpha}_u^2 \geq (\sum_{u\in\bar{\mathcal{U}}\setminus\mathcal{U}^{(t-1)}}\bar{\alpha}_u)^2/\bar{l}$. By combining the above with (2.6.1) we get

$$\xi G(x^{(t)};\lambda) \leq \xi G(x^{(t-1)};\lambda) - \eta\left(G(x^{(t-1)};\lambda) - G(\bar{x};\lambda) + \frac{\rho_{s+\bar{s}}\bar{\beta}\xi^2}{2\bar{l}}\right) + 2\eta^2 r^2 \xi L.$$

Let us choose $\eta = \xi\zeta \leq 1$ 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

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

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

2.6.3 Proof of Lemma 2

Proof. Let $r(w) := (\|w\|_k^{sp})^2$. From the definition of $F(w;\lambda)$ and the strong convexity of f we know that

$$\begin{aligned}
F(w;\lambda) =& f(w) + \lambda r(w) \\
\geq& f(\bar{w}) + \langle \nabla f(\bar{w}), w - \bar{w}\rangle + \frac{\rho_{s+\bar{s}}}{2}\|w-\bar{w}\|^2 + \lambda(r(\bar{w}) + \langle \partial r(\bar{w}), w - \bar{w}\rangle) \\
=& f(\bar{w}) + \lambda r(\bar{w}) + \langle \nabla f(\bar{w}) + \lambda\partial r(\bar{w}), w - \bar{w}\rangle + \frac{\rho_{s+\bar{s}}}{2}\|w-\bar{w}\|^2 \\
=& F(\bar{w};\lambda) + \langle \nabla f(\bar{w}) + \lambda\partial r(\bar{w}), w - \bar{w}\rangle + \frac{\rho_{s+\bar{s}}}{2}\|w-\bar{w}\|^2.
\end{aligned}$$

Since $\bar{w}$ is optimal, it satisfies

$$\nabla f(\bar{w}) + \lambda\partial r(\bar{w}) = 0.$$

Therefore

$$F(w) \geq F(\bar{w}) + \frac{\rho_{s+\bar{s}}}{2}\|w-\bar{w}\|^2,$$

which implies the desired result. □

Chapter 3

Dual Iterative Hard Thresholding Method for ℓ_0-Constrained Minimization

3.1 Introduction

The k-support-norm as a regularizer in sparse learning is a convex relaxation of ℓ_0-norm. In this chapter we consider solving the sparse model learning problem that involves the parameter ℓ_0-norm as a constraint. Given a set of training samples $\{(x_i, y_i)\}_{i=1}^N$ in which $x_i \in \mathbb{R}^p$ is the feature representation and $y_i \in \mathbb{R}$ the corresponding label, the following sparsity-constrained ℓ_2-norm regularized loss minimization problem is often considered in high-dimensional analysis:

$$\min_{\|w\|_0 \le k} P(w) := \frac{1}{N} \sum_{i=1}^{N} f(w^\top x_i, y_i) + \frac{\lambda}{2}\|w\|^2. \tag{3.1.1}$$

Here $f(\cdot;\cdot)$ is a convex loss function, $w \in \mathbb{R}^p$ is the model parameter vector and λ controls the regularization strength. For example, the squared loss

$$f(w^\top x_i, y_i) = (y_i - w^\top x_i)^2$$

is used in linear regression and the hinge loss

$$f(w^\top x_i, y_i) = \max\{0, 1 - y_i w^\top x_i\}$$

in support vector machines. Due to the presence of cardinality constraint $\|w\|_0 \le k$, the problem (3.1.1) is simultaneously non-convex and NP-hard in general, and thus is challenging for optimization.

We are interested in algorithms that directly minimize the non-convex formulation in (3.1.1). Early efforts mainly lie in compressed sensing for signal recovery, which is a special case of (3.1.1) with squared loss. Among others, a family of the so called

Iterative Hard Thresholding (IHT) methods [11, 25] have gained significant interests and they have been witnessed to offer the fastest and most scalable solutions in many cases. More recently, IHT-style methods have been generalized to handle generic convex loss functions [4, 93, 41] as well as structured sparsity constraints [40]. The common theme of these methods is to iterate between gradient descent and hard thresholding to maintain sparsity of solution while minimizing the objective value.

Although IHT-style methods have been extensively studied, the state-of-the-art is only designed for the primal formulation (3.1.1). It remains an open problem to investigate the feasibility of solving the original NP-hard/non-convex formulation in a dual space that might potentially further improve computational efficiency. To fill this gap, inspired by the recent success of dual methods in regularized learning problems, we systematically build a sparse duality theory and propose an IHT-style algorithm along with its stochastic variant for dual optimization.

3.1.1 Overview of Our Contribution

The core contribution of this work is two-fold in theory and algorithm. As the theoretical contribution, we have established a novel sparse Lagrangian duality theory for the NP-hard/non-convex problem (3.1.1) which to the best of our knowledge has not been reported elsewhere in literature. We provide in this part a set of *sufficient and necessary* conditions under which one can safely solve the original non-convex problem through maximizing its concave dual objective function. As the algorithmic contribution, we propose the dual IHT (DIHT) algorithm as a super-gradient method to maximize the non-smooth dual objective. In high level description, DIHT iterates between dual gradient ascent and primal hard thresholding pursuit until convergence. A stochastic variant of DIHT is proposed to handle large-scale learning problems. For both algorithms, we provide non-asymptotic convergence analysis on parameter estimation error, sparsity recovery, and primal-dual gap as well. In sharp contrast to the existing analysis for primal IHT-style algorithms, our analysis is not relying on Restricted Isometry Property (RIP) conditions and thus is less restrictive in real-life

high-dimensional statistical settings. Numerical results on synthetic datasets and machine learning benchmark datasets demonstrate that dual IHT significantly outperforms the state-of-the-art primal IHT algorithms in accuracy and efficiency. The theoretical and algorithmic contributions of this paper are highlighted in below:

- Sparse Lagrangian duality theory: we established a sparse saddle point theorem (Theorem 2), a sparse mini-max theorem (Theorem 3) and a sparse strong duality theorem (Theorem 4).
- Dual optimization: we proposed an IHT-style algorithm along with its stochastic extension for non-smooth dual maximization. These algorithms have been shown to converge at sub-linear rates when the individual loss functions are Lipschitz smooth, and at linear rates if further assuming strongly convexity of loss functions.

3.1.2 Notation and Organization

Notation. Before continuing, we define some notations to be used. Let $x \in \mathbb{R}^p$ be a vector and F be an index set. We use $\mathrm{H}_F(x)$ to denote the truncation operator that restricts x to the set F. $\mathrm{H}_k(x)$ is a truncation operator which preserves the top k (in magnitude) entries of x and sets the remaining to be zero. The notation $\mathrm{supp}(x)$ represents the index set of nonzero entries of x. We conventionally define $\|x\|_\infty = \max_i |[x]_i|$ and define $x_{\min} = \min_{i \in \mathrm{supp}(x)} |[x]_i|$. For a matrix A, $\sigma_{\max}(A)$ ($\sigma_{\min}(A)$) denotes its largest (smallest) singular value.

Organization. The rest of this chapter is organized as follows: In §3.2 we briefly review some relevant work. In §3.3 we develop a Lagrangian duality theory for sparsity-constrained minimization problems. The dual IHT-style algorithms along with convergence analysis are presented in §**??**. The numerical evaluation results are reported in §3.4. Finally, the concluding remarks are made in §3.5. All the technical proofs are deferred to §3.6.

3.2 Related Work

3.2.1 ℓ_0-Constrained Sparse Learning

In signal processing, the compressed sensing problem has been extensively studied [21, 15]. It can be viewed as a special case of ℓ_0-constrained sparse learning task with squared training loss. The representative algorithms include matching pursuit [65], orthogonal matching pursuit [89] and subspace pursuit [19].

For generic convex objective beyond quadratic loss, the rate of convergence and parameter estimation error of IHT-style methods were analyzed under proper RIP (or restricted strong condition number) bounding conditions [10, 93]. In [41], several relaxed variants of IHT-style algorithms were presented for which the high-dimensional estimation consistency can be established without requiring the RIP conditions. The support recovery performance of IHT-style methods were investigated in [94, 85, 86] to understand when the algorithm can exactly recover the support of a sparse signal from its compressed measurements. In large-scale settings where a full gradient evaluation on all data samples becomes a bottleneck, stochastic and variance reduction techniques have been adopted to improve the computational efficiency of IHT [70, 59, 18]. Recently, a Nesterov's momentum based hard thresholding method was proposed to further improve the efficiency of IHT [43].

3.2.2 Dual Methods

Dual optimization algorithms have gained considerable popularity in various learning tasks including SVMs [34], multi-task learning [52] and graphical models learning [66]. In recent years, many stochastic dual coordinate ascent (SDCA) methods have been proposed for solving large-scale regularized loss minimization problems [81, 82, 83]. All these methods exhibit fast convergence rate in theory and highly competitive numerical performance in practice. A dual free variant of SDCA that supports non-regularized objectives and non-convex individual loss functions was investigated in [78]. To further improve computational efficiency, some primal-dual methods are developed to alternately minimize the primal objective and maximize the dual objective. The successful

examples of primal-dual methods include learning total variation regularized model [16] and generalized Dantzig selector [55]. More recently, a number of stochastic variants were developed to make the primal-dual algorithms more scalable and efficient [99, 92].

Our work lies at the intersection of the above two disciplines of research. Although dual optimization methods have long been understood in machine learning, it still remains largely unknown, in both theory and algorithms, how to apply dual methods to the non-convex and NP-hard problem (3.1.1), where the non-convexity arises from the cardinality constraint rather than the loss function. We are going to close this gap by presenting a sparse Lagrangian duality theory and a dual variant of IHT for solving the related dual maximization problem with provable primal-dual convergence and support recovery guarantees.

3.3 A Sparse Lagrangian Duality Theory

In this section, we establish weak and strong duality theory that guarantees the original non-convex and NP-hard problem in (3.1.1) can be equivalently solved in a dual space. The results in this part build the theoretical foundation of developing dual IHT methods.

From now on we abbreviate $f_i(w^\top x_i) = f(w^\top x_i, y_i)$. The convexity of $f(w^\top x_i, y_i)$ implies that $f_i(u)$ is also convex. Let $f_i^*(\alpha_i) = \max_u\{\alpha_i u - f_i(u)\}$ be the convex conjugate of $f_i(u)$ and $\mathcal{F} \subseteq \mathbb{R}$ be the feasible set of α_i. According to the well-known expression of $f_i(u) = \max_{\alpha_i \in \mathcal{F}} \{\alpha_i u - f_i^*(\alpha_i)\}$, the problem (3.1.1) can be reformulated into the following mini-max formulation:

$$\min_{\|w\|_0 \le k} \frac{1}{N} \sum_{i=1}^{N} \max_{\alpha_i \in \mathcal{F}} \{\alpha_i w^\top x_i - f_i^*(\alpha_i)\} + \frac{\lambda}{2}\|w\|^2. \tag{3.3.1}$$

The following Lagrangian form will be useful in analysis:

$$L(w, \alpha) = \frac{1}{N} \sum_{i=1}^{N} \Big(\alpha_i w^\top x_i - f_i^*(\alpha_i)\Big) + \frac{\lambda}{2}\|w\|^2,$$

where $\alpha = [\alpha_1, ..., \alpha_N] \in \mathcal{F}^N$ is the vector of dual variables. We now introduce the following concept of sparse saddle point which is a restriction of the conventional saddle point to the setting of sparse optimization.

Definition 2 (Sparse Saddle Point)**.** *A pair* $(\bar{w}, \bar{\alpha}) \in \mathbb{R}^p \times \mathcal{F}^N$ *is said to be a* k*-sparse saddle point for* L *if* $\|\bar{w}\|_0 \leq k$ *and the following holds for all* $\|w\|_0 \leq k, \alpha \in \mathcal{F}^N$*:*

$$L(\bar{w}, \alpha) \leq L(\bar{w}, \bar{\alpha}) \leq L(w, \bar{\alpha}). \tag{3.3.2}$$

Different from the conventional definition of saddle point, the k-sparse saddle point only requires the inequality (3.3.2) holds for any arbitrary k-sparse vector w. The following result is a basic sparse saddle point theorem for L. Throughout the chapter, we will use $f'(\cdot)$ to denote a sub-gradient (or super-gradient) of a convex (or concave) function $f(\cdot)$, and use $\partial f(\cdot)$ to denote its sub-differential (or super-differential).

Theorem 2 (Sparse Saddle Point Theorem)**.** *Let* $\bar{w} \in \mathbb{R}^p$ *be a* k*-sparse primal vector and* $\bar{\alpha} \in \mathcal{F}^N$ *be a dual vector. Then the pair* $(\bar{w}, \bar{\alpha})$ *is a sparse saddle point for* L *if and only if the following conditions hold:*

(a) $\bar{w}$ *solves the primal problem in* (3.1.1)*;*

(b) $\bar{\alpha} \in [\partial f_1(\bar{w}^\top x_1), ..., \partial f_N(\bar{w}^\top x_N)]$*;*

(c) $\bar{w} = \mathrm{H}_k\left(-\frac{1}{\lambda N}\sum_{i=1}^N \bar{\alpha}_i x_i\right)$.

Proof. A proof of this result is given in § 3.6.1. □

Remark 3. *Theorem 2 shows that the conditions (a)∼(c) are sufficient and necessary to guarantee the existence of a sparse saddle point for the Lagrangian form* L*. This result is different from the traditional saddle point theorem which requires the use of the Slater Constraint Qualification to guarantee the existence of saddle point.*

Remark 4. *Let us consider* $P'(\bar{w}) = \frac{1}{N}\sum_{i=1}^N \bar{\alpha}_i x_i + \lambda \bar{w} \in \partial P(\bar{w})$*. Denote* $\bar{F} = supp(\bar{w})$*. It is easy to verify that the condition (c) in Theorem 2 is equivalent to*

$$\mathrm{H}_{\bar{F}}(P'(\bar{w})) = 0, \quad \bar{w}_{\min} \geq \frac{1}{\lambda}\|P'(\bar{w})\|_\infty. \tag{3.3.3}$$

We use a quadratic optimization problem as an example to illustrate this condition. Consider

$$P(w) = \frac{1}{2N}\sum_{i=1}^N (y_i - w_i)^2 + \frac{\lambda}{2}\|w\|^2$$

where the model parameter $w \in \mathbb{R}^N$. *In this case we can derive that* $\bar{w} = \mathrm{H}_k(\frac{y}{1+\lambda N})$ *and* $P'(w) = (\lambda + \frac{1}{N})\bar{w} - \frac{1}{N}y$. *Denote* $[|y|]_{(j)}$ *the* j*-th largest entry of* $|y|$, *i.e.,* $[|y|]_{(1)} \geq [|y|]_{(2)} \cdots \geq [|y|]_{(N)}$. *The condition* (3.3.3) *is essentially* $\frac{[|y|]_{(k)}}{1+\lambda N} \geq \frac{[|y|]_{(k+1)}}{\lambda N}$ *which requires* $\lambda \geq \frac{[|y|]_{(k+1)}}{N([|y|]_{(k)} - [|y|]_{(k+1)})}$. *Obviously, we need* $[|y|]_{(k)}$ *to be strictly larger than* $[|y|]_{(k+1)}$ *to guarantee the existence of* λ.

We next propose the following sparse mini-max theorem that guarantees the min and max in (3.3.1) can be safely switched if and only if there exists a sparse saddle point for $L(w, \alpha)$.

Theorem 3 (Sparse Mini-Max Theorem)**.** *The mini-max relationship*

$$\max_{\alpha \in \mathcal{F}^N} \min_{\|w\|_0 \leq k} L(w, \alpha) = \min_{\|w\|_0 \leq k} \max_{\alpha \in \mathcal{F}^N} L(w, \alpha) \tag{3.3.4}$$

holds if and only if there exists a sparse saddle point $(\bar{w}, \bar{\alpha})$ *for* L.

Proof. A proof of this result is given in § 3.6.2. □

The sparse mini-max result in Theorem 3 provides sufficient and necessary conditions under which one can safely exchange a min-max for a max-min, in the presence of the cardinality constraint for primal variable parameter w. By applying Theorem 2 to Theorem 3 we can derive the following corollary.

Corollary 2. *The mini-max relationship*

$$\max_{\alpha \in \mathcal{F}^N} \min_{\|w\|_0 \leq k} L(w, \alpha) = \min_{\|w\|_0 \leq k} \max_{\alpha \in \mathcal{F}^N} L(w, \alpha)$$

holds if and only if there exist a k*-sparse primal vector* $\bar{w} \in \mathbb{R}^p$ *and a dual vector* $\bar{\alpha} \in \mathcal{F}^N$ *such that the conditions (a)∼(c) in Theorem 2 are satisfied.*

The mini-max result in Theorem 3 can be used as a basis for establishing sparse duality theory. Indeed, we have already shown the following:

$$\min_{\|w\|_0 \leq k} \max_{\alpha \in \mathcal{F}^N} L(w, \alpha) = \min_{\|w\|_0 \leq k} P(w).$$

This is called the *primal* minimization problem and it is the min-max side of the sparse mini-max theorem. The other side, the max-min problem, will be called as the *dual*

maximization problem with dual objective function $D(\alpha) := \min_{\|w\|_0 \leq k} L(w, \alpha)$, i.e.,

$$\max_{\alpha \in \mathcal{F}^N} D(\alpha) = \max_{\alpha \in \mathcal{F}^N} \min_{\|w\|_0 \leq k} L(w, \alpha). \tag{3.3.5}$$

The following Proposition 1 shows that the dual objective function $D(\alpha)$ is concave and explicitly gives the expression of its super-differential.

Proposition 1. *The dual objective function $D(\alpha)$ is given by*

$$D(\alpha) = \frac{1}{N} \sum_{i=1}^{N} -f_i^*(\alpha_i) - \frac{\lambda}{2} \|w(\alpha)\|^2,$$

where $w(\alpha) = \mathrm{H}_k\left(-\frac{1}{N\lambda} \sum_{i=1}^{N} \alpha_i x_i\right)$. Moreover, $D(\alpha)$ is concave and its super-differential is given by

$$\partial D(\alpha) = \frac{1}{N}[w(\alpha)^\top x_1 - \partial f_1^*(\alpha_1), ..., w(\alpha)^\top x_N - \partial f_N^*(\alpha_N)].$$

Particularly, if $w(\alpha)$ is unique at α and $\{f_i^*\}_{i=1,...,N}$ are differentiable, then $\partial D(\alpha)$ is unique and it is the super-gradient of $D(\alpha)$.

Proof. A proof of this result is given in §3.6.3. □

Based on Theorem 2 and Theorem 3, we are able to further establish a sparse strong duality theorem which gives the sufficient and necessary conditions under which the optimal values of the primal and dual problems coincide.

Theorem 4 (Sparse Strong Duality Theorem)**.** *Let $\bar{w} \in \mathbb{R}^p$ is a k-sparse primal vector and $\bar{\alpha} \in \mathcal{F}^N$ be a dual vector. Then $\bar{\alpha}$ solves the dual problem in (3.3.5), i.e., $D(\bar{\alpha}) \geq D(\alpha)$, $\forall \alpha \in \mathcal{F}^N$, and $P(\bar{w}) = D(\bar{\alpha})$ if and only if the pair $(\bar{w}, \bar{\alpha})$ satisfies the conditions (a)∼(c) in Theorem 2.*

Proof. A proof of this result is given in § 3.6.4. □

We define the sparse primal-dual gap $\epsilon_{PD}(w, \alpha) := P(w) - D(\alpha)$. The main message conveyed by Theorem 4 is that the sparse primal-dual gap reaches zero at the primal-dual pair $(\bar{w}, \bar{\alpha})$ if and only if the conditions (a)∼(c) in Theorem 2 hold.

3.3.1 On the Dual Sufficient Conditions for Sparse Strong Duality

The previously established strong sparse duality theory relies on the sparsity constraint qualification condition (c) in Theorem 2. This key condition is essentially imposed on the underlying primal sparse minimizer $\bar{w}$ one would like to recover. To make the results more comprehensive, we further provide in the following theorem a sufficient condition imposed on the dual maximizer of $D(\alpha)$ to guarantee zero primal-dual gap. From now on we denote $X = [x_1, ..., x_N] \in \mathbb{R}^{p\times N}$ the data matrix of which the N data samples are columns.

Theorem 5. *Assume that each f_i is differentiable and smooth, and each dual feasible set $\mathcal{F}_i$ is convex. Let $\bar{\alpha}_i = \arg\max_\alpha D(\alpha)$ be a dual maximizer. If $w(\bar{\alpha}) = \mathrm{H}_k(-\frac{1}{N\lambda}\sum_{i=1}^{N}\bar{\alpha}_i x_i)$ is unique with respect to $\bar{\alpha}$, then $(w(\bar{\alpha}), \bar{\alpha})$ is a sparse saddle point and $w(\bar{\alpha})$ is a primal minimizer of $P(w)$ satisfying $P(w(\bar{\alpha})) = D(\bar{\alpha})$.*

Remark 5. *The dual sufficient condition given in Theorem 5 basically shows that under mild conditions if $w(\bar{\alpha})$ constructed from a dual maximizer $\bar{\alpha}$ is unique, then sparse strong duality holds. Such a uniqueness condition is computationally more verifiable than the condition (c) in Theorem 2 as maximizing the dual concave program is easier than minimizing the primal non-convex problem.*

We use sparse linear regression and logistic regression model learning tasks as example to provide intuition of Theorem 5. Those two models are commonly used in statistical machine learning.

Example I: Sparse strong duality for linear regression. The first example is to show that strong sparse duality hold mildly for the sparse linear regression model which is widely applied in compressed sensing. Consider the special case of the primal problem (3.1.1) with least square loss $f(w^\top x_i, y_i) = \frac{1}{2}(y_i - w^\top x_i)^2$. Let us write $f(w^\top x_i, y_i) = f_i(w^\top x_i)$ with $f_i(a) = \frac{1}{2}(a - y_i)^2$. It is standard to know that the convex conjugate of $f_i(a)$ is $f_i^*(\alpha_i) = \frac{\alpha_i^2}{2} + y_i\alpha_i$ and $\mathcal{F}_i \in \mathbb{R}$. Obviously, f_i^* is differentiable and $\mathcal{F}_i$ is convex. By directly applying Theorem 5 to this case we obtain the following corollary showing that strong sparse duality holds for sparse linear regression given that $w(\bar{\alpha})$ is unique.

Corollary 3 (Sparse strong duality for linear regression)**.** *Consider the special case of the primal problem* (3.1.1) *with least square loss* $f(w^\top x_i, y_i) = \frac{1}{2}(y_i - w^\top x_i)^2$*. Let* $\bar{\alpha} = \arg\max_{\alpha} D(\alpha)$ *be a dual maximizer. If* $w(\bar{\alpha}) = \mathrm{H}_k\left(-\frac{1}{N\lambda}\sum_{i=1}^{N}\bar{\alpha}_i x_i\right)$ *is unique with respect to* $\bar{\alpha}$*, then* $w(\bar{\alpha})$ *is a primal minimizer of* $P(w)$ *satisfying* $P(w(\bar{\alpha})) = D(\bar{\alpha})$*.*

Remark 6. *To illustrate the result in the above corollary, we consider the same example as presented in Remark 4. In this case, we have the dual objective function written by*

$$D(\alpha) = \frac{1}{N}\sum_{i=1}^{N}\left\{-\frac{\alpha_i^2}{2} - \alpha_i y_i\right\} - \frac{1}{2\lambda N^2}\|\mathrm{H}_k(\alpha)\|^2$$

Provided that $\frac{\lambda N[|y|]_{(k)}}{1+\lambda N} > [|y|]_{(k+1)}$*, we can directly verify that the dual solution is*

$$[\bar{\alpha}]_{(i)} = \begin{cases} -\frac{\lambda N}{1+\lambda N}[y]_i & i \in \{1, ..., k\} \\ -[y]_{(i)} & i \in \{k+1, ..., N\} \end{cases}$$

and $w(\bar{\alpha}) = \mathrm{H}_k(-\frac{1}{\lambda N}\bar{\alpha})$ *is by definition unique. According to the discussion in Remark 4,* $w(\bar{\alpha})$ *is exactly the primal minimizer. This verifies the validness of Corollary 3 on the considered example.*

Example II: Sparse strong duality for logistic regression. We further show that strong sparse duality hold mildly for the sparse logistic regression model, In logistic regression model, given a k-sparse parameter vector $\bar{w}$, the relation between the random feature vector $x \in \mathbb{R}^p$ and its associated random binary label $y \in \{-1, +1\}$ is determined by the conditional probability $\mathbb{P}(y|x; \bar{w}) = \exp(2y\bar{w}^\top x)/(1 + \exp(2y\bar{w}^\top x))$. The logistic loss over a sample (x_i, y_i) is written by $f(w^\top x_i, y_i) = f_i(w^\top x_i) = \log\left(1 + \exp(-y_i w^\top x_i)\right)$, where $f_i(a) = \log(1 + \exp(-ay_i))$. In this case we have $f^*(\alpha_i) = -\alpha_i y_i \log(-\alpha_i y_i) + (1 + \alpha_i y_i)\log(1 + \alpha_i y_i)$ with $\alpha_i y_i \in [-1, 0]$. Note that f_i^* is differentiable and $\mathcal{F}_i$ is convex. Therefore Theorem 5 implied the following corollary for sparse logistic regression models.

Corollary 4 (Sparse strong duality for logistic regression)**.** *Consider the special case of the primal problem* (3.1.1) *with logistic loss* $f(w^\top x_i, y_i) = \log\left(1 + \exp(-y_i w^\top x_i)\right)$*. Let* $\bar{\alpha} = \arg\max_{\alpha} D(\alpha)$ *be a dual maximizer. If* $w(\bar{\alpha}) = \mathrm{H}_k\left(-\frac{1}{N\lambda}\sum_{i=1}^{N}\bar{\alpha}_i x_i\right)$ *is unique with respect to* $\bar{\alpha}$*, then* $w(\bar{\alpha})$ *is a primal minimizer of* $P(w)$ *satisfying* $P(w(\bar{\alpha})) = D(\bar{\alpha})$*.*

Algorithm 2: Dual Iterative Hard Thresholding (DIHT)

Input : Training set $\{x_i, y_i\}_{i=1}^N$. Regularization strength parameter λ. Cardinality constraint k. Step-size η.

Initialization $w^{(0)} = 0$, $\alpha_1^{(0)} = ... = \alpha_N^{(0)} = 0$.

for $t = 1, 2, ..., T$ **do**

(**S1**) Dual projected super-gradient ascent: $\forall\ i \in \{1, 2, ..., N\}$,

$$\alpha_i^{(t)} = \mathrm{P}_{\mathcal{F}}\left(\alpha_i^{(t-1)} + \eta^{(t-1)} g_i^{(t-1)}\right), \tag{3.3.6}$$

where $g_i^{(t-1)} = \frac{1}{N}(x_i^\top w^{(t-1)} - f_i^{*\prime}(\alpha_i^{(t-1)}))$ is the super-gradient and $\mathrm{P}_{\mathcal{F}}(\cdot)$ is the Euclidian projection operator with respect to feasible set $\mathcal{F}$.

(**S2**) Primal hard thresholding:

$$w^{(t)} = \mathrm{H}_k\left(-\frac{1}{\lambda N}\sum_{i=1}^N \alpha_i^{(t)} x_i\right). \tag{3.3.7}$$

end

Output: $w^{(T)}$.

3.3.2 The Dual Iterative Hard Thresholding Algorithm

The Dual Iterative Hard Thresholding (DIHT) algorithm, as outlined in Algorithm 2, is essentially a projected super-gradient method for maximizing $D(\alpha)$. The procedure generates a sequence of prima-dual pairs $(w^{(0)}, \alpha^{(0)}), (w^{(1)}, \alpha^{(1)}), \ldots$ from an initial pair $w^{(0)} = 0$ and $\alpha^{(0)} = 0$. At the t-th iteration, the dual update step **S1** conducts the projected super-gradient ascent in (3.3.6) to update $\alpha^{(t)}$ from $\alpha^{(t-1)}$ and $w^{(t-1)}$. Then in the primal update step **S2**, the primal variable $w^{(t)}$ is constructed from $\alpha^{(t)}$ using a k-sparse truncation operation in (3.3.7).

3.3.3 The Stochastic Dual Iterative Hard Thresholding Algorithm

When a batch estimation of super-gradient $D'(\alpha)$ becomes expensive in large-scale applications, it is natural to consider the stochastic implementation of DIHT, namely

SDIHT, as outlined in Algorithm 3. Different from the batch computation in Algorithm 2, the dual update step **S1** in Algorithm 3 randomly selects a block of samples (from a given block partition of samples) and update their corresponding dual variables according to (3.3.8). Then in the primal update step **S2.1**, we incrementally update an intermediate accumulation vector $\tilde{w}^{(t)}$ which records $-\frac{1}{\lambda N}\sum_{i=1}^{N}\alpha_i^{(t)}x_i$ as a weighted sum of samples. In **S2.2**, the primal vector $w^{(t)}$ is updated by applying k-sparse truncation on $\tilde{w}^{(t)}$. The SDIHT is essentially a block-coordinate super-gradient method for the dual problem. Particularly, in the extreme case $m = 1$, SDIHT reduces to the batch DIHT. At the opposite extreme end with $m = N$, i.e., each block contains one sample, SDIHT becomes a stochastic coordinate-wise super-gradient method.

The dual update (3.3.8) in SDIHT is much more efficient than DIHT as the former only needs to access a small subset of samples at a time. If the hard thresholding operation in primal update becomes a bottleneck, e.g., in high-dimensional settings, we suggest to use SDIHT with relatively smaller number of blocks so that the hard thresholding operation in **S2.2** can be less frequently called.

3.3.4 Convergence Analysis of DIHT

We now analyze the non-asymptotic convergence behavior of DIHT. In the following analysis, we will denote $\bar{w} = \arg\min_{\|w\|_0\le k} P(w)$ and use the abbreviation $\epsilon_{PD}^{(t)} := \epsilon_{PD}(w^{(t)}, \alpha^{(t)})$. Let $r = \max_{a\in\mathcal{F}}|a|$ be the bound of the dual feasible set $\mathcal{F}$ and $\rho = \max_{i,a\in\mathcal{F}}|f_i^{*\prime}(a)|$. For example, such quantities exist when f_i and f_i^* are Lipschitz continuous [82]. The following is our theorem on the sub-linear convergence of dual parameter estimation error and primal-dual gap of DIHT. The support recovery can be guaranteed after sufficient iteration.

Theorem 6. *Assume that f_i is $1/\mu$-smooth and $\bar{\epsilon} := \bar{w}_{\min} - \frac{1}{\lambda}\|P'(\bar{w})\|_\infty > 0$. Set the step size as $\eta^{(t)} = \frac{N}{\mu(t+2)}$.*

(a) **Parameter estimation error and primal dual gap**: *The sequence $\{\alpha^{(t)}\}_{t\ge1}$ generated by Algorithm 2 satisfies the following estimation error inequality:*

$$\|\alpha^{(t)} - \bar{\alpha}\| \le \frac{r\sqrt{k}\|X\|_{2,\infty}^2 + \lambda\sqrt{N}\rho}{\lambda\mu}\left(\frac{1}{\sqrt{t+2}}\right).$$

Algorithm 3: Stochastic Dual Iterative Hard Thresholding (SDIHT)

Input : Training set $\{x_i, y_i\}_{i=1}^N$. Regularization strength parameter λ. Cardinality constraint k. Step-size η. A block disjoint partition $\{B_1, ..., B_m\}$ of the sample index set $[1, ..., N]$.

Initialization $w^{(0)} = \tilde{w}^{(0)} = 0$, $\alpha_1^{(0)} = ... = \alpha_N^{(0)} = 0$.

for $t = 1, 2, ..., T$ **do**

(**S1**) Dual projected super-gradient ascent: Uniformly randomly select an index block $B_i^{(t)} \in \{B_1, ..., B_m\}$. For all $j \in B_i^{(t)}$ update $\alpha_j^{(t)}$ as

$$\alpha_j^{(t)} = \mathrm{P}_{\mathcal{F}}\left(\alpha_j^{(t-1)} + \eta^{(t-1)} g_j^{(t-1)}\right). \tag{3.3.8}$$

Set $\alpha_j^{(t)} = \alpha_j^{(t-1)}$, $\forall j \notin B_i^{(t)}$.

(**S2**) Primal hard thresholding:

– (**S2.1**) Intermediate update:

$$\tilde{w}^{(t)} = \tilde{w}^{(t-1)} - \frac{1}{\lambda N} \sum_{j \in B_i^{(t)}} (\alpha_j^{(t)} - \alpha_j^{(t-1)}) x_j. \tag{3.3.9}$$

– (**S2.2**) Hard thresholding: $w^{(t)} = \mathrm{H}_k(\tilde{w}^{(t)})$.

end

Output: $w^{(T)}$.

Moreover, the primal-dual gap is bounded as

$$\epsilon_{PD}^{(t)} \leq \frac{(r\sqrt{k}\|X\|_{2,\infty}^2 + \lambda\sqrt{N}\rho)^2}{\lambda^2 \mu N} \left(\frac{\sqrt{k}\|X\|_{2,\infty}^2}{\lambda\sqrt{N}\mu}(1 + \frac{4\|X\|_{2,\infty}\|\bar{\alpha}\|}{\lambda N \bar{\epsilon}}) + 1 \right) \left(\frac{1}{\sqrt{t+2}} \right)$$

(b) **Support recovery**. *The exact support recovery* $supp(w^{(t)}) = supp(\bar{w})$ *holds if*

$$t \geq \left\lceil \frac{4\|X\|_{2,\infty}^2 (r\sqrt{k}\|X\|_{2,\infty}^2 + \lambda\sqrt{N}\rho)^2}{\lambda^4 \mu^2 N^2 \bar{\epsilon}^2} \right\rceil$$

Proof. A proof of this result is given in § 3.6.6. □

Remark 7. *To gain better intuition of the bounds in the theorem, if conventionally choosing the regularization parameter* $\lambda = \mathcal{O}(\frac{1}{\sqrt{N}})$, *then* $\frac{1}{\sqrt{N}}\|\alpha^{(t)} - \bar{\alpha}\| = \mathcal{O}\left(\sqrt{\frac{k}{t}}\right)$, $\epsilon_{PD}^{(t)} = \mathcal{O}\left(\sqrt{\frac{k}{t}}\right)$ *and* $supp(w^{(t)}) = supp(\bar{w})$ *is guaranteed after* $\mathcal{O}(\frac{k}{\bar{\epsilon}^2})$ *steps of iteration.*

Theorem 6 also suggests a computationally tractable termination criterion for DIHT: the algorithm can be stopped when the primal-dual gap becomes sufficiently small and $supp(w^{(t)})$ becomes stable.

Consider primal sub-optimality $\epsilon_P^{(t)} := P(w^{(t)}) - P(\bar{w})$. Since $\epsilon_P^{(t)} \leq \epsilon_{PD}^{(t)}$ always holds, the convergence rates in Theorem 6 are applicable to the primal sub-optimality as well. An interesting observation is that these convergence results on $\epsilon_P^{(t)}$ are not relying on the Restricted Isometry Property (RIP) (or restricted strong condition number) which is required in most existing analysis of IHT-style algorithms [11, 93]. In [41], several relaxed variants of IHT-style algorithms are presented for which the estimation consistency can be established without requiring the RIP conditions. In contrast to the RIP-free sparse recovery analysis in [41], our Theorem 6 does not require the sparsity level k to be relaxed.

3.3.5 Convergence Analysis of SDIHT

When the primal loss functions are Lipschitz continuous, we can similarly establish sub-linear convergence rate bounds for SDIHT, as summarized in the following theorem.

Theorem 7. *Assume that the primal loss functions $\{f_i(\cdot)\}_{i=1}^N$ are $1/\mu$-smooth and $\bar{\epsilon} := \bar{w}_{\min} - \frac{1}{\lambda}\|P'(\bar{w})\| > 0$. Set $\eta^{(t)} = \frac{mN}{\mu(t+2)}$.*

(a) **Parameter estimation error**: *Let $\bar{\alpha} = [f_1'(\bar{w}^\top x_1), ..., f_N'(\bar{w}^\top x_N)]$. The sequence $\{\alpha^{(t)}\}_{t\geq 1}$ generated by Algorithm 3 satisfies the following expected estimation error inequality:*

$$\mathbb{E}[\|\alpha^{(t)} - \bar{\alpha}\|^2] \leq \frac{m(r\sqrt{k}\|X\|_{2,\infty}^2 + \lambda\sqrt{N}\rho)^2}{\lambda^2\mu^2}\left(\frac{1}{t+2}\right),$$

Moreover, the primal-dual gap is bounded in expectation by

$$\mathbb{E}[\epsilon_{PD}^{(t)}] \leq \frac{\sqrt{m}(r\sqrt{k}\|X\|_{2,\infty}^2 + \lambda\sqrt{N}\rho)^2}{\lambda^2\mu N}\left(\frac{\sqrt{k}\|X\|_{2,\infty}^2}{\lambda\mu\sqrt{N}}(1 + \frac{4\|X\|_{2,\infty}\|\bar{\alpha}\|}{\lambda N\bar{\epsilon}}) + 1\right)(\frac{1}{\sqrt{t+2}})$$

(b) **Support recovery**: *For any $\delta \in (0,1)$, it holds with probability at least $1-\delta$ that $supp(w^{(t)}) = supp(\bar{w})$ when*

$$t \geq \left\lceil \frac{4m\|X\|_{2,\infty}^2(r\sqrt{k}\|X\|_{2,\infty}^2 + \lambda\sqrt{N}\rho)^2}{\delta^2\lambda^4\mu^2N^2\bar{\epsilon}} \right\rceil$$

Proof. A proof of this result is given in Appendix 3.6.7. □

Remark 8. *Theorem 7 shows that, up to scaling factors, the expected or high probability iteration complexity of SDIHT is almost identical to that of DIHT. The scaling factor m reflects a trade-off between the decreased per-iteration cost and the increased iteration complexity.*

Remark 9. *The part(b) of Theorem 7 show that $supp(w^{(t)}) = supp(\bar{w})$ occurs with high probability when t is sufficiently large. When this event occurs, SDIHT (with $m = N$) reduces to a restricted version of SDCA [82] over $supp(\bar{w})$, and thus we are able to obtain improved primal-dual gap convergence rate by straightforwardly applying the analysis of SDCA over $supp(\bar{w})$. However, we do not pursue further in that direction as the final state convergence behavior of SDIHT, after exact support recovery, is not the primal gola of this work.*

3.4 Experiments

Numerical experiments are conducted to verify the proposed theory and evaluate the algorithm. We verify the proposed strong sparse duality theorems established in §3.3 through a sparse linear regression model learning task on simulated data. Then we evaluate the efficiency of DIHT/SDIHT on sparse ℓ_2-regularized smooth SVM loss and Hinge loss minimization tasks using real-world datasets.

3.4.1 Theory Verification

For theory verification, we consider the sparse ridge regression model with quadratic loss function $f(w^\top x_i, y_i) = \frac{1}{2}(y_i - w^\top x_i)^2$. The feature points $\{x_i\}_{i=1}^N$ are sampled from standard normal distribution. The responses $\{y_i\}_{i=1}^N$ are generated according to a linear model $y_i = \bar{w}^\top x_i + \varepsilon_i$ with a $\bar{k}$-sparse parameter $\bar{w} \in \mathbb{R}^p$ and random Gaussian noise $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. For this simulation study, we tested with two baseline dimensionality-sparse configurations $(p, \bar{k}) \in \{(30, 5), (500, 50)\}$. For each configuration, we fixed the baseline sample size $N = p$ and studied the effect of varying regularization strength λ,

signal strength $\bar{w}_{\min}$ and noise level σ on the optimal primal-dual gap between primal minimum and dual maximum.

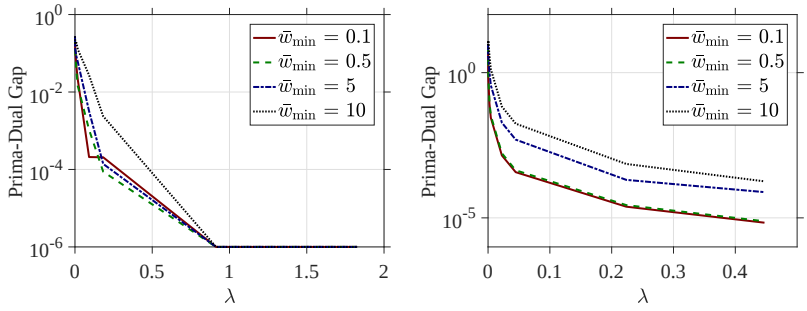
Setup and results of strong sparse duality theory verification. The strong sparse duality theory relies on the sparsity constraint qualification condition (c) in Theorem 2, which essentially requires $\bar{w}_{\min} \geq \frac{1}{\lambda}\|P'(\bar{w})\|_\infty$. In this group of simulation study, keeping all other quantities fixed, we tested how the optimal primal-dual gap evolves under varying $\bar{w}_{\min} \in \{0.1, 0.5, 5, 10\}$ and $\lambda = \frac{\lambda_0}{\sqrt{N}}$ for a wide range of $\lambda_0 \in [10^{-3}, 10]$. To compute the optimal primal-dual gap, we need to find ways to estimate the primal and dual optimal values. For the configuration $(p, \bar{k}) = (30, 5)$, the primal minimizer can be exactly determined via brute-force search among the optimal values over all the feasible index sets of cardinality $\bar{k}$, and the dual maximizer is estimated via running the proposed DIHT algorithm until convergence. For $(p, \bar{k}) = (500, 50)$, it is computationally prohibitive to compute the exact primal minimum. In this case, we just run DIHT on the dual problem until convergence and compute the suboptimal primal-dual gap at the estimated dual maximizer. Figure 3.4.1 shows the optimal primal-dual gap evolving curves as functions of λ under different values of signal strength $\bar{w}_{\min}$. From this group of curves we can make the following key observations:

- For each curve with fixed $\bar{w}_{\min}$, the optimal primal-dual gap decreases as λ increases and the gap reaches zero when λ is sufficiently large. This is as expected because the larger λ is, the easier the condition $\bar{w}_{\min} \geq \frac{1}{\lambda}\|P'(\bar{w})\|_\infty$ can be fulfilled so as to guarantee strong sparse duality.

- For a fixed λ, we note that the primal-dual gap is insensitive to $\bar{w}_{\min}$ especially when λ is sufficiently large and $\bar{w}_{\min}$ is relatively small. This is partially because $\bar{w}$ appears on both sides of the inequality $\bar{w}_{\min} \geq \frac{1}{\lambda}\|P'(\bar{w})\|_\infty$ so that its scale seems not having significant impact on the validness of this inequality.

3.4.2 Algorithm Evaluation

We now turn to evaluate the effectiveness and efficiency of the proposed DIHT/SDIHT algorithm for dual sparse optimization. We begin with a simulation study to confirm



(a) $p = 30$, $\bar{k} = 5$ (b) $p = 500$, $\bar{k} = 50$

Figure 3.4.1: Verification of strong sparse duality theory on linear regression: optimal primal-dual gap evolving curves as functions of regularization strength λ under different levels of signal strength $\bar{w}_{\min}$. For the sake of semi-log curve plotting, we set the primal-dual gap as 10^{-6} when the gap is exactly zero.

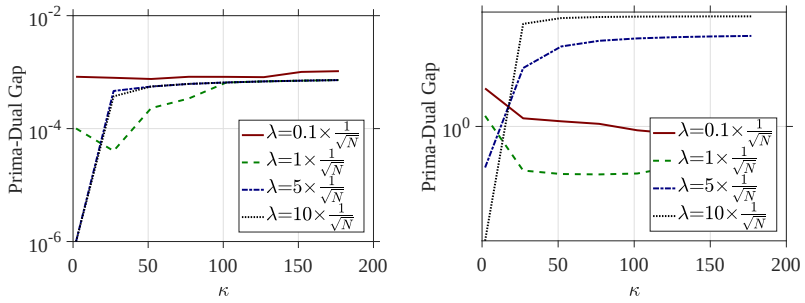
some theoretical properties of DIHT. After that we conduct a set of real data experiments to demonstrate the computational efficiency of DIHT/SDIHT when applied to sparse hinge loss minimization problems.

Simulated Study

The basic setting of this simulation study is identical to the one as described in the theory verification part. As we pointed out at the end of Section 4.1, an interesting theoretical property of DIHT is that its convergence is not relying on the RIP-type conditions which in contrast are usually required by primal IHT-style algorithms. To confirm this point, for each configuration $(p, \bar{k})$, we studied the effect of varying regularization strength λ and condition number of design matrix on the optimal primal-dual gap achieved by DIHT, and make a comparison to some baseline primal IHT-style methods as well.

Convergence of DITH under varying condition number. In this simulation, when λ is fixed and given a desirable condition number κ, we generate feature points $\{x_i\}_{i=1}^{N}$ from multivariate Gaussian distribution $\mathcal{N}(0, \Sigma)$ of which the covariance matrix is carefully designed such that the condition number of $\Sigma + \lambda I$ is κ. In this way of data generation, the condition number of the primal Hessian matrix $\frac{1}{N} X X^{\top} + \lambda I$ is close

to κ. Keeping all other quantities fixed, we tested how the optimal primal-dual gap output by DIHT evolves under varying $\kappa \in [1, 200]$ and regularization strength $\lambda = \frac{\lambda_0}{\sqrt{N}}$ for $\lambda_0 \in \{0.1, 1, 5, 10\}$. Figure 3.4.2 shows the corresponding optimal primal-dual gap evolving curves. From these curves we can observe that the optimal primal-dual gap curves are not sensitive to κ in most cases, especially in badly conditioned cases when $\kappa \geq 50$. This numerical observation confirms our theoretical claim that the convergence behavior of DIHT is not relying on the condition number of problem.

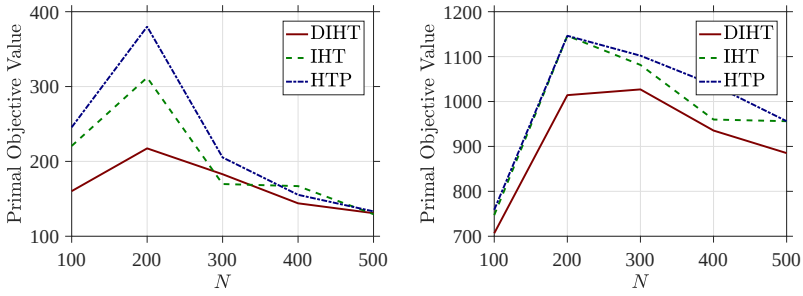


(a) $d = 30$, $\bar{k} = 5$ (b) $d = 500$, $\bar{k} = 50$

Figure 3.4.2: Convergence of DITH under varying condition number of problem: optimal primal-dual gap evolving curves as functions of condition number κ of the Hessian matrix $\frac{1}{N}XX^\top + \lambda I$, under different regularization strength λ.

DIHT versus primal methods on ill-conditioned problems. We further ran experiments to compare DIHT against primal IHT and HTP methods [93, 41] in high condition number setting. For this simulation study, we tested with the dimensionality-sparsity configuration $(p, \bar{k}) = (500, 50)$. To make the problem badly conditioned, we followed a protocol introduced by [41] to select $\bar{k}/2$ random coordinates from the support of nominal parameter vector $\bar{w}$ and $\bar{k}/2$ random coordinates outside its support and constructed a covariance matrix with heavy correlations between these chosen coordinates. The condition number of the resulting matrix was around 50. Keeping all other quantities fixed, we tested how the primal objective values output by the considered algorithms evolve under varying sample size $N \leq p$ and regularization strength $\lambda = \frac{\lambda_0}{\sqrt{N}}$ for $\lambda_0 \in \{1, 10\}$. The resulting curves are plot in Figure 3.4.3. It can be seen from these curves that in most cases DIHT is able to achieve more optimal primal

objective values than IHT and HTP in the considered ill-conditioned problems. We attribute such a numerical benefit of DIHT to its invariance to the condition number of problem.



(a) $\lambda = \frac{1}{\sqrt{N}}$ (b) $\lambda = \frac{10}{\sqrt{N}}$

Figure 3.4.3: DIHT versus primal IHT-style methods on badly conditioned problems: primal objective value evolving curves as functions of sample size N with regularization strength chosen as $\lambda = \frac{1}{\sqrt{N}}$ (left panel) and $\lambda = \frac{10}{\sqrt{N}}$ (right panel).

Real Data: Computational Efficiency Evaluation

For real data experiment, since there is no ground truth model to compare against, we will mainly evaluate the computational efficiency of DIHT. We test with the hinge loss functions which are commonly used by support vector machines. Two binary benchmark datasets from LibSVM data repository[1], RCV1 ($p = 47,236$) [57] and News20 ($p = 1,355,191$) [51], are used for algorithm efficiency evaluation and comparison. We select 0.5 million samples from RCV1 dataset for model training ($N \gg p$). For News20 dataset, all of the $19,996$ samples are used as the training data ($p \gg N$).

Smoothed Hinge Loss. We consider the sparse learning model (3.1.1) with the following smoothed hinge loss function

$$f(w^\top x_i, y_i) = \begin{cases} 0 & y_i w^\top x_i \geq 1 \\ 1 - y_i w^\top x_i - \frac{\gamma}{2} & y_i w^\top x_i < 1 - \gamma \\ \frac{1}{2\gamma}(1 - y_i w^\top x_i)^2 & \text{otherwise} \end{cases} .$$

[1] https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html

Its convex conjugate can be shown as

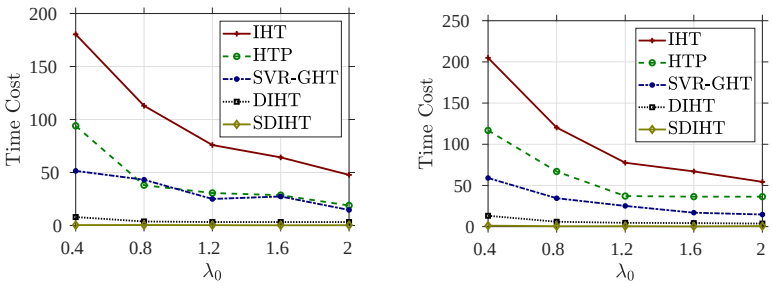
$$f^*(\alpha_i) = \begin{cases} y_i\alpha_i + \frac{\gamma}{2}\alpha_i^2 & \text{if } \ y_i\alpha_i \in [-1, 0] \\ +\infty & \text{otherwise} \end{cases} .$$

We set $\gamma = 0.25$ throughout our experiment. The computational efficiency of DIHT and SDIHT is evaluated by comparing their running time against three primal baseline algorithms: IHT, HTP, and SVR-GHT which is a stochastic variance reduction variant of IHT [59]. Particularly, the following protocol is used for running time comparison: we first run IHT until the sub-optimality criterion $\frac{|P(w^{(t)}) - P(w^{(t-1)})|}{P(w^{(t)})} \leq 10^{-4}$ or a maximum number of iteration is reached, and then test the running time cost spend by other algorithms to reach the same level of primal objective value $P(w^{(t)})$. The parameter update step-size of all the considered algorithms is tuned by grid search. For running the two stochastic algorithms SDIHT and SVR-GHT, we uniformly randomly divide the training data into $|B| = 10$ mini-batches.

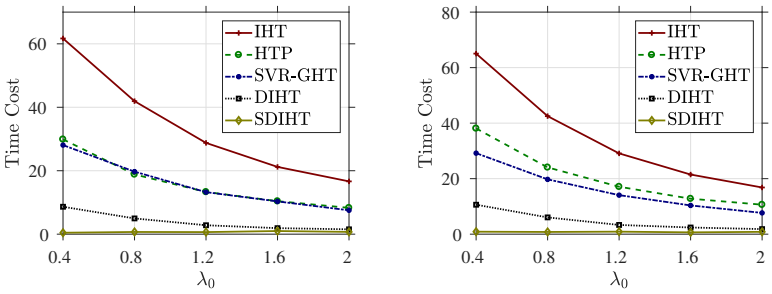
Figure 3.4.4 shows the running time curves on both datasets under varying sparsity level k and regularization strength $\lambda = \frac{\lambda_0}{\sqrt{N}}, \lambda_0 = \{0.4, 0.8, 1.2, 1.6, 2\}$. It is obvious that under all tested (k, λ) configurations on both datasets, DIHT and SDIHT need much less time than their respective primal baseline algorithms IHT, HTP and SVR-GHT to reach the same primal sub-optimality. Also, it can be seen that SDIHT is more efficient than DIHT which matches the consensus that stochastic dual coordinate methods often outperform their batch counterpart [34, 82].

We next evaluate the primal-dual convergence behavior of DIHT and SDIHT, we plot the primal-dual gap evolving curves with respect to the number of epochs. Figure 3.4.5 illustrates the primal-dual gap convergence on both datasets, under sparsity level $k = 1K$ for RCV1 dataset and $k = 50K$ for News20 dataset. The regularization parameters are set to be $\lambda = \frac{\lambda_0}{\sqrt{N}}, \lambda_0 = \{0.4, 1.2, 2\}$, respectively. The results again verify the superior efficiency of SDIHT over DIHT as it uses less number of training sample pass to reach comparable primal-dual gap.

Non-smooth Hinge Loss. Finally, we test the efficiency of our algorithms when applied to the vanilla hinge loss $f(w^\top x_i, y_i) = \max(0, 1 - y_i w^\top x_i)$. It is standard to



(a) RCV1, $k = 500$ (b) RCV1, $k = 1K$



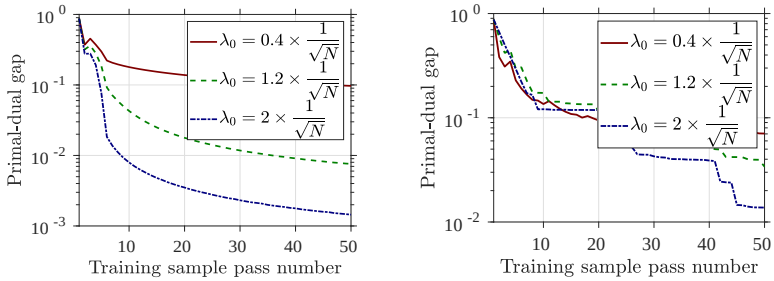
(c) News20, $k = 10K$ (d) News20, $k = 50K$

Figure 3.4.4: Smoothed hinge loss: Running time (in second) comparison of the considered algorithms.

know

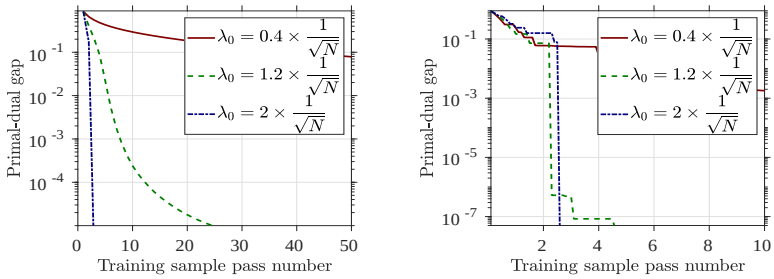
$$f^*(\alpha_i) = \begin{cases} y_i\alpha_i & \text{if } \ y_i\alpha_i \in [-1, 0] \\ +\infty & \text{otherwise} \end{cases} .$$

We follow the same experiment protocol in the smoothed hinge loss model training experiment to compare the considered algorithms on the benchmark datasets. In this non-smooth model training task, we set the step-size in DIHT and SDIHT to be $\eta^{(t)} = \frac{c}{t+2}$, where c is a constant determined by grid search for optimal efficiency. The time cost comparison is illustrated in Figure 3.4.6 and the prima-dual gap sub-optimality is illustrated in Figure 3.4.7. This group of results indicate that DIHT and SDIHT still exhibit remarkable efficiency advantage over the considered primal IHT algorithms even when the loss function is non-smooth.



(a) DIHT on RCV1

(b) SDIHT on RCV1



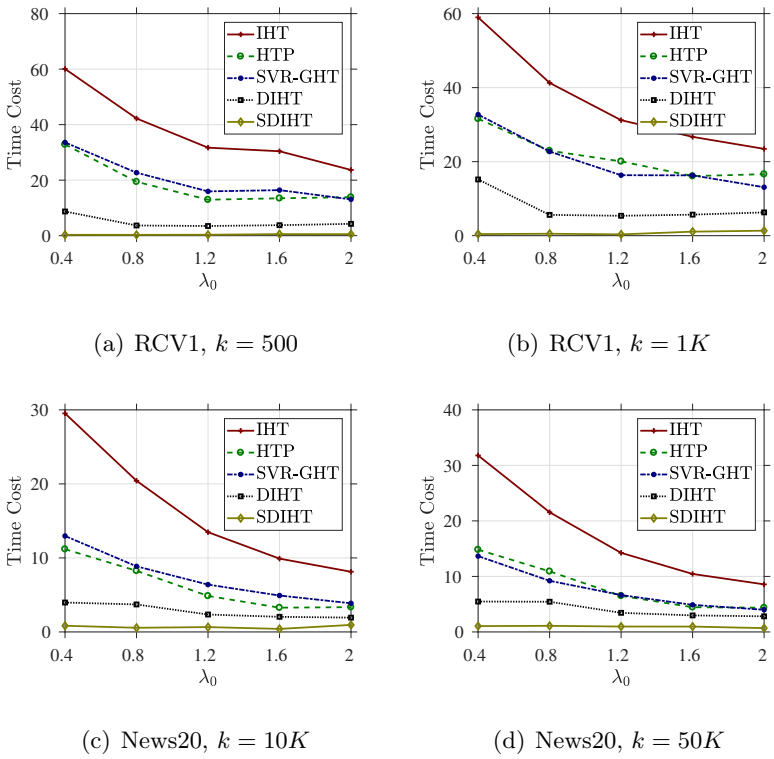
(c) DIHT on News20

(d) SDIHT on News20

Figure 3.4.5: Smoothed hinge loss: The primal-dual gap evolving curves of DIHT and SDIHT. We use sparsity level $k = 1K$ for RCV1 and $k = 50K$ for News20.

3.5 Conclusion and Future Work

In this chapter, we systematically investigated duality theory and optimization algorithms for solving the sparsity-constrained minimization problem which is NP-hard and non-convex in its primal formulation. As the core theoretical contribution, we established a sparse Lagrangian duality theory which guarantees strong duality in sparse settings under certain sufficient and necessary conditions. For the cases when strong duality can be violated, we further developed an approximate duality theory to upper bound the prima-dual gap with statistical estimation error of model. Our theory opens the gate to solve the original NP-hard/non-convex problem equivalently in a dual formulation. We then propose DIHT as a first-order method to maximize the non-smooth dual concave formulation. The algorithm is characterized by dual super-gradient ascent and primal hard thresholding. To further improve iteration efficiency in large-scale settings, we propose SDIHT as a block-coordinate stochastic variant of DIHT. For both



(a) RCV1, $k = 500$

(b) RCV1, $k = 1K$

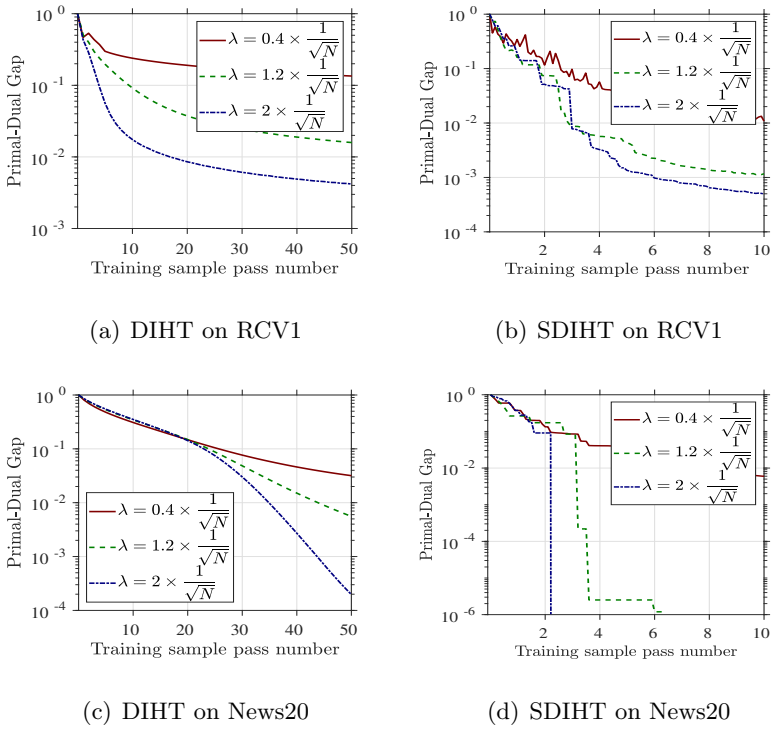
(c) News20, $k = 10K$

(d) News20, $k = 50K$

Figure 3.4.6: Hinge loss: Running time (in second) comparison of the considered algorithms.

algorithms we have proved sub-linear primal-dual gap convergence rate when the loss is smooth, and the improved linear rate can be obtained when the loss is additionally strongly convex. An interesting finding is that these convergence results are valid without assuming RIP-style conditions which are usually required by the existing IHT methods. Based on our theoretical findings and numerical results, we conclude that DIHT and SDIHT are theoretically sound and computationally attractive alternatives to the conventional primal IHT algorithms, especially when the sample size is smaller than feature dimensionality.

Our work leaves several open issues for future exploration. First, it remains an open question on how to verify the key condition (c) in Theorem 2 for general sparse learning models. It will be interesting to provide some more intuitive ways to understand this condition in popular statistical learning models such as linear regression and logistic regression. Second, our convergence results in Theorem 7 merely indicate that SDIHT



(a) DIHT on RCV1

(b) SDIHT on RCV1

(c) DIHT on News20

(d) SDIHT on News20

Figure 3.4.7: Hinge loss: The primal-dual gap evolving curves of DIHT and SDIHT. We use sparsity level $k = 1K$ for RCV1 and $k = 50K$ for News20.

is not worse than DIHT in convergence rate, but without showing that its dependence of scaling factors on sample size N and regularization strength λ can be significantly improved as what has been achieved by SDCA for unconstrained regularized learning [82]. In our opinion, a main challenge we are facing here is the non-smoothness of the dual objective $D(\alpha)$, which prevents us from mimicking the analysis of SDCA to SDIHT. It will be interesting to develop some new proof approaches to justify why SDIHT often outperforms DIHT in practice. Finally, it would be an interesting future work to apply our duality theory and algorithms to communication-efficient distributed sparse learning problems which have recently gained considerable attention in large-scale machine learning [37, 90].

3.6 Appendix

3.6.1 Proof of Theorem 2

Proof. "⇐": If the pair $(\bar{w}, \bar{\alpha})$ is a sparse saddle point for L, then from the definition of conjugate convexity and inequality (3.3.2) we have

$$P(\bar{w}) = \max_{\alpha \in \mathcal{F}} L(\bar{w}, \alpha) \leq L(\bar{w}, \bar{\alpha}) \leq \min_{\|w\|_0 \leq k} L(w, \bar{\alpha}).$$

On the other hand, we know that for any $\|w\|_0 \leq k$ and $\alpha \in \mathcal{F}$

$$L(w, \alpha) \leq \max_{\alpha' \in \mathcal{F}} L(w, \alpha') = P(w).$$

By combining the preceding two inequalities we obtain

$$P(\bar{w}) \leq \min_{\|w\|_0 \leq k} L(w, \bar{\alpha}) \leq \min_{\|w\|_0 \leq k} P(w) \leq P(\bar{w}).$$

Therefore $P(\bar{w}) = \min_{\|w\|_0 \leq k} P(w)$, i.e., $\bar{w}$ solves the problem in (3.1.1), which proves the necessary condition (a). Moreover, the above arguments lead to

$$P(\bar{w}) = \max_{\alpha \in \mathcal{F}} L(\bar{w}, \alpha) = L(\bar{w}, \bar{\alpha}).$$

Then from the maximizing argument property of convex conjugate we know that $\bar{\alpha}_i \in \partial f_i(\bar{w}^\top x_i)$. Thus the necessary condition (b) holds. Note that

$$L(w, \bar{\alpha}) = \frac{\lambda}{2} \left\| w + \frac{1}{N\lambda} \sum_{i=1}^{N} \bar{\alpha}_i x_i \right\|^2 - \frac{1}{N} \sum_{i=1}^{N} f_i^*(\bar{\alpha}_i) + C, \tag{3.6.1}$$

where C is a quantity not dependent on w. Let $\bar{F} = \mathrm{supp}(\bar{w})$. Since the above analysis implies $L(\bar{w}, \bar{\alpha}) = \min_{\|w\|_0 \leq k} L(w, \bar{\alpha})$, it must hold that

$$\bar{w} = \mathrm{H}_{\bar{F}} \left(-\frac{1}{N\lambda} \sum_{i=1}^{N} \bar{\alpha}_i x_i \right) = \mathrm{H}_k \left(-\frac{1}{N\lambda} \sum_{i=1}^{N} \bar{\alpha}_i x_i \right).$$

This validates the necessary condition (c).

"⇒": Conversely, let us assume that $\bar{w}$ is a k-sparse solution to the problem (3.1.1) (i.e., conditio(a)) and let $\bar{\alpha}_i \in \partial f_i(\bar{w}^\top x_i)$ (i.e., condition (b)). Again from the maximizing argument property of convex conjugate we know that $f_i(\bar{w}^\top x_i) = \bar{\alpha}_i \bar{w}^\top x_i - f_i^*(\bar{\alpha}_i)$. This leads to

$$L(\bar{w}, \alpha) \leq P(\bar{w}) = \max_{\alpha \in \mathcal{F}} L(\bar{w}, \alpha) = L(\bar{w}, \bar{\alpha}). \tag{3.6.2}$$

The sufficient condition (c) guarantees that $\bar{F}$ contains the top k (in absolute value) entries of $-\frac{1}{N\lambda}\sum_{i=1}^{N}\bar{\alpha}_i x_i$. Then based on the expression in (3.6.1) we can see that the following holds for any k-sparse vector w

$$L(\bar{w},\bar{\alpha}) \leq L(w,\bar{\alpha}). \tag{3.6.3}$$

By combining the inequalities (3.6.2) and (3.6.3) we get that for any $\|w\|_0 \leq k$ and $\alpha \in \mathcal{F}$,

$$L(\bar{w},\alpha) \leq L(\bar{w},\bar{\alpha}) \leq L(w,\bar{\alpha}).$$

This shows that $(\bar{w},\bar{\alpha})$ is a sparse saddle point of the Lagrangian L. □

3.6.2 Proof of Theorem 3

Proof. "⇒": Let $(\bar{w},\bar{\alpha})$ be a saddle point for L. On one hand, note that the following holds for any k-sparse w' and $\alpha' \in \mathcal{F}$

$$\min_{\|w\|_0\leq k} L(w,\alpha') \leq L(w',\alpha') \leq \max_{\alpha\in\mathcal{F}} L(w',\alpha),$$

which implies

$$\max_{\alpha\in\mathcal{F}} \min_{\|w\|_0\leq k} L(w,\alpha) \leq \min_{\|w\|_0\leq k} \max_{\alpha\in\mathcal{F}} L(w,\alpha). \tag{3.6.4}$$

On the other hand, since $(\bar{w},\bar{\alpha})$ is a saddle point for L, the following is true:

$$\begin{aligned}\min_{\|w\|_0\leq k} \max_{\alpha\in\mathcal{F}} L(w,\alpha) &\leq \max_{\alpha\in\mathcal{F}} L(\bar{w},\alpha)\\ &\leq L(\bar{w},\bar{\alpha}) \leq \min_{\|w\|_0\leq k} L(w,\bar{\alpha})\\ &\leq \max_{\alpha\in\mathcal{F}} \min_{\|w\|_0\leq k} L(w,\alpha).\end{aligned} \tag{3.6.5}$$

By combining (3.6.4) and (3.6.5) we prove the equality in (3.3.4).

"⇐": Assume that the equality in (3.3.4) holds. Let us define $\bar{w}$ and $\bar{\alpha}$ such that

$$\begin{aligned}\max_{\alpha\in\mathcal{F}} L(\bar{w},\alpha) &= \min_{\|w\|_0\leq k} \max_{\alpha\in\mathcal{F}} L(w,\alpha)\\ \min_{\|w\|_0\leq k} L(w,\bar{\alpha}) &= \max_{\alpha\in\mathcal{F}} \min_{\|w\|_0\leq k} L(w,\alpha)\end{aligned}.$$

Then we can see that for any $\alpha \in \mathcal{F}$,

$$L(\bar{w},\bar{\alpha}) \geq \min_{\|w\|_0\leq k} L(w,\bar{\alpha}) = \max_{\alpha'\in\mathcal{F}} L(\bar{w},\alpha') \geq L(\bar{w},\alpha),$$

where the "=" is due to (3.3.4). In the meantime, for any $\|w\|_0 \leq k$,

$$L(\bar{w}, \bar{\alpha}) \leq \max_{\alpha \in \mathcal{F}} L(\bar{w}, \alpha) = \min_{\|w'\|_0 \leq k} L(w', \bar{\alpha}) \leq L(w, \bar{\alpha}).$$

This shows that $(\bar{w}, \bar{\alpha})$ is a sparse saddle point for L. □

3.6.3 Proof of Proposition 1

Proof. For any fixed $\alpha \in \mathcal{F}$, then it is easy to verify that the k-sparse minimum of $L(w, \alpha)$ with respect to w is attained at the following point:

$$w(\alpha) = \arg\min_{\|w\|_0 \leq k} L(w, \alpha) = \mathrm{H}_k \left(-\frac{1}{N\lambda} \sum_{i=1}^{N} \alpha_i x_i \right).$$

Thus we have

$$\begin{aligned} D(\alpha) &= \min_{\|w\|_0 \leq k} L(w, \alpha) = L(w(\alpha), \alpha) \\ &= \frac{1}{N} \sum_{i=1}^{N} \left(\alpha_i w(\alpha)^\top x_i - f_i^*(\alpha_i) \right) + \frac{\lambda}{2} \|w(\alpha)\|^2 \\ &\overset{\zeta_1}{=} \frac{1}{N} \sum_{i=1}^{N} -f_i^*(\alpha_i) - \frac{\lambda}{2} \|w(\alpha)\|^2, \end{aligned}$$

where "ζ_1" follows from the above definition of $w(\alpha)$.

Now let us consider two arbitrary dual variables $\alpha', \alpha'' \in \mathcal{F}$ and any $g(\alpha'') \in \frac{1}{N}[w(\alpha'')^\top x_1 - \partial f_1^*(\alpha_1''), ..., w(\alpha'')^\top x_N - \partial f_N^*(\alpha_N'')]$. From the definition of $D(\alpha)$ and the fact that $L(w, \alpha)$ is concave with respect to α at any fixed w we can derive that

$$\begin{aligned} D(\alpha') &= L(w(\alpha'), \alpha') \\ &\leq L(w(\alpha''), \alpha') \\ &\leq L(w(\alpha''), \alpha'') + \left\langle g(\alpha''), \alpha' - \alpha'' \right\rangle. \end{aligned}$$

This shows that $D(\alpha)$ is a concave function and its super-differential is as given in the theorem.

If we further assume that $w(\alpha)$ is unique and $\{f_i^*\}_{i=1,...,N}$ are differentiable at any α, then $\partial D(\alpha) = \frac{1}{N}[w(\alpha)^\top x_1 - \partial f_1^*(\alpha_1), ..., w(\alpha)^\top x_N - \partial f_N^*(\alpha_N)]$ becomes unique, which implies that $\partial D(\alpha)$ is the unique super-gradient of $D(\alpha)$. □

3.6.4 Proof of Theorem 4

Proof. "⇒": Given the conditions in the theorem, it can be known from Theorem 2 that the pair $(\bar{w}, \bar{\alpha})$ forms a sparse saddle point of L. Thus based on the definitions of sparse saddle point and dual function $D(\alpha)$ we can show that

$$D(\bar{\alpha}) = \min_{\|w\|_0 \leq k} L(w, \bar{\alpha}) \geq L(\bar{w}, \bar{\alpha}) \geq L(\bar{w}, \alpha) \geq D(\alpha).$$

This implies that $\bar{\alpha}$ solves the dual problem in (3.3.5). Furthermore, Theorem 3 guarantees the following

$$D(\bar{\alpha}) = \max_{\alpha \in \mathcal{F}} \min_{\|w\|_0 \leq k} L(w, \alpha) = \min_{\|w\|_0 \leq k} \max_{\alpha \in \mathcal{F}} L(w, \alpha) = P(\bar{w}).$$

This indicates that the primal and dual optimal values are equal to each other.

"⇐": Assume that $\bar{\alpha}$ solves the dual problem in (3.3.5) and $D(\bar{\alpha}) = P(\bar{w})$. Since $D(\bar{\alpha}) \leq P(w)$ holds for any $\|w\|_0 \leq k$, $\bar{w}$ must be the sparse minimizer of $P(w)$. It follows that

$$\max_{\alpha \in \mathcal{F}} \min_{\|w\|_0 \leq k} L(w, \alpha) = D(\bar{\alpha}) = P(\bar{w}) = \min_{\|w\|_0 \leq k} \max_{\alpha \in \mathcal{F}} L(w, \alpha).$$

From the "⇐" argument in the proof of Theorem 3 and Corollary 2 we get that the conditions (a)∼(c) in Theorem 2 should be satisfied for $(\bar{w}, \bar{\alpha})$. □

3.6.5 Proof of Theorem 5

Proof. Recall the dual objective function $D(\alpha)$ is

$$D(\alpha) = \frac{1}{N} \sum_{i=1}^{N} -f_i^*(\alpha_i) - \frac{\lambda}{2} \|w(\alpha)\|^2.$$

Since $w(\bar{\alpha})$ is unique and each f_i^* is differentiable, according to Proposition 1 it is true that the super-gradient of $D(\alpha)$ at $\bar{\alpha}$ is given by $D'(\bar{\alpha}) = \frac{1}{N}[w(\bar{\alpha})^\top x_1 - f_1^{*\prime}(\bar{\alpha}_1), ..., w(\bar{\alpha})^\top x_N - f_N^{*\prime}(\bar{\alpha}_N)]$. Under the conditions in the theorem, we are going to show that for sufficiently small η, the following must hold:

$$\bar{\alpha}_i = \mathrm{P}_{\mathcal{F}_i}\left(\bar{\alpha}_i + \eta \bar{g}_i\right), \tag{3.6.6}$$

where $\bar{g}_i = \frac{1}{N}(x_i^\top w(\bar{\alpha}) - f_i^{*\prime}(\bar{\alpha}_i))$ and $\mathrm{P}_{\mathcal{F}_i}(\cdot)$ is the Euclidian projection operator with respect to feasible set $\mathcal{F}_i$. Before proving this, we need to present a few preliminaries.

For any $\alpha \in \mathcal{F}$, let us define $\tilde{w}(\alpha) = -\frac{1}{N\lambda}\sum_{i=1}^{N}\alpha_i x_i$. For a vector $x \in \mathbb{R}^p$, denote $[x]_{(j)}$ the j-th largest entry (in absolute value) of x, i.e., $|[x]_{(1)}| \geq |[x]_{(2)}| \geq ... \geq |[x]_{(p)}|$. Since $w(\bar{\alpha})$ is unique, or equivalently, the top k entries of $\tilde{w}(\bar{\alpha})$ is unique, we must have $\bar{\epsilon} := [\tilde{w}(\bar{\alpha})]_{(k)} - [\tilde{w}(\bar{\alpha})]_{(k+1)} > 0$. Let $\bar{F} = \mathrm{supp}(w(\bar{\alpha}))$ and define $\mathcal{B}(\bar{\alpha}) = \left\{\alpha \in \mathbb{R}^N : \|\alpha - \bar{\alpha}\| \leq \frac{\lambda N \bar{\epsilon}}{2\|X\|_{2,\infty}}\right\}$.

We prove the equation (3.6.6) by contradiction. Note that for any $\alpha \in \mathcal{B}(\bar{\alpha})$ we have

$$\|\tilde{w}(\alpha) - \tilde{w}(\bar{\alpha})\|_\infty = \frac{1}{N\lambda}\|X(\alpha - \bar{\alpha})\|_\infty \leq \frac{\|X\|_{2,\infty}}{\lambda N}\|\alpha - \bar{\alpha}\| \leq \frac{\bar{\epsilon}}{2}.$$

This indicates that $\mathrm{supp}(w(\alpha)) = \bar{F} = \mathrm{supp}(w(\bar{\alpha}))$. That is, $\bar{F}$ still contains the (unique) top k entries of $\tilde{w}(\alpha)$ for all $\alpha \in \mathcal{B}(\bar{\alpha})$. Consider the vector β with $\beta_i = \bar{\alpha}_i + \eta\bar{g}_i$ with a sufficiently small step-size $\eta > 0$ such that $\beta \in \mathcal{B}(\bar{\alpha})$. Let α' be a vector such that $\alpha'_i = \mathrm{P}_{\mathcal{F}_i}(\beta_i)$. From the non-expanding property of Euclidian projection we know that and $\|\alpha' - \bar{\alpha}\| \leq \|\beta - \bar{\alpha}\|$ and thus $\alpha' \in \mathcal{B}(\bar{\alpha})$. Therefore $\mathrm{supp}(w(\alpha')) = \bar{F}$. Assume that $\alpha' \neq \bar{\alpha}$. Assume that l_i^* is ℓ-smooth. Then

$$\begin{aligned}
D(\alpha') =& \frac{1}{N}\sum_{i=1}^{N} -f_i^*(\alpha'_i) - \frac{\lambda}{2}\|w(\alpha')\|^2 \\
=& \frac{1}{N}\sum_{i=1}^{N} -f_i^*(\alpha'_i) - \frac{\lambda}{2}\left\|\mathrm{H}_{\bar{F}}\left(-\frac{1}{N\lambda}\sum_{i=1}^{N}\alpha'_i x_i\right)\right\|^2 \\
\geq& \frac{1}{N}\sum_{i=1}^{N}\left(-f_i^*(\bar{\alpha}_i) - f_i^{*'}(\alpha_i)(\alpha'_i - \bar{\alpha}_i) - \frac{\ell}{2}(\alpha'_i - \bar{\alpha}_i)^2\right) - \frac{\lambda}{2}\left\|\mathrm{H}_{\bar{F}}\left(-\frac{1}{N\lambda}\sum_{i=1}^{N}\alpha'_i x_i\right)\right\|^2 \\
=& \frac{1}{N}\sum_{i=1}^{N}\left(-f_i^*(\bar{\alpha}_i) - f_i^{*'}(\alpha_i)(\alpha'_i - \bar{\alpha}_i) - \frac{\ell}{2}(\alpha'_i - \bar{\alpha}_i)^2\right) - \frac{\lambda}{2}\|w(\bar{\alpha})\|^2 \\
&+ \frac{1}{N}\sum_{i=1}^{N} x_i^\top w(\bar{\alpha})(\alpha'_i - \bar{\alpha}_i) - \frac{1}{2\lambda N^2}(\alpha' - \bar{\alpha})^\top X_{\bar{F}}^\top X_{\bar{F}}(\alpha' - \bar{\alpha}) \\
\overset{\zeta_1}{\geq}& D(\bar{\alpha}) + \langle D'(\bar{\alpha}), \alpha' - \bar{\alpha}\rangle - \frac{\lambda N\ell + \|X\|^2}{2\lambda N^2}\|\alpha' - \bar{\alpha}\|^2 \\
\overset{\zeta_2}{\geq}& D(\bar{\alpha}) + \left(\frac{1}{2\eta} - \frac{\lambda N\ell + \|X\|^2}{2\lambda N^2}\right)\|\alpha' - \bar{\alpha}\|^2,
\end{aligned}$$

where in "ζ_1" we have used $(\alpha' - \bar{\alpha})^\top X_{\bar{F}}^\top X_{\bar{F}}(\alpha' - \bar{\alpha}) \leq \|X\|^2\|\alpha' - \bar{\alpha}\|^2$, "$\zeta_2$" is due to the fact that $\|\alpha' - \beta\|^2 \leq \|\bar{\alpha} - \beta\|^2$ which then implies $\|\alpha' - \bar{\alpha}\|^2 - 2\eta\langle\alpha' - \bar{\alpha}, D'(\bar{\alpha})\rangle \leq 0$. Since we have assumed $\|\alpha' - \bar{\alpha}\| \neq 0$, by choosing sufficiently small $\eta < \frac{\lambda N^2}{\lambda N\ell + \|X\|^2}$, we can always find $D(\alpha') > D(\bar{\alpha})$, which contradicts the optimality of $\bar{\alpha}$. Therefore $\alpha' = \bar{\alpha}$, i.e., the equation (3.6.6) must hold for sufficiently small η.

Next we show that $(w(\bar{\alpha}), \bar{\alpha})$ forms a sparse saddle point of the Lagrangian of the form:

$$L(w, \alpha) = \frac{1}{N} \sum_{i=1}^{N} \left(\alpha_i w^\top x_i - f_i^*(\alpha_i) \right) + \frac{\lambda}{2} \|w\|^2.$$

Since (3.6.6) holds and $\mathcal{F}_i$ is convex, it must hold that either $\bar{g}_i = \frac{1}{N}(x_i^\top w(\bar{\alpha}) - f_i^{*'}(\bar{\alpha}_i)) = 0$ for $\bar{\alpha}_i$ lies in the interior of $\mathcal{F}_i$, or $\bar{\alpha}_i$ lies on the boundary of $\mathcal{F}_i$ (if it is closed) and it maximizes the function $\frac{1}{N}\left(\alpha_i w^\top x_i - f_i^*(\alpha_i)\right)$. In any case, we always have that $\bar{\alpha}_i$ is a maximizer of $\frac{1}{N}\left(\alpha_i x_i^\top w(\bar{\alpha}) - f_i^*(\alpha_i)\right)$ over the feasible set $\mathcal{F}_i$, which implies $L(w(\bar{\alpha}), \alpha) \leq L(w(\bar{\alpha}), \bar{\alpha})$ holds for any $\alpha \in \mathcal{F}$. From the definition of $w(\bar{\alpha})$ we know that $L(w(\bar{\alpha}), \bar{\alpha}) \leq L(w, \bar{\alpha})$ is valid for all k-sparse primal vector w. Therefore $(w(\bar{\alpha}), \bar{\alpha})$ is a sparse saddle point, and consequently according to Theorem 4 that $w(\bar{\alpha})$ admits a primal k-sparse minimizer. □

3.6.6 Proof of Theorem 6

We need a series of technical lemmas to prove this theorem. The following lemma bounds the estimation error $\|\alpha - \bar{\alpha}\|^2 = \mathcal{O}(\langle D'(\alpha) - D'(\bar{\alpha}), \bar{\alpha} - \alpha \rangle)$ when the primal loss $\{f_i\}_{i=1}^N$ are Lipschitz smooth.

Lemma 3. *Assume that the primal loss functions $\{f_i(\cdot)\}_{i=1}^N$ are $1/\mu$-smooth. Then the following inequality holds for any $\alpha, \alpha'' \in \mathcal{F}$ and $g(\alpha') \in \partial D(\alpha')$, $g(\alpha'') \in \partial D(\alpha'')$:*

$$\|\alpha' - \alpha''\|^2 \leq \frac{N}{\mu} \langle g(\alpha') - g(\alpha''), \alpha'' - \alpha' \rangle.$$

Proof. Recall that

$$D(\alpha) = \frac{1}{N} \sum_{i=1}^{N} -f_i^*(\alpha_i) - \frac{\lambda}{2} \|w(\alpha)\|^2,$$

Now let us consider two arbitrary dual variables $\alpha', \alpha'' \in \mathcal{F}$. The assumption of f_i being $1/\mu$-smooth implies that its convex conjugate function l_i^* is μ-strongly-convex.

Let $F'' = \mathrm{supp}(w(\alpha''))$. We have

$$\begin{aligned}
D(\alpha') =& \frac{1}{N}\sum_{i=1}^{N} -f_i^*(\alpha_i') - \frac{\lambda}{2}\|w(\alpha')\|^2 \\
=& \frac{1}{N}\sum_{i=1}^{N} -f_i^*(\alpha_i') - \frac{\lambda}{2}\left\|\mathrm{H}_k\left(-\frac{1}{N\lambda}\sum_{i=1}^{N}\alpha_i' x_i\right)\right\|^2 \\
\leq& \frac{1}{N}\sum_{i=1}^{N}\left(-f_i^*(\alpha_i'') - f_i^{*'}(\alpha_i'')(\alpha_i' - \alpha_i'') - \frac{\mu}{2}(\alpha_i' - \alpha_i'')^2\right) - \frac{\lambda}{2}\left\|\mathrm{H}_{F''}\left(-\frac{1}{N\lambda}\sum_{i=1}^{N}\alpha_i' x_i\right)\right\|^2 \\
\leq& \frac{1}{N}\sum_{i=1}^{N}\left(-f_i^*(\alpha_i'') - f_i^{*'}(\alpha_i'')(\alpha_i' - \alpha_i'') - \frac{\mu}{2}(\alpha_i' - \alpha_i'')^2\right) - \frac{\lambda}{2}\|w(\alpha'')\|^2 \\
&+ \frac{1}{N}\sum_{i=1}^{N} x_i^\top w(\alpha'')(\alpha_i' - \alpha_i'') - \frac{1}{2\lambda N^2}(\alpha' - \alpha'')^\top X_{F''}^\top X_{F''}(\alpha' - \alpha'') \\
\leq& D(\alpha'') + \langle g(\alpha''), \alpha' - \alpha''\rangle - \frac{\mu}{2N}\|\alpha' - \alpha''\|^2.
\end{aligned}$$

By adding two copies of the above inequality with α and α' interchanged we arrive at

$$\frac{\mu}{N}\|\alpha' - \alpha''\|^2 \leq \langle g(\alpha') - g(\alpha''), \alpha'' - \alpha'\rangle.$$

This leads to the desired inequality in the lemma. □

The following lemma gives a simple expression of the gap for properly connected primal-dual pairs.

Lemma 4. *Given a dual variable $\alpha \in \mathcal{F}$ and the related primal variable*

$$w = \mathrm{H}_k\left(-\frac{1}{N\lambda}\sum_{i=1}^{N}\alpha_i x_i\right).$$

The primal-dual gap $\epsilon_{PD}(w, \alpha)$ can be expressed as:

$$\epsilon_{PD}(w, \alpha) = \frac{1}{N}\sum_{i=1}^{N}\left(f_i(w^\top x_i) + f_i^*(\alpha_i) - \alpha_i w^\top x_i\right).$$

Proof. It is directly to know from the definitions of $P(w)$ and $D(\alpha)$ that

$$\begin{aligned}
& P(w) - D(\alpha) \\
=& \frac{1}{N}\sum_{i=1}^{N} f_i(w^\top x_i) + \frac{\lambda}{2}\|w\|^2 - \left(\frac{1}{N}\sum_{i=1}^{N}\left(\alpha_i w^\top x_i - f_i^*(\alpha_i)\right) + \frac{\lambda}{2}\|w\|^2\right) \\
=& \frac{1}{N}\sum_{i=1}^{N}\left(f_i(w^\top x_i) + f_i^*(\alpha_i) - \alpha_i w^\top x_i\right).
\end{aligned}$$

This shows the desired expression. □

Based on Lemma 4, we can derive the following lemma which establishes a bound on the primal-dual gap.

Lemma 5. *Consider a primal-dual pair* (w, α) *satisfying*

$$w = \mathrm{H}_k\left(-\frac{1}{N\lambda}\sum_{i=1}^{N}\alpha_i x_i\right).$$

Then the following inequality holds for any $g(\alpha) \in \partial D(\alpha)$ *and* $\beta \in [\partial f_1(w^\top x_1), ..., \partial f_N(w^\top x_N)]$*:*

$$P(w) - D(\alpha) \leq \langle g(\alpha), \beta - \alpha\rangle.$$

Proof. For any $i \in [1, ..., N]$, from the maximizing argument property of convex conjugate we have

$$f_i(w^\top x_i) = w^\top x_i f_i'(w^\top x_i) - f_i^*(l_i'(w^\top x_i)),$$

and

$$f_i^*(\alpha_i) = \alpha_i f_i^{*'}(\alpha_i) - f_i(f_i^{*'}(\alpha_i)).$$

By summing both sides of above two equalities we get

$$\begin{aligned}
&f_i(w^\top x_i) + f_i^*(\alpha_i)\\
=&w^\top x_i f_i'(w^\top x_i) + \alpha_i f_i^{*'}(\alpha_i) - (f_i(f_i^{*'}(\alpha_i)) + f_i^*(f_i'(w^\top x_i)))\\
\overset{\zeta_1}{\leq}&w^\top x_i f_i'(w^\top x_i) + \alpha_i f_i^{*'}(\alpha_i) - f_i^{*'}(\alpha_i) f_i'(w^\top x_i),
\end{aligned} \tag{3.6.7}$$

where "ζ_1" follows from Fenchel-Young inequality. Therefore

$$\begin{aligned}
&\langle g(\alpha), \beta - \alpha\rangle\\
=&\frac{1}{N}\sum_{i=1}^{N}(w^\top x_i - f_i^{*'}(\alpha_i))(f_i'(w^\top x_i) - \alpha_i)\\
=&\frac{1}{N}\sum_{i=1}^{N}\left(w^\top x_i f_i'(w^\top x_i) - f_i^{*'}(\alpha_i) f_i'(w^\top x_i) - \alpha_i w^\top x_i + \alpha_i f_i^{*'}(\alpha_i)\right)\\
\overset{\zeta_2}{\geq}&\frac{1}{N}\sum_{i=1}^{N}(f_i(w^\top x_i) + f_i^*(\alpha_i) - \alpha_i w^\top x_i)\\
\overset{\zeta_3}{=}&P(w) - D(\alpha),
\end{aligned}$$

where "ζ_2" follows from (3.6.7) and "ζ_3" follows from Lemma 4. This proves the desired bound. □

The following lemma shows that under proper conditions, $w(\alpha)$ is locally smooth around $\bar{w} = w(\bar{\alpha})$.

Lemma 6. *Assume that* $\{f_i\}_{i=1,\dots,N}$ *are differentiable and* $\bar{\epsilon} := \bar{w}_{\min} - \frac{1}{\lambda}\|P'(\bar{w})\|_\infty > 0$. *Let* $\bar{\alpha} = [f_1'(\bar{w}^\top x_1), ..., f_N'(\bar{w}^\top x_N)]$. *If* $\|\alpha - \bar{\alpha}\| \leq \frac{\lambda N \bar{\epsilon}}{2\|X\|_{2,\infty}}$, *then* $supp(w(\alpha)) = supp(\bar{w})$ *and*

$$\|w(\alpha) - \bar{w}\| \leq \frac{\sqrt{k}\|X\|_{2,\infty}}{\lambda N}\|\alpha - \bar{\alpha}\|.$$

Moreover if $\|\alpha - \bar{\alpha}\| > \frac{\lambda N \bar{\epsilon}}{2\|X\|_{2,\infty}}$, *then*

$$\|w(\alpha) - \bar{w}\| \leq \frac{\sqrt{k}\|X\|_{2,\infty}}{\lambda N}\left(1 + \frac{4\|X\|_{2,\infty}\|\bar{\alpha}\|}{\lambda N \bar{\epsilon}}\right)\|\alpha - \bar{\alpha}\|.$$

Proof. For any $\alpha \in \mathcal{F}$, let us define

$$\tilde{w}(\alpha) = -\frac{1}{N\lambda}\sum_{i=1}^{N}\alpha_i x_i.$$

Consider $\bar{F} = \text{supp}(\bar{w})$. Given $\bar{\epsilon} > 0$, it is known from Theorem 4 that $\bar{w} = \text{H}_{\bar{F}}(\tilde{w}(\bar{\alpha}))$ and $\frac{P'(\bar{w})}{\lambda} = \text{H}_{\bar{F}^c}(-\tilde{w}(\bar{\alpha}))$. Then $\bar{\epsilon} > 0$ implies $\bar{F}$ is unique, i.e., the top k entries of $\tilde{w}(\bar{\alpha})$ is unique, and $\bar{w} = w(\bar{\alpha})$. Given that $\|\alpha - \bar{\alpha}\| \leq \frac{\lambda N \bar{\epsilon}}{2\|X\|_{2,\infty}}$, from the consistency of matrix $\ell_{2,\infty}$-norm we can show that

$$\|\tilde{w}(\alpha) - \tilde{w}(\bar{\alpha})\|_\infty = \frac{1}{N\lambda}\|X(\alpha - \bar{\alpha})\|_\infty \leq \frac{\|X\|_{2,\infty}}{\lambda N}\|\alpha - \bar{\alpha}\| \leq \frac{\bar{\epsilon}}{2}.$$

This indicates that $\bar{F}$ still contains the (unique) top k entries of $\tilde{w}(\alpha)$. Therefore,

$$\text{supp}(w(\alpha)) = \bar{F} = \text{supp}(\bar{w}).$$

Consequently we have

$$\begin{aligned}
\|w(\alpha) - w(\bar{\alpha})\| &= \|\text{H}_{\bar{F}}(\tilde{w}(\alpha)) - \text{H}_{\bar{F}}(\tilde{w}(\bar{\alpha}))\| \\
&\leq \sqrt{k}\|\tilde{w}(\alpha) - \tilde{w}(\bar{\alpha})\|_\infty \\
&= \frac{\sqrt{k}}{N\lambda}\|X(\alpha - \bar{\alpha})\|_\infty \\
&\leq \frac{\sqrt{k}\|X\|_{2,\infty}}{\lambda N}\|\alpha - \bar{\alpha}\|.
\end{aligned}$$

This proves the first desired bound.

Next let us consider the case $\|\alpha - \bar{\alpha}\| > \frac{\lambda N \bar{\epsilon}}{2\|X\|_{2,\infty}}$. From the expression of $w(\alpha)$ we can verify that $\|w(\alpha)\| \leq \frac{\sqrt{k}}{\lambda N}\|X\alpha\|_\infty \leq \frac{\sqrt{k}}{\lambda N}\|X\|_{2,\infty}\|\alpha\|$. Then we have

$$\begin{aligned}
\|w(\alpha) - w(\bar{\alpha})\| &\leq \frac{\sqrt{k}\|X\|_{2,\infty}}{\lambda N}(\|\alpha\| + \|\bar{\alpha}\|) \\
&\leq \frac{\sqrt{k}\|X\|_{2,\infty}}{\lambda N}(\|\alpha - \bar{\alpha}\| + 2\|\bar{\alpha}\|) \\
&\leq \frac{\sqrt{k}\|X\|_{2,\infty}}{\lambda N}\left(\|\alpha - \bar{\alpha}\| + \frac{4\|X\|_{2,\infty}}{\lambda N \bar{\epsilon}}\|\bar{\alpha}\|\|\alpha - \bar{\alpha}\|\right) \\
&= \frac{\sqrt{k}\|X\|_{2,\infty}}{\lambda N}\left(1 + \frac{4\|X\|_{2,\infty}}{\lambda N \bar{\epsilon}}\|\bar{\alpha}\|\right)\|\alpha - \bar{\alpha}\|.
\end{aligned}$$

This shows the second bound. □

We are now in the position to prove Theorem 6.

of Theorem 6. **Part(a):** Let us consider $g^{(t)} \in \partial D(\alpha^{(t)})$ with $g_i^{(t)} = \frac{1}{N}(x_i^\top w^{(t)} - f_i^{*'}(\alpha_i^{(t)}))$. From the expression of $w^{(t)}$ we can verify

$$\|w^{(t)}\| \leq \frac{\sqrt{k}}{\lambda N}\|X\alpha^{(t)}\|_\infty \leq \frac{\sqrt{k}\|X\|_{2,\infty}\|\alpha^{(t)}\|}{\lambda N} \leq \frac{r\sqrt{k}\|X\|_{2,\infty}}{\lambda\sqrt{N}}.$$

Since $\|x_i\| \leq \|X\|_{2,\infty}$, we can show that

$$\|g^{(t)}\| \leq \frac{r\sqrt{k}\|X\|_{2,\infty}^2 + \lambda\sqrt{N}\rho}{\lambda N}. \tag{3.6.8}$$

Let $\bar{g} \in \partial D(\bar{\alpha})$ with $\bar{g}_i = \frac{1}{N}(x_i^\top w(\bar{\alpha}) - f_i^{*'}(\bar{\alpha}_i))$. We will now claim $\bar{g} = 0$. Indeed, Since $\bar{\epsilon} = \bar{w}_{\min} - \frac{1}{\lambda}\|P'(\bar{w})\|_\infty > 0$, from the strong sparse duality theory we can show that $\bar{w} = w(\bar{\alpha})$. Then, according to the fact $f^{*'}(f'(a)) = a$ we can derive $g_i^{(t)} = \frac{1}{N}(x_i^\top \bar{w} - f_i^{*'}(f_i'(x_i^\top \bar{w}))) = \frac{1}{N}(x_i^\top \bar{w} - x_i^\top \bar{w}) = 0$, and thus $\bar{g} = 0$.

Let $h^{(t)} = \|\alpha^{(t)} - \bar{\alpha}\|$ and $v^{(t)} = \langle g^{(t)} - \bar{g}, \bar{\alpha} - \alpha^{(t)}\rangle$. From Lemma 3 we know that $(h^{(t)})^2 \leq N v^{(t)}/\mu$. Then

$$\begin{aligned}
(h^{(t)})^2 =& \|\mathrm{P}_{\mathcal{F}}\left(\alpha^{(t-1)} + \eta^{(t-1)} g^{(t-1)}\right) - \bar{\alpha}\|^2 \\
\leq& \|\alpha^{(t-1)} + \eta^{(t-1)} g^{(t-1)} - \bar{\alpha}\|^2 \\
=& (h^{(t-1)})^2 - 2\eta^{(t-1)}\langle g^{(t-1)}, \bar{\alpha} - \alpha^{(t-1)}\rangle + (\eta^{(t-1)})^2\|g^{(t-1)}\|^2 \\
=& (h^{(t-1)})^2 - 2\eta^{(t-1)}\langle g^{(t-1)} - \bar{g}, \bar{\alpha} - \alpha^{(t-1)}\rangle + (\eta^{(t-1)})^2\|g^{(t-1)}\|^2 \\
\leq& (h^{(t-1)})^2 - \eta^{(t-1)}\frac{2\mu}{N}(h^{(t-1)})^2 + (\eta^{(t-1)})^2\frac{(r\sqrt{k}\|X\|_{2,\infty}^2 + \lambda\sqrt{N}\rho)^2}{\lambda^2 N^2},
\end{aligned}$$

where the first inequality is permitted by the non-expansion property of convex projection operator. Let $\eta^{(t)} = \frac{N}{\mu(t+2)}$. Then we obtain

$$(h^{(t)})^2 \leq \left(1 - \frac{2}{t+1}\right)(h^{(t-1)})^2 + \frac{(r\sqrt{k}\|X\|_{2,\infty}^2 + \lambda\sqrt{N}\rho)^2}{\lambda^2\mu^2(t+1)^2}. \tag{3.6.9}$$

We will now use induction over $t \geq 1$ to prove our claimed bound, i.e., for all $t \geq 1$,

$$(h^{(t)})^2 \leq \frac{c_0}{t+2}.$$

where $c_0 = \frac{(r\sqrt{k}\|X\|_{2,\infty}^2 + \lambda\sqrt{N}\rho)^2}{\lambda^2\mu^2}$. The base-case $t = 1$ follows immediately from (3.6.9). Now considering $t \geq 2$, the bound in (3.6.9) reads as

$$\begin{aligned}(h^{(t)})^2 &\leq \left(1 - \frac{2}{t+1}\right)(h^{(t-1)})^2 + \frac{c_0}{(t+1)^2} \\ &\leq \left(1 - \frac{2}{t+1}\right)\frac{c_0}{t+1} + \frac{c_0}{(t+1)^2} \\ &= \left(1 - \frac{1}{t+1}\right)\frac{c_0}{t+1} \leq \frac{c_0}{t+2},\end{aligned}$$

which is our claimed estimation error bound when $t \geq 2$.

To prove the convergence of primal-dual gap, we consider $\beta^{(t)} := [f_1'(x_1^\top w^{(t)}), ..., f_N'(x_N^\top w^{(t)})]$. According to Lemma 5 we have

$$\begin{aligned}\epsilon_{PD}^{(t)} &= P(w^{(t)}) - D(\alpha^{(t)}) \\ &\leq \langle g^{(t)}, \beta^{(t)} - \alpha^{(t)}\rangle \\ &\leq \|g^{(t)}\|(\|\beta^{(t)} - \bar{\alpha}\| + \|\bar{\alpha} - \alpha^{(t)}\|).\end{aligned}$$

From the smoothness of l_i and Lemma 6 we get

$$\|\beta^{(t)} - \bar{\alpha}\| \leq \frac{\sqrt{N}\|X\|_{2,\infty}}{\mu}\|w^{(t)} - \bar{w}\| \leq \frac{\sqrt{k}\|X\|_{2,\infty}^2}{\lambda\mu\sqrt{N}}\left(1 + \frac{4\|X\|_{2,\infty}\|\bar{\alpha}\|}{\lambda N\bar{\epsilon}}\right)\|\alpha - \bar{\alpha}\|,$$

where in the first "$\leq$" we have used the assumption $\|x_i\| \leq \|X\|_{2,\infty}$. By combining the above with the bound in (3.6.8) we obtain

$$\begin{aligned}\epsilon_{PD}^{(t)} &\leq \|g^{(t)}\|(\|\beta^{(t)} - \bar{\alpha}\| + \|\bar{\alpha} - \alpha^{(t)}\|) \\ &\leq \frac{r\sqrt{k}\|X\|_{2,\infty}^2 + \lambda\sqrt{N}\rho}{\lambda N}\left(\frac{\sqrt{k}\|X\|_{2,\infty}^2}{\lambda\mu\sqrt{N}}\left(1 + \frac{4\|X\|_{2,\infty}\|\bar{\alpha}\|}{\lambda N\bar{\epsilon}}\right) + 1\right)\|\alpha^{(t)} - \bar{\alpha}\| \\ &\leq \frac{(r\sqrt{k}\|X\|_{2,\infty}^2 + \lambda\sqrt{N}\rho)^2}{\lambda^2\mu N}\left(\frac{\sqrt{k}\|X\|_{2,\infty}^2}{\lambda\mu\sqrt{N}}\left(1 + \frac{4\|X\|_{2,\infty}\|\bar{\alpha}\|}{\lambda N\bar{\epsilon}}\right) + 1\right)\left(\frac{1}{\sqrt{t+2}}\right).\end{aligned}$$

This completes the proof of Part(a).

Part(b): Let us consider $\epsilon_0 = \frac{\lambda N \bar{\epsilon}}{2\|X\|_{2,\infty}}$. From Part(a) we obtain

$$\|\alpha^{(t)} - \bar{\alpha}\| \leq \epsilon_0$$

after $t \geq t_0 = \frac{c_0}{\epsilon_0^2}$. In this case, it is known from Lemma 6 that $\text{supp}(w^{(t)}) = \text{supp}(\bar{w})$. This proves the desired bound in Part(b). □

3.6.7 Proof of Theorem 7

Proof. **Part(a):** The proof argument mostly mimics that of Theorem 6. Let $h^{(t)} = \|\alpha^{(t)} - \bar{\alpha}\|$ and $v^{(t)} = \langle g^{(t)} - \bar{g}, \bar{\alpha} - \alpha^{(t)} \rangle$. From Lemma 3 we know that $(h^{(t)})^2 \leq N v^{(t)}/\mu$. For an index set B, denote $g_B^{(t)} := \mathrm{H}_B(g^{(t)})$ and $v_B^{(t)} := \langle g_B^{(t)} - \bar{g}_B, \bar{\alpha} - \alpha^{(t)} \rangle$. Then from the non-expansion property of convex projection operator and the fact of $\bar{g} = 0$ we can show

$$\begin{aligned}
(h^{(t)})^2 =& \|\mathrm{P}_{\mathcal{F}}\left(\alpha^{(t-1)} + \eta^{(t-1)} g^{(t-1)}_{B_{i(t-1)}}\right) - \bar{\alpha}\|^2 \\
\leq& \|\alpha^{(t-1)} + \eta^{(t-1)} g^{(t-1)}_{B_{i(t-1)}} - \bar{\alpha}\|^2 \\
=& (h^{(t-1)})^2 - 2\eta^{(t-1)} v^{(t-1)}_{B_{i(t-1)}} + (\eta^{(t-1)})^2 \|g^{(t-1)}_{B_{i(t-1)}}\|^2.
\end{aligned}$$

By taking conditional expectation (with respect to uniform random block selection, conditioned on $\alpha^{(t-1)}$) on both sides of the above inequality we get

$$\begin{aligned}
&\mathbb{E}[(h^{(t)})^2 \mid \alpha^{(t-1)}] \\
\leq& (h^{(t-1)})^2 - \frac{1}{m}\sum_{i=1}^{m} 2\eta^{(t-1)} v^{(t-1)}_{B_i} + \frac{1}{m}\sum_{i=1}^{m} (\eta^{(t-1)})^2 \|g^{(t-1)}_{B_i}\|^2 \\
=& (h^{(t-1)})^2 - \frac{2\eta^{(t-1)}}{m} v^{(t-1)} + \frac{(\eta^{(t-1)})^2}{m} \|g^{(t-1)}\|^2 \\
\leq& (h^{(t-1)})^2 - \frac{2\eta^{(t-1)}\mu}{mN} (h^{(t-1)})^2 + (\eta^{(t-1)})^2 \frac{(r\sqrt{k}\|X\|_{2,\infty}^2 + \lambda\sqrt{N}\rho)^2}{m\lambda^2 N^2}.
\end{aligned}$$

Let us choose $\eta^{(t)} = \frac{mN}{\mu(t+2)}$. Then we obtain

$$\mathbb{E}[(h^{(t)})^2 \mid \alpha^{(t-1)}] \leq \left(1 - \frac{2}{t+1}\right)(h^{(t-1)})^2 + \frac{m(r\sqrt{k}\|X\|_{2,\infty}^2 + \lambda\sqrt{N}\rho)^2}{\lambda^2\mu^2(t+1)^2}.$$

By taking expectation on both sides of the above over $\alpha^{(t-1)}$, we further get

$$\mathbb{E}[(h^{(t)})^2] \leq \left(1 - \frac{2}{t+1}\right)\mathbb{E}[(h^{(t-1)})^2] + \frac{m(r\sqrt{k}\|X\|_{2,\infty}^2 + \lambda\sqrt{N}\rho)^2}{\lambda^2\mu^2(t+1)^2}.$$

This recursive inequality leads to

$$\mathbb{E}[(h^{(t)})^2] \leq \frac{m(r\sqrt{k}\|X\|_{2,\infty}^2 + \lambda\sqrt{N}\rho)^2}{\lambda^2\mu^2}\left(\frac{1}{t+2}\right).$$

Moreover, similar to the argument in the proof of Theorem 6 we obtain

$$\begin{aligned}
\mathbb{E}[\epsilon_{PD}^{(t)}] &\leq \mathbb{E}[\|g^{(t)}\|(\|\beta^{(t)} - \bar{\alpha}\| + \|\bar{\alpha} - \alpha^{(t)}\|)] \\
&\leq \frac{r\sqrt{k}\|X\|_{2,\infty}^2 + \lambda\sqrt{N}\rho}{\lambda N}\left(\frac{\sqrt{k}\|X\|_{2,\infty}^2}{\lambda\mu\sqrt{N}}\left(1 + \frac{4\|X\|_{2,\infty}\|\bar{\alpha}\|}{\lambda N\bar{\epsilon}}\right) + 1\right)\mathbb{E}[\|\alpha^{(t)} - \bar{\alpha}\|] \\
&\leq \frac{\sqrt{m}(r\sqrt{k}\|X\|_{2,\infty}^2 + \lambda\sqrt{N}\rho)^2}{\lambda^2\mu N}\left(\frac{\sqrt{k}\|X\|_{2,\infty}^2}{\lambda\mu\sqrt{N}}\left(1 + \frac{4\|X\|_{2,\infty}\|\bar{\alpha}\|}{\lambda N\bar{\epsilon}}\right) + 1\right)\left(\frac{1}{\sqrt{t+2}}\right).
\end{aligned}$$

This proves the results in Part(a).

Part(b): Let us consider $\epsilon_0 = \frac{\lambda N\bar{\epsilon}}{2\|X\|_{2,\infty}}$. From Part(a) we obtain

$$\mathbb{E}[\|\alpha^{(t)} - \bar{\alpha}\|] \leq \delta\epsilon_0$$

after $t \geq t_1 = \left\lceil \frac{m(r\sqrt{k}\|X\|_{2,\infty}^2 + \lambda\sqrt{N}\rho)^2}{\lambda^2\mu^2\delta^2\epsilon_0^2} \right\rceil$. Then from the Markov inequality we know that $\|\alpha^{(t)} - \bar{\alpha}\| \leq \mathbb{E}[\|\alpha^{(t)} - \bar{\alpha}\|]/\delta \leq \epsilon_0$ holds with probability at least $1 - \delta$. Lemma 6 shows that $\|\alpha^{(t)} - \bar{\alpha}\| \leq \epsilon_0$ implies $\text{supp}(w^{(t)}) = \text{supp}(\bar{w})$. Therefore when $t \geq t_1$, the event $\text{supp}(w^{(t)}) = \text{supp}(\bar{w})$ occurs with probability at least $1 - \delta$. This proves the desired result in Part(b). □

Chapter 4

Distributed Inexact Newton-type Pursuit for Non-convex Sparse Learning

4.1 Introduction

In chapter 3 we have introduced the work about using dual method to solve the ℓ_2-regularized minimization problem with model parameter ℓ_0-constraint. Now let's consider the following model parameter cardinality-constrained empirical risk minimization (ERM) problem:

$$\min_{w\in\mathbb{R}^p} F(w) := \frac{1}{N}\sum_{i=1}^{N} f(w; x_i, y_i), \quad \text{subject to} \quad \|w\|_0 \leq k, \tag{4.1.1}$$

where $\{x_i, y_i\}_{i=1}^N$ are training samples, f is a general loss function, $\|w\|_0$ represents the number of non-zero entries in w, and k is an integer controlling the cardinality. Due to the presence of cardinality constraint, the problem is non-convex and NP-hard even when f is convex. In this work, we are interested in distributed computing methods for solving such a non-convex ERM problem. In particular, we assume the training data $\mathcal{D} = \{D_1, ..., D_m\}$ with $N = mn$ samples is evenly and randomly distributed over m different machines; each machine j locally stores and accesses n training samples $D_j = \{x_{ji}, y_{ji}\}_{i=1}^n$. Let $F_j(w) := \frac{1}{n}\sum_{i=1}^n f(w; x_{ji}, y_{ji})$ be the local empirical risk evaluated on D_j. The global goal is to minimize the average of these local objectives under cardinality constraint:

$$\min_{w\in\mathbb{R}^p} F(w) = \frac{1}{m}\sum_{j=1}^{m} F_j(w), \quad \text{s.t. } \|w\|_0 \leq k. \tag{4.1.2}$$

We will refer to the above model as ℓ_0-ERM in this chapter.

4.1.1 Iterative Hard Thresholding

The iterative hard thresholding (IHT) methods have demonstrated superior scalability in ℓ_0-ERM problems [4, 93, 41]. The iteration procedure of IHT is as simple as a truncated version of gradient descent step: $w^{(t)} = \mathrm{H}_k\left(w^{(t-1)} - \eta\nabla F(w^{(t-1)})\right)$, where $\mathrm{H}_k(x)$ is a truncation operator which preserves the top k (in magnitude) entries of vector x and sets the remaining to be zero. Let $\bar{w}$ be a $\bar{k}$-sparse target solution. If $F(w)$ is L-smooth and μ_s-strongly-convex over any s-sparse vector space with $s = O(k)$, then it is known from [41] that with some sparsity level $k = O\left(\frac{L^2}{\mu_s^2}\bar{k}\right)$, IHT-style methods reach the estimation error level $\|w^{(t)} - \bar{w}\| = \mathcal{O}\left(\sqrt{k}\|\nabla F(\bar{w})\|_\infty/\mu_s\right)$ after

$$\mathcal{O}\left(\frac{L}{\mu_s}\log\left(\frac{\mu_s\|w^{(0)} - \bar{w}\|}{\sqrt{k}\|\nabla F(\bar{w})\|_\infty}\right)\right) \tag{4.1.3}$$

rounds of iteration. A direct approach for distributed ℓ_0-ERM is a centralized map-reduce implementation of IHT: (1) *map step*: each machine calculates local gradient $\nabla F_j(w^{(t-1)})$ at $w^{(t-1)}$ then send $\nabla F_j(w^{(t-1)})$ to master; (2) *reduce step*: parameter update $w^{(t)} = \mathrm{H}_k(w^{(t-1)} - \eta\frac{1}{m}\sum_{j=1}^m F_j(w^{(t-1)}))$ on a master machine then send $w^{(t)}$ to workers. This distributed IHT approach was first introduced in [72] for compressive sensing. However, as suggested by (4.1.3), the linear dependence of the iteration complexity on the restricted condition number L/μ_s obviously makes the distributed IHT communication inefficient in ill-conditioned problems.

4.1.2 Distributed Approximate Newton-type Methods

For classical distributed ERM problems, the iteration complexity of first-order distributed approaches including gradient descent and ADMM [13] also suffer from the unsatisfactory polynomial dependence on condition number. To tackle this problem, [84] proposed a distributed approximate Newton-type (DANE) method that takes advantage of the stochastic nature of problem: the i.i.d. data samples $\{x_i, y_i\}_{i=1}^N$ are uniformly distributed and each local problem will become sufficiently similar to the global problem when data size increases. If $F(w)$ is quadratic with condition number L/μ, the communication complexity (in high probability) of DANE to reach ϵ-precision was shown to be $\mathcal{O}\left(\frac{L^2}{\mu^2 n}\log(mp)\log\left(\frac{1}{\epsilon}\right)\right)$, which has an improved dependence

on the condition number L/μ which could scale as large as $\mathcal{O}(\sqrt{mn})$ in regularized learning problems. By applying Nesterov's acceleration technique, AIDE [75] further reduces the communication complexity of DANE to $\mathcal{O}\left(\sqrt{\frac{L}{\mu n^{1/2}}}\log(mp)\log\left(\frac{1}{\epsilon}\right)\right)$ in the quadratic case, while allowing the local optimization to be inexact. For more general self-concordant empirical risk functions, [98] proposed DiSCO as a distributed inexact damped Newton method with comparable communication complexity to AIDE. More recently, the EDSL and Two-way Truncation (TWT) method [90, 76] extend DANE to solving ℓ_1-norm regularized ERM problems, obtaining similarly improved dependence of communication cost on condition number. The main finding here is: when the local functions are well structured and sufficiently correlated, distributed Newton-type methods are able to approximate the optimal solution in considerably fewer rounds of communication than those first-order methods.

4.1.3 Overview of Our Approach

In this work, we propose a DANE-type algorithm for distributed ℓ_0-ERM. The method iterates between two main steps: 1) each worker machine (inexactly) solves a variance-reduced local ℓ_0-ERM which is constructed based on the difference between global and local gradients of loss; and 2) the master machine generates the next iterate via properly aggregating the local solutions from workers. In practice, the proposed method has been implemented on parameter server platform [58] with actual performance evaluated on synthetic and real data high dimensional statistical learning tasks.

Although our method shares a similar algorithmic framework with DANE, its iteration complexity analysis turns out to be more challenging due to the presence of non-convex cardinality constraint $\|w\|_0 \leq k$ and potentially non-convex objective functions. Provided that n is sufficiently large and $F(w)$ is convex with restricted Lipschitz continuous Hessian (see Definition 4) and restricted condition number L/μ_s, we show in Theorem 9 that the estimation error $\|w^{(t)} - \bar{w}\| = \mathcal{O}\left(\sqrt{k}\|\nabla F(\bar{w})\|_\infty/\mu_s\right)$ can be guaranteed in high probability after

$$\mathcal{O}\left(\frac{1}{1-\frac{L}{\mu_s}\sqrt{\frac{\log(mp)}{n}}}\log\left(\frac{\mu_s\|w^{(0)}-\bar{w}\|}{\sqrt{k}\|\nabla F(\bar{w})\|_\infty}\right)\right) \tag{4.1.4}$$

rounds of communication. In sharp contrast to the analysis of DANE [84] and AIDE [75] which are restricted to the quadratic case, our bound in (4.1.4) is applicable to a much wider problem spectrum in machine learning. Given that $n = \mathcal{O}\left(\frac{L^2 \log(mp)}{\mu_s^2}\right)$ is sufficiently large and equipped with proper initialization, the bound also implies that in some popular statistical learning models the communication complexity scales in $O(\log(m))$ with respect to the number of machines. In comparison, the sample complexity in [90, 76] for ℓ_1-regularized ERM is $n = \mathcal{O}\left(\frac{s^2 L^2 \log p}{\mu_s^2}\right)$ which is inferior to ours. As another highlight of analysis, we have analyzed our method for non-convex functions, which to our knowledge has not been touched in previous DANE-type methods.

4.1.4 Notation

We denote $\mathrm{H}_k(x)$ as a truncation operator which preserves the top k (in magnitude) entries of vector x and forces the remaining to be zero. The notation $\mathrm{supp}(x)$ represents the index set of nonzero entries of x. We conventionally define $\|x\|_\infty = \max_i |[x]_i|$ and define $x_{\min} = \min_{i \in \mathrm{supp}(x)} |[x]_i|$. For an index set S, we define $[x]_S$ and $[A]_{SS}$ as the restriction of x to S and the restriction of rows and columns of A to S, respectively. For an integer n, we abbreviate the set $\{1, ..., n\}$ to $[n]$.

4.2 The DINPS Method

We now introduce the Distributed Inexact Newton-type PurSuit (DINPS) method. The high level algorithmic procedure of DINPS is outlined in Algorithm 4. Starting from an initial k-sparse approximation $w^{(0)}$, the procedure generates a sequence of intermediate k-sparse iterate $\{w^{(t)}\}_{t \geq 1}$ via distributed local sparse estimation and global synchronization among machines. More precisely, each iteration loop of DINPS can be decomposed into the following three consequent main steps:

Map-reduce gradient computation. In this step, the global gradient $\nabla F(w^{(t-1)}) = \frac{1}{m}\sum_{j=1}^m \nabla F_j(w^{(t-1)})$ is evaluated at the current iterate via simple map-reduce averaging and distributed to all machines for local computation.

Local inexact sparse approximation. Based on the received gradient $\nabla F(w^{(t-1)})$,

each machine j constructs at the current iterate a local objective function (4.2.1) and then inexactly estimate a local k-sparse solution $w_j^{(t)} \approx \arg\min_{\|w\|_0 \le k} P_j(w; w^{(t-1)}, \eta, \gamma)$ up to sparsity level $\bar{k} \le k$ and ϵ-precision. This inexact sparse optimization step can be implemented using IHT-style algorithms which have been witnessed to offer fast and accurate solutions for ℓ_0-estimation [93, 41].

Centralized results aggregation. We compute the truncated average

$$w^{(t)} = \mathrm{H}_k\left(\frac{1}{m}\sum_{j=1}^{m} w_j^{(t)}\right)$$

as the next iterate generated from local sparse predictors. Here the truncation operation is needed to maintain sparsity of output.

Algorithm 4: Distributed Inexact Newton-type PurSuit (DINPS)

Input : Loss functions $\{F_j(w)\}_{j=1}^m$, sparsity level k, parameter $\gamma \ge 0$ and $\eta > 0$.

Initialization $w^{(0)} = 0$ *or* $w^{(0)} \approx \arg\min_{\|w\|_0 \le k} F_1(w)$.

for $t = 1, 2, \ldots$ **do**

Compute $\nabla F(w^{(t-1)}) = \frac{1}{m}\sum_{j=1}^m \nabla F_j(w^{(t-1)})$ and broadcast it to all workers;

for *all the workers* $j = 1, \ldots, m$ *in parallel* **do**

(i) Construct a local objective function:

$$P_j(w; w^{(t-1)} \mid \eta, \gamma) := F_j(w) + \langle \eta \nabla F(w^{(t-1)}) - \nabla F_j(w^{(t-1)}), w\rangle + \frac{\gamma}{2}\|w - w^{(t-1)}\|^2, \tag{4.2.1}$$

(ii) Estimate a k-sparse vector $w_j^{(t)}$ such that for any $\bar{k}$-sparse $\bar{w}$ with $\bar{k} \le k$:

$$P_j(w_j^{(t)}; w^{(t-1)} \mid \eta, \gamma) \le P_j(\bar{w}; w^{(t-1)} \mid \eta, \gamma) + \epsilon;$$

end

Compute $w^{(t)} = \mathrm{H}_k\left(\frac{1}{m}\sum_{j=1}^m w_j^{(t)}\right)$.

end

Output: $w^{(t)}$.

The construction of the local objective (4.2.1) is inspired by the idea of leveraging the first-order gradient information and local higher-order information for local processing as originally introduced in DANE [84]. Compared to those first-order distributed methods [13, 37], such a way of local computation is known to be able to take advantage of inter-machine statistical correlation to dramatically reduce the frequency of

communication. Similar local optimization strategy was also considered by [90, 76] for l_1-regularized sparse learning. Different from these existing DANE-type approaches for convex optimization, our method is designed for ℓ_0-ERM of which the constraint and objective function can both be non-convex.

For initialization, the simplest way is to set $w^{(0)} = 0$, i.e., starting the iteration from scratch. Since the data samples are assumed to be evenly and randomly distributed on machines, another reasonable option of initialization is to minimize one of the local ℓ_0-ERM problems, say $w^{(0)} \approx \arg\min_{\|w\|_0 \leq k} F_1(w)$, which is expected to be close to the global solution.

4.3 Analysis for Convex Functions

In this section, we analyze the rate-of-convergence performance of DINPS for convex objective functions.

4.3.1 Preliminaries

We start by introducing the concept of restricted strong convexity and smoothness which are conventionally used in analyzing greedy pursuit methods [80, 93, 41].

Definition 3 (Restricted Strong Convexity/Smoothness)**.** *For any integer $s > 0$, we say $f(w)$ is restricted μ_s-strongly-convex and L_s-smooth if $\frac{\mu_s}{2}\|w-w'\|^2 \leq f(w)-f(w')-\langle \nabla f(w'), w-w'\rangle \leq \frac{L_s}{2}\|w-w'\|^2$ holds for $\forall w, w'$ with $\|w-w'\|_0 \leq s$.*

We next introduce the concept of restricted Lipschitz continuous gradient and Hessian which characterizes the continuity of the gradient vector and Hessian matrix over sparse subspaces. To simplify the notation, we will use abbreviations $\nabla_S f := [\nabla f]_S$ and $\nabla^2_{SS} f := [\nabla^2 f]_{SS}$.

Definition 4 (Restricted Lipschitz Gradient/Hessian)**.** *We say $f(w)$ has Restricted Lipschitz Gradient with constant $\alpha_s \geq 0$ (or α_s-RLG) if $\|\nabla_S f(w) - \nabla_S f(w')\| \leq \alpha_s \|w - w'\|$ holds for all w, w' with $\|w - w'\|_0 \leq s$ and $S = supp(w) \cup supp(w')$. Moreover, suppose that $f(w)$ is twice continuously differentiable. We say $f(w)$ has Restricted Lipschitz Hessian with constant $\beta_s \geq 0$ (or β_s-RLH) if $\left\|\nabla^2_{SS} f(w) - \nabla^2_{SS} f(w')\right\| \leq$*

$\beta_s\|w - w'\|$.

The RLH property of logistic loss function. Consider the logistic loss $f(w) = \frac{1}{n}\sum_{i=1}^{n}\log(1+\exp(-2y_i w^\top x_i))$ for some $y = (y_i) \in \{-1,+1\}^n$ and $X^n = (x_i) \in \mathbb{R}^{n\times p}$. We need to access the gradient and Hessian of the logistic loss $f(w)$. Let $\sigma(z) = 1/(1+\exp(-z))$ be the sigmoid function. It is easy to show that the gradient $\nabla f(w) = Xa(w)/n$ where $a(w) \in \mathbb{R}^n$ with $[a(w)]_i = -2y_i(1-\sigma(2x_i w^\top u_i))$; and the Hessian $\nabla^2 f(w) = X\Lambda(w)X^\top/n$ where $\Lambda(w)$ is an $n\times n$ diagonal matrix whose diagonal entries $[\Lambda(w)]_{ii} = 4\sigma(2v_i w^\top u_i)(1-\sigma(2v_i w^\top u_i))$. The following proposition further shows that the logistic loss has RLH.

Proposition 2. *Given a cardinality number s. Assume that $\|[x_i]_s\| \le r_s$ holds for all x_i. Let $\Sigma_n = \frac{1}{n}XX^\top$ be the sample covariance matrix. Then the logistic loss $f(w)$ has β_s-RLH with $\beta_s = 24 r_s \rho_s^{\max}(\Sigma_n)$.*

Proof. See Appendix 4.7.4 for a proof of this result. □

4.3.2 Results for quadratic objective functions

Here we present our results in a special setting where $F(w)$ is quadratic with RLH strength parameter $\beta_s \equiv 0$ for all s. The widely applied sparse least square regression model belongs to this case. We need in our analysis the concept of sparse largest/smallest eigenvalue of a square matrix.

Definition 5 (Sparse Largest/Smallest Eigenvalues)**.** *Let $H \in \mathbb{R}^{p\times p}$ be a square matrix. we define the largest s-sparse eigenvalue of H as*

$$\rho_s^{\max}(H) = \max_{w\in\mathbb{R}^p}\left\{w^\top H w \mid \|w\|_0 \le s, \|w\| = 1\right\},$$

and the smallest s-sparse eigenvalue of H as

$$\rho_s^{\min}(H) = \min_{w\in\mathbb{R}^p}\left\{w^\top H w \mid \|w\|_0 \le s, \|w\| = 1\right\}.$$

The following is a **deterministic result** on sparse parameter estimation error of DINPS when the objective $F(w)$ is quadratic.

Theorem 8. *Let $\bar{w}$ be a $\bar{k}$-sparse target vector with $\bar{k} \leq k$. Assume that each component $F_j(w)$ is quadratic with a Hessian matrix H_j and $\rho_{3k}^{\min}(H_j) \geq \mu_{3k} > 0$. Let $H = \frac{1}{m}\sum_{j=1}^{m} H_j$. Assume that $\max_j \|H_j - \eta H\| \leq \frac{\theta \mu_{3k}}{3.24}$ for some $\theta \in (0,1)$ and $\epsilon \leq \frac{k\eta^2 \|\nabla F(\bar{w})\|_\infty^2}{5.29\mu_{3k}}$. Set $\gamma = 0$. Then Algorithm 4 will output solution $w^{(t)}$ satisfying*

$$\|w^{(t)} - \bar{w}\| \leq \frac{7.62\eta\sqrt{k}\|\nabla F(\bar{w})\|_\infty}{(1-\theta)\mu_{3k}}$$

after $t \geq \frac{1}{1-\theta}\log\left(\frac{\mu_{3k}\|w^{(0)}-\bar{w}\|}{\eta\sqrt{k}\|\nabla F(\bar{w})\|_\infty}\right)$ rounds of iteration.

Proof. A proof of this result is given in Appendix 4.7.2. □

The result established in Theorem 8 shows that under proper conditions: 1) the estimation error of DINPS is controlled by the multiplier of $\sqrt{k}\|\nabla F(\bar{w})\|_\infty$ which usually represents the optimal statistical error in high dimensional learning models; and 2) the rate of convergence before reaching the error region is linear.

We now turn to a **stochastic setting** where the samples are uniformly randomly distributed over m machines. The following lemma, which is based on a matrix concentration bound [88], shows that the Hessian H_j is close to H when sample size is sufficiently large. The same result appears in [84].

Lemma 7. *Assume that $\|\nabla^2 f(w^\top x_{ji}, y_{ji})\| \leq L$ holds for all $j \in [m]$ and $i \in [n]$. Let $H_j = \frac{1}{n}\sum_{i=1}^{n} \nabla^2 f(w^\top x_{ji}, y_{ji})$ and $H = \frac{1}{m}\sum_{j=1}^{m} H_j$. Then for each j, with probability at least $1-\delta$ over the samples, $\max_j \|H_j - H\| \leq \sqrt{\frac{32L^2\log(mp/\delta)}{n}}$.*

Equipped with Lemma 7, we are able to derive the following result as a specialization of Theorem 8 to the considered stochastic setting.

Corollary 5. *Let $\bar{w}$ be a $\bar{k}$-sparse target vector with $\bar{k} \leq k$. Assume that the samples are uniformly randomly distributed on m machines and the conditions in Theorem 8 hold. Assume $\|\nabla^2 f(w^\top x_{ji}, y_{ji})\| \leq L$ holds for all $j \in [m]$ and $i \in [n]$. Set $\gamma = 0$ and $\eta = 1$. For any $\delta \in (0,1)$, if $n > \frac{336L^2\log(mp/\delta)}{\mu_{3k}^2}$, then with probability at least $1-\delta$, Algorithm 4 will output solution $w^{(t)}$ satisfying $\|w^{(t)} - \bar{w}\| \leq \frac{7.62\eta\sqrt{k}\|\nabla F(\bar{w})\|_\infty}{(1-\theta)\mu_{3k}}$ after $t \geq \frac{1}{1-\theta}\log\left(\frac{\mu_{3k}\|w^{(0)}-\bar{w}\|}{\sqrt{k}\|\nabla F(\bar{w})\|_\infty}\right)$ rounds of iteration with $\theta = \frac{L}{\mu_{3k}}\sqrt{\frac{336\log(mp/\delta)}{n}}$.*

Proof. See Appendix 4.7.2 for a proof of this corollary. □

The main message conveyed by Corollary 5 is that in the quadratic case, the contraction factor θ can be arbitrarily small given that the sample size $n = \mathcal{O}\left(\frac{L^2 \log(mp)}{\mu_{3k}^2}\right)$ is sufficiently large. This sample size complexity is clearly superior to the corresponding $n = \mathcal{O}\left(\frac{k^2 L^2 \log p}{\mu_{3k}^2}\right)$ complexity established in [90, 76] for ℓ_1-regularized sparse linear regression models.

4.3.3 Results for objective functions with RLH

We now study the case where the objective functions are twice differentiable with RLH. The following is a **deterministic result** on sparse parameter estimation error of DINPS in the considered setting.

Theorem 9. *Let $\bar{w}$ be a $\bar{k}$-sparse target vector with $\bar{k} \leq k$. Let $\bar{H}_j = \nabla^2 F_j(\bar{w})$ and $\bar{H} = \frac{1}{m}\sum_{j=1}^m \bar{H}_j$. Assume that: **(a)** $F_j(w)$ is μ_{3k}-strongly-convex and has β_{3k}-RLH; **(b)** $\max_j \|\bar{H}_j - \eta\bar{H}\| \leq \frac{\theta\mu_{3k}}{6.48}$ for some $\theta \in (0,1)$, $\|\nabla F(\bar{w})\|_\infty \leq \frac{\theta(1-\theta)\mu_{3k}^2}{21.45\eta(1+\eta)\beta_{3k}\sqrt{k}}$, and $\epsilon \leq \frac{k\eta^2\|\nabla F(\bar{w})\|_\infty^2}{5.29\mu_{3k}}$; **(c)** $\|w^{(0)} - \bar{w}\| \leq \frac{\theta\mu_{3k}}{3.24(1+\eta)\beta_{3k}}$. Set $\gamma = 0$. Then Algorithm 4 will output $w^{(t)}$ satisfying*

$$\|w^{(t)} - \bar{w}\| \leq \frac{7.62\eta\sqrt{k}\|\nabla F(\bar{w})\|_\infty}{(1-\theta)\mu_{3k}}$$

after $t \geq \frac{1}{1-\theta}\log\left(\frac{\mu_{3k}\|w^{(0)}-\bar{w}\|}{\eta\sqrt{k}\|\nabla F(\bar{w})\|_\infty}\right)$ rounds of iteration.

Proof. A proof of this result is given in Appendix 4.7.3. □

Given that $w^{(0)}$ is properly initialized and the gradient infinity norm $\|\nabla F(\bar{w})\|_\infty$ is sufficiently small, the estimation error of DINPS for RLH objectives is controlled by the multiplier of $\sqrt{k}\|\nabla F(\bar{w})\|_\infty$ which typically represents the optimal statistical error in sparse learning models; and the rate of convergence towards this error level is linear.

As a direct consequence of Theorem 9, if we further assume $\bar{w}_{\min} > \frac{7.62\eta\sqrt{k}\|\nabla F(\bar{w})\|_\infty}{(1-\theta)\mu_{3k}}$, then *support recovery* $\mathrm{supp}(w^{(t)}) \supseteq \mathrm{supp}(\bar{w})$ can be guaranteed at $w^{(t)}$.

Stochastic result. By plugging Lemma 7 to Theorem 9 we obtain the following stochastoc result of DINPS for objectives with RLH.

Corollary 6. *Let $\bar{w}$ be a $\bar{k}$-sparse target vector with $\bar{k} \leq k$. Assume that the samples are uniformly randomly distributed on m machines and the conditions in Theorem 9 and Lemma 7 hold. Set $\gamma = 0$ and $\eta = 1$. For any $\delta \in (0,1)$, if $n > \frac{1344L^2 \log(mp/\delta)}{\mu_{3k}^2}$, then with probability at least $1-\delta$, Algorithm 4 will output $w^{(t)}$ satisfying $\|w^{(t)} - \bar{w}\| \leq \frac{7.62\sqrt{k}\|\nabla F(\bar{w})\|_\infty}{(1-\theta)\mu_{3k}}$ after $t \geq \frac{1}{1-\theta}\log\left(\frac{\mu_{3k}\|w^{(0)}-\bar{w}\|}{\sqrt{k}\|\nabla F(\bar{w})\|_\infty}\right)$ rounds of iteration with $\theta = \frac{L}{\mu_{3k}}\sqrt{\frac{1344\log(mp/\delta)}{n}}$.*

Proof. See § 4.7.3 for a proof of this corollary. □

Corollary 6 shows that when objective functions have RLH, provided that sample size $n = \mathcal{O}\left(\frac{L^2\log(mp)}{\mu_{3k}^2}\right)$ is sufficiently large, the contraction factor θ can be well controlled to remove the dependency on condition number L/μ_{3k}. This sample complexity improves the corresponding $n = \mathcal{O}\left(\frac{k^2L^2\log p}{\mu_{3k}^2}\right)$ complexity presented in [90, 76] for distributed sparse learning.

On local initialization. The iteration complexity results established in Theorem 9 and Corollary 6 rely on the initialization error $\|w^{(0)} - \bar{w}\|$. Let us consider an ideal local initialization strategy of $w^{(0)} = \arg\min_{\|w\|_0 \leq k} F_1(w)$. If the component $F_1(w)$ is μ_{3k}-strongly convex then it can be verified that $\|w^{(0)} - \bar{w}\| \leq \frac{2.84\sqrt{k}\|\nabla F_1(\bar{w})\|_\infty}{\mu_{3k}}$. By plugging this error bound to Corollary 6, the iteration complexity of DINPS for RLH objectives can be bounded from above by

$$\mathcal{O}\left(\frac{1}{1-\frac{L}{\mu_s}\sqrt{\frac{\log(mp)}{n}}}\log\left(\frac{\|\nabla F_1(\bar{w})\|_\infty}{\|\nabla F(\bar{w})\|_\infty}\right)\right). \tag{4.3.1}$$

In the following example, we will show that the term $\log\left(\frac{\|\nabla F_1(\bar{w})\|_\infty}{\|\nabla F(\bar{w})\|_\infty}\right)$ scales as $\log(m)$ in logistic regression.

Implications for distributed sparse logistic regression. As an example, we briefly discuss the implications of our results for distributed sparse logistic regression models. The logistic loss over data D_j is defined as $F_j(w) = \frac{1}{n}\sum_{i=1}^n \log\left(1+\exp(-y_{ji}w^\top x_{ji})\right)$. Let $F(w) = \frac{1}{m}\sum_{j=1}^m F_j(w)$ be the average of local loss. From Proposition 2 we know that each local logistic loss has RLH. Suppose x_{ji} are sub-Gaussian with parameter σ. It is known that $\|\nabla F(\bar{w})\|_\infty = \mathcal{O}\left(\sigma\sqrt{\log p/(mn)}\right)$ and $\|\nabla F_j(\bar{w})\|_\infty = \mathcal{O}\left(\sigma\sqrt{\log p/n}\right)$ hold with high probability [93]. Then with the local initialization

$w^{(0)} \approx \arg\min_{\|w\|_0 \le k} F_1(w)$, the bound in (4.3.1) suggests that DINPS essentially needs $O(\log m)$ rounds of iteration/communication to reach the statistical error barrier $\mathcal{O}\left(\sigma\sqrt{k\log p/(mn)}\right)$.

4.3.4 Results for general strongly-convex functions

For more general strongly convex $F(w)$ without necessarily having RLH, the following result further shows that by using properly relaxed sparsity levels, DINPS can accurately estimate parameters. For the purpose of readability, we only consider the ideal case where the local subproblems are solved exactly with $\epsilon = 0$, although the result generalizes easily to the inexact case of $\epsilon > 0$.

Theorem 10. *Let $\bar{w}$ be a $\bar{k}$-sparse vector. Assume that:* ***(a)*** *$F_j(w)$ is μ_{3k}-strongly-convex and L_{3k}-smooth;* ***(b)*** *$k > \left(1 + \left(\frac{L_{3k}^2+\mu_{3k}L_{3k}-\mu_{3k}^2}{\mu_{3k}^2}\right)^2\right)\bar{k}$ and $\epsilon = 0$. Set $\gamma = \frac{L_{3k}^2-\mu_{3k}^2}{2\mu_{3k}}$ and $\eta = \frac{L_{3k}}{\mu_{3k}}$. Then the Algorithm 4 will output*

$$\|w^{(t)} - \bar{w}\| \le \frac{4.47L_{3k}\sqrt{k}}{(1-\theta)(L_{3k}^2-\mu_{3k}^2)}\|\nabla F(\bar{w})\|_\infty$$

after $t \ge \frac{1}{1-\theta}\log\frac{(1-\theta)(L_{3k}^2-\mu_{3k}^2)\|w^{(0)}-\bar{w}\|}{\theta L_{3k}\sqrt{k}\|\nabla F(\bar{w})\|_\infty}$ rounds of iteration with the contraction factor

$$\theta = \frac{(L_{3k}^2-\mu_{3k}^2)(\sqrt{k-\bar{k}}+\sqrt{\bar{k}})}{(L_{3k}^2+\mu_{3k}^2)\sqrt{k-\bar{k}}-\sqrt{\bar{k}}(L_{3k}^2-\mu_{3k}^2+2\mu_{3k}L_{3k})}.$$

Proof. A proof of this result is given in § 4.7.5. □

Remark 10. *Theorem 10 actually generalizes those RIP-free convergence results for IHT [41] to DINPS.*

Particularly, if $F(w)$ has bounded restricted strong condition number, then we can establish linear convergence of DINPS without relaxing the sparsity level k, as formally stated in the following Theorem.

Theorem 11. *Let $\bar{w}$ be a $\bar{k}$-sparse target vector with $\bar{k} \le k$. Assume that each component $F_j(w)$ is μ_{3k}-strongly-convex and L_{3k}-smooth. Assume that $\frac{L_{3k}}{\mu_{3k}} < 1.37$ and $\epsilon \le \frac{kL_{3k}^2\|\nabla F(\bar{w})\|_\infty^2}{10.58(L_{3k}^2+\mu_{3k}^2)\mu_{3k}}$. Set $\gamma = \frac{L_{3k}^2-\mu_{3k}^2}{2\mu_{3k}}$ and $\eta = \frac{L_{3k}}{\mu_{3k}}$. Then Algorithm 4 will output*

solution $w^{(t)}$ *satisfying*

$$\|w^{(t)} - \bar{w}\| \leq \frac{8.47 L_{3k}\sqrt{k}}{L_{3k}^2 + \mu_{3k}^2} \|\nabla F(\bar{w})\|_\infty$$

after $t \geq \frac{1}{1-\theta} \log\left(\frac{(1-\theta)(L_{3k}^2+\mu_{3k}^2)\|w^{(0)}-\bar{w}\|}{L_{3k}\sqrt{k}\|\nabla F(\bar{w})\|_\infty}\right)$ *rounds of iteration, where* $\theta = \frac{3.24(L_{3k}^2-\mu_{3k}^2)}{L_{3k}^2+\mu_{3k}^2} \in [0, 1)$.

Proof. See § 4.7.6 for a proof of this result. □

Comparing to Theorem 9, the results of Theorem 10 and Theorem 11 are not as strong because the contraction factor has polynomial dependence on the restricted condition number.

4.4 Analysis for Non-Convex Functions

We now turn to study the case when the objective function is non-convex. To analyze the global convergence of general non-convex problems, we follow the convention to use the value $\|\nabla F(w)\|^2$ as a measurement of quality for approximate stationary solutions, keeping in mind that the estimation error criterion for convex problems is not applicable due to the hardness of non-convex problems [74]. For our global analysis, we make two slight modifications of Algorithm 4 to adapt to non-convexity: i) estimate a k-sparse vector $w_j^{(t)}$ such that $\|\nabla P_j(w_j^{(t)}; w^{(t-1)} \mid \eta^{(t)}, \gamma)\| \leq \epsilon$; and ii) update $w^{(t)} = w_1^{(t)}$, that is, we always set $w^{(t)}$ as the local solution of the first (or any other fixed) machine.

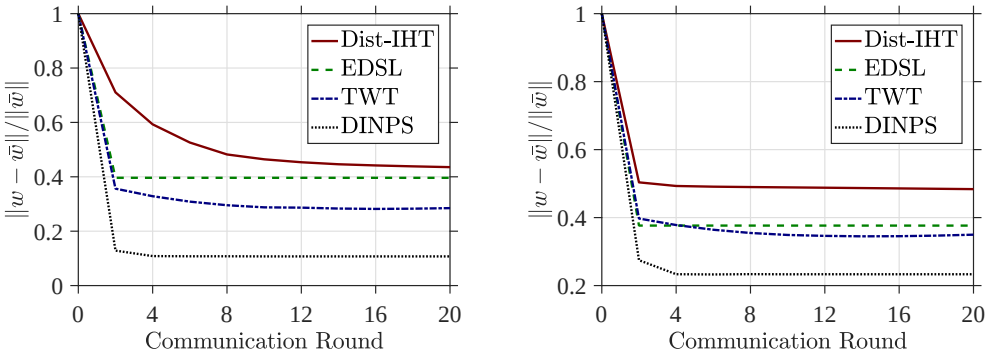
Theorem 12. *Assume that for all* j, $F_j(w)$ *is* L_{2k}*-smooth. Set* $\gamma = (\eta + 2)L_{2k}$. *Then*

$$\min_{1 \leq \tau \leq t} \|\nabla F(w^{(\tau)})\|^2 \leq \left(\frac{8(\eta+3)^2 L_{2k}(F(w^{(0)}) - F(w^*))}{\eta}\right)\frac{1}{t} + \frac{18(\eta+3)^2}{\eta^2}\epsilon^2,$$

where $F(w^*) = \min_{\|w\|_0 \leq k} F(w)$.

Proof. A proof of this result is given in Appendix 4.7.7. □

Remark 11. *The precision barrier* $\mathcal{O}(\epsilon^2)$ *appeared in the above bound is introduced by the local sparse solution whose gradient is generally non-vanishing. In the extreme case of dense learning where the cardinality constraint is inactive, the local solution precision* ϵ *can be arbitrarily small. This leads to a sub-linear convergence result for the original DANE method with non-convex objective functions.*



(a) $\gamma = 2$. (b) $\gamma = 8$.

Figure 4.5.1: Simulation study on sparse linear regression: communication efficiency comparison with varying γ values.

To our knowledge, Theorem 12 is the first convergence result for IHT-style methods with non-convex objective functions.

4.5 Experiments

In this section, we present empirical results of DINPS on a number of synthetic and real-world sparse learning problems, including sparse linear/logistic regression, sparse bilinear regression and training skinny neural networks. The considered algorithms are implemented with C++ and tested on multiple machines with 3.0GHz CPU interconnected by Ethernet. Machine communication interface is implemented by parameter server [58]. In this section, we present empirical results of DINPS on a number of synthetic and real-world sparse learning problems, including sparse linear/logistic regression, sparse bilinear regression and deep neural nets pruning. The considered algorithms are implemented with C++ and tested on multiple machines with 3.0GHz CPU interconnected by Ethernet. The machine communication interface is implemented by parameter server [58].

4.5.1 Sparse linear regression

We first compare DINPS with distributed IHT (Dist-IHT) [72], efficient distributed sparse learning (EDSL) [90] and two-way trauncation (TWT) [76] on simulated sparse

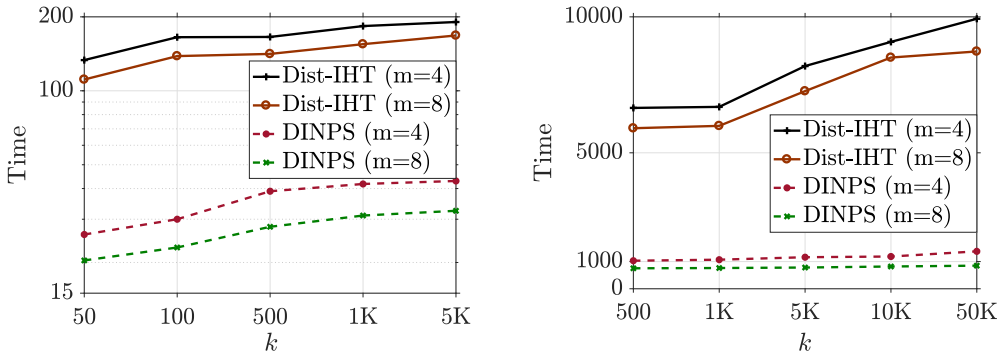
linear regression tasks. Recollect that EDSL and TWT are DANE-type distributed computing methods for solving the Lasso-type estimation problem. A synthetic $N \times p$ design matrix is generated with each data sample x_i drawn from Gaussian distribution $\mathcal{N}(0, \Sigma)$ with $\Sigma_{j,k} = \begin{cases} 1 & \text{if } j = k \\ 1.1^{-\frac{|j-k|}{\gamma}} & \text{otherwise} \end{cases}$. A $\bar{k}$-sparse model parameter $\bar{w} \in \mathbb{R}^p$ is generated with the top $\bar{k}$ entries uniformly randomly valued in interval $(0, 1)$ and all the other entries set to be zero. The response variables $\{y_i\}_{i=1}^N$ are generated by $y_i = \bar{w}^\top x_i + \epsilon_i$ with $\epsilon_i \sim \mathcal{N}(0, 1)$. The convergence is measured by relative estimation error $\|w - \bar{w}\|/\|\bar{w}\|$. The algorithm hyper-parameters are tuned by grid search for optimal performance. We fix the training sample size to be $N = 5 \times 10^3$, $p = 10^4$, $\bar{k} = 100$, the number of machines to be $m = 8$ and vary the value of γ to be 2 and 8.

The convergence curves of the considered algorithms with respect to round of communication are shown in Figure 4.5.1. From this group of results we can see: 1) DINPS, EDSL and TWT converge quickly after a few rounds of master-worker communication, while Dist-IHT method needs thousands more rounds of communication to reach the comparable accuracy of DINPS; 2) When convergence is attained, DINPS outputs more accurate sparse solution than EDSL and TWT, mainly because DINPS directly works on the cardinality-constrained formulation while the EDSL and TWT work a relaxed ℓ_1-formulation which tends to introduce bias in sparse learning. In conclusion, DINPS simultaneously achieves higher communication efficiency and model estimation accuracy than the two state-of-the-art baseline methods.

4.5.2 Sparse ℓ_2-regularized logistic regression

Next we evaluate the performance of DINPS in sparse ℓ_2-regularized binary logistic regression tasks. We compare the training time of DINPS with Dist-IHT on two real-world datasets: rcv1 ($N = 6 \times 10^5$, $p \approx 4.7 \times 10^5$) and kdd2010-algebra ($N \approx 8 \times 10^6$, $p \approx 2 \times 10^7$). For both datasets, the training samples are evenly distributed onto $m = 4$ and 8 machines, and the ℓ_2-regularization strength is set as 10^{-5}.

Figure 4.5.2 shows the running time of algorithms under varying sparsity level $k \in \{0.05, 0.1, 0.5, 1, 5\} \times 10^3$ for rcv1 and $k \in \{0.05, 0.1, 5, 1, 5\} \times 10^4$ for kdd2010-algebra,



(a) rcv1. (b) kdd2010-algebra.

Figure 4.5.2: Sparse logistic regression model training: time cost (in second) comparison and kdd2010-algebra.

	$k = 100$			$k = 1K$		
	EDSL	TWT	DINPS	EDSL	TWT	DINPS
$m = 2$	0.3237	0.2820	**0.2709**	0.2201	0.1823	**0.1551**
$m = 4$	0.3255	0.2828	**0.2717**	0.2225	0.1842	**0.1554**
$m = 8$	0.3298	0.2830	**0.2723**	0.2236	0.1861	**0.1555**

Table 4.5.1: Distributed ℓ_2-sparse logistic regression: model training loss comparison on rcv1, with $k = 100$ and $1K$.

	$k = 100$			$k = 1K$		
	EDSL	TWT	DINPS	EDSL	TWT	DINPS
$m = 2$	0.3959	0.3832	**0.3709**	0.3503	0.3422	**0.3314**
$m = 4$	0.4049	0.3874	**0.3712**	0.3521	0.3460	**0.3347**
$m = 8$	0.4060	0.3902	**0.3723**	0.3526	0.3463	**0.3356**

Table 4.5.2: Distributed ℓ_2-sparse logistic regression: model training loss comparison on kdd-algebra, with $k = 100$ and $1K$.

with machine number $m = 4$ and 8. For any sparsity level, We first run Dist-IHT until it reaches a sub-optimality $|F(w^{(t)}) - F(w^{(t-1)})|/|F(w^{(t)})| \leq 10^{-4}$ or maximum number of iteration, then record the running time of DINPS with different machine number m to reach the training loss level below this. Each model training is repeated 5 times to calculate the average time cost. It can be clearly seen that DINPS is consistently more efficient than Dist-IHT under varying sparsity level and number of machines.

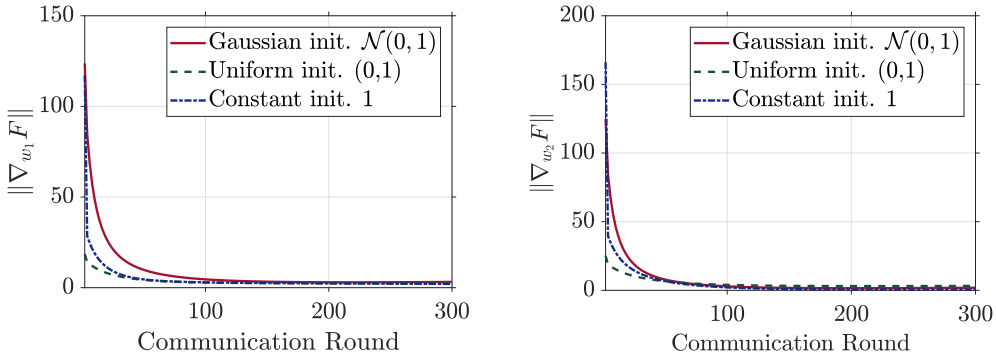
We compare the sparse logistic regression model training accuracy between D-INPS, EDSL and TWT. We run both algorithms to the convergence state evaluated by $|F(w^{(t)}) - F(w^{(t-1)})| - /F(w^{(t)}) \leq 10^{-4}$. The average training loss comparison over 5 splits, given $k = 100$ and $1K$, $m = 2, 4$ and 8 on RCV1 and kdd-algebra datasets, is shown in Table 4.5.1 and 4.5.2. It is observable that DINPS achieves superior model training accuracy than EDSL and TWT.

4.5.3 Sparse bilinear regression

This is a simulated experiment to verify our convergence analysis of DINPS for non-convex functions. Here we consider a non-convex regression problem in which the training samples $\{X_i, y_i\}_{i=1}^N, X_i \in \mathbb{R}^{p_1 \times p_2}, y_i \in \mathbb{R}$ are generated according to a bilinear model $y_i = \bar{w}_1^\top X_i \bar{w}_2 + \varepsilon_i$, where $\bar{w}_1 \in \mathbb{R}^{p_1}$ and $\bar{w}_2 \in \mathbb{R}^{p_2}$ are two sparse vectors whose non-zero entries are uniformly drawn from interval (0,1), $X_i \sim \mathcal{N}(0, I)$ and $\varepsilon_i \sim \mathcal{N}(0, 0.5)$. The objective is to minimize $F(w_1, w_2) := \frac{1}{2N} \sum_{i=1}^N \|y_i - w_1^\top X_i w_2\|^2$ with constraint $\|w_1\|_0 \leq k_1, \|w_2\|_0 \leq k_2$. We test with $p_1 = 40$, $\|\bar{w}_1\|_0 = 20$, $p_2 = 20$, $\|\bar{w}_1\|_0 = 10$, $k_1 = \|\bar{w}_1\|$, $k_2 = \|\bar{w}_2\|$ and $N = 10^4$.

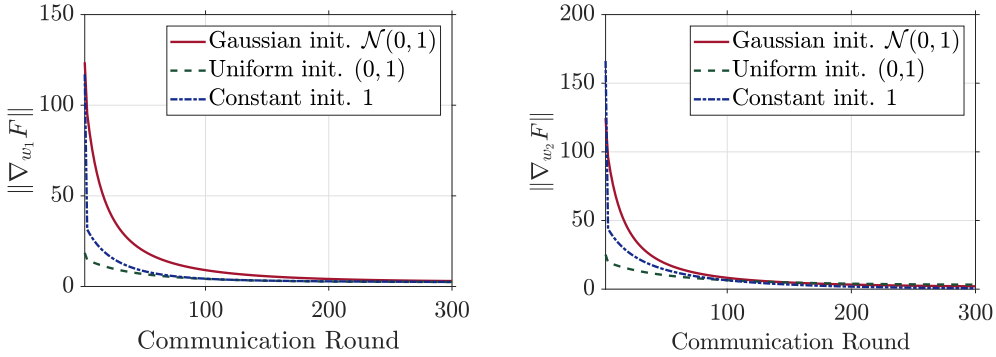
We study the global convergence of DINPS under three different initialization schemes: (1) Gaussian random initialization $\mathcal{N}(0, 1)$, (2) uniform random initialization $(0, 1)$, and (3) constant initialization 1. For solving the local ℓ_0-minimization problem (4.2.1), we alternately optimize w_1 and w_2 using IHT. The convergence curves of $\|\nabla_{w_1} F\|$ and $\|\nabla_{w_2} F\|$ with respect to round of communication are respectively plot in Figure 4.5.3 for machine number $m = 4$ and 8. From this group of curves we can see that the ℓ_2-norm of parameter gradient converges quickly to a stable state after sufficient communication among machines. This is consistent with our global convergence results

stated in Theorem 12.



(a) Convergence of $\|\nabla_{w_1} F\|$, $m = 4$.

(b) Convergence of $\|\nabla_{w_2} F\|$, $m = 4$.



(c) Convergence of $\|\nabla_{w_1} F\|$, $m = 8$.

(d) Convergence of $\|\nabla_{w_2} F\|$, $m = 8$.

Figure 4.5.3: Distributed sparse bilinear regression: global convergence of gradients $\|\nabla_{w_j} F\|$, $j = 1, 2$, with respect to communication round under different initialization. The number of machine is $m = 4$ and 8.

4.5.4 Sparse deep neural networks

Finally, we apply DINPS to distributed learning of neural networks under layer-wise sparsity constraint over neuron connections. Such skinny neural networks have recently been shown to be able to efficiently compress model size without sacrificing accuracy such as in image classification problems [31, 42, 91]. In our experiment, we test with two LeNet structures, denoted by LeNet-1 (3 fully-connected layers) and LeNet-2 (2 convolutional layers and 2 fully-connected layers) [53], on mnist digit dataset. For both

	LeNet-1		LeNet-2	
	FedAvg	DINPS	FedAvg	DINPS
$m = 2$	1.49	1.43	0.71	0.63
$m = 4$	1.51	1.44	0.70	0.68
$m = 8$	1.55	1.46	0.73	0.69
model size	266K	**53K**	430K	**94K**

Table 4.5.3: Distributed skinny neural networks learning: validation set classification error (in %) and model size.

networks, we prune 50% of the parameters in convolutional layers and 80% of the parameters in fully connected layers. To initialize DINPS, we train a dense network by applying the Federated-Averaging (FedAvg) method [68] on the given data partition. The algorithm proposed in [42] is adopted for local training. We compare the sparse network output by DINPS against the dense network by FedAvg in prediction accuracy and compression ratio. The experiment is replicated 5 times with average results reported. The considered algorithms are implemented on Apache MXNet platform and tested on a cluster with Nvidia K80 GPUs.

Table 4.5.3 lists the experimental results on $m = 2, 4$ and 8 machines. It can be observed from these results that the sparse networks trained by the DINPS have quite competitive or even superior prediction accuracy to the dense ones obtained by FedAvg, while the former has $\sim 80\%$ fewer model parameters than the latter. This set of empirical results confirm that DINPS is an accurate and communication-efficient method for distributed sparse neural networks learning.

4.6 Conclusion

We proposed DINPS as a Newton-type communication-efficient distributed computing method for ℓ_0-constrained sparse learning. The algorithm iterates between: (1) solving an inexact variance reduced ℓ_0-constrained minimization problem on each local worker machine; and (2) parameter and gradient aggregation on master machine. For

generic convex loss functions, DINPS has been shown to exhibit logarithmical communication complexity and lower sample complexity than prior methods. For non-convex loss functions, we have established global sublinear convergence for DINPS under proper conditions. Extensive empirical results confirmed our theoretical predictions and demonstrated the advantages of DINPS over the state-of-the-art methods.

4.7 Appendix

4.7.1 Technical Lemmas

The following lemma shows that the estimation error of the truncated average of estimators is well upper bounded by the average error of estimators.

Lemma 8. *Let $\bar{w}$ be $\bar{k}$-sparse vector. For a set of k-sparse vectors $\{w_j\}_{j=1}^m$ with $k \geq \bar{k}$, it holds that*

$$\left\| \mathrm{H}_k\left(\frac{1}{m}\sum_{j=1}^m w_j\right) - \bar{w} \right\| \leq \frac{1.62}{m}\sum_{j=1}^m \|w_j - \bar{w}\|.$$

Moreover, if $k > \bar{k}$, then

$$\left\| \mathrm{H}_k\left(\frac{1}{m}\sum_{j=1}^m w_j\right) - \bar{w} \right\| \leq \frac{1}{m}\sqrt{1 + 2\sqrt{\frac{\bar{k}}{k-\bar{k}}}}\sum_{j=1}^m \|w_j - \bar{w}\|.$$

Proof. The first claim follows directly from [87, Theorem 1] and triangle inequality. The second claim is due to the result in [59, Lemma 3.3]. □

The following lemma summarizes some important properties of restricted strong smoothness/convexity, restricted Lipschitz gradient and restricted Lipschitz Hessian.

Lemma 9. *Assume that $f(w)$ is differentiable. Then $f(w)$ has α_s-RLG if and only if for any w, w' with $\|w - w'\|_0 \leq s$,*

$$\left| f(w) - f(w') - \langle \nabla_S f(w'), w - w' \rangle \right| \leq \frac{\alpha_s}{2}\|w - w'\|^2,$$

where $S = supp(w) \cup supp(w')$. Moreover, assume that $f(w)$ has β_s-RLH. Then

$$\left\| \nabla_S f(w) - \nabla_S f(w') - \nabla^2_{SS} f(w')(w - w') \right\| \leq \frac{\beta_s}{2}\|w - w'\|^2.$$

Proof. **The "⇒" direction of the first part:** Assume that

$$\left|f(w) - f(w') - \langle \nabla_S f(w'), w - w' \rangle\right| \leq \frac{\alpha_s}{2}\|w - w'\|^2$$

. We first claim and prove the following inequality which is key to our analysis:

$$f(w') \leq f(w) + \frac{\alpha_s}{4}\|w - w'\|^2 + \frac{1}{2}\langle \nabla_S f(w') + \nabla_S f(w), w' - w \rangle - \frac{1}{4\alpha_s}\left\|\nabla_S f(w) - \nabla_S f(w')\right\|^2. \tag{4.7.1}$$

Let $g(w) = f(w) + \frac{\alpha_s}{2}\|w\|^2$. The assumption on $f(w)$ implies

$$0 \leq g(w) - g(w') - \langle \nabla_S g(w'), w - w' \rangle \leq \alpha_s \|w - w'\|^2. \tag{4.7.2}$$

That is, $g(w)$ is 0-strongly-convex and $2\alpha_s$-smooth. Let us fix w' and allow w to vary under the constraint $\|w - w'\|_0 \leq s$. Consider the function $\phi(w) = g(w) - \langle \nabla_S g(w'), w \rangle$. From (4.7.2) we can verify w' is an optimal point of $\phi(w)$ over $\{w : \|w - w'\|_0 \leq s\}$. Therefore, in view of (4.7.2), we get that for any w satisfying $\|w - w'\|_0 \leq s$

$$\phi(w') \leq \phi\left(w - \frac{1}{2\alpha_s}\nabla_S \phi(w)\right) \leq \phi(w) - \frac{1}{4\alpha_s}\|\nabla_S \phi(w)\|^2.$$

By substituting the expression of $\phi(w)$ into the above we have

$$g(w') \leq g(w) + \langle \nabla_S g(w'), w' - w \rangle - \frac{1}{4\alpha_s}\|\nabla_S g(w) - \nabla_S g(w')\|^2.$$

By substituting the expression of $g(w)$ into the above and with proper elementary calculation we further get

$$\begin{aligned} f(w') &\leq f(w) + \frac{\alpha_s}{2}\|w\|^2 - \frac{\alpha_s}{2}\|w'\|^2 + \langle \nabla_S f(w') + \alpha_s w', w' - w \rangle \\ &\quad - \frac{1}{4\alpha_s}\left\|\nabla_S f(w) - \nabla_S f(w') + \alpha_s(w - w')\right\|^2 \\ &= f(w) + \frac{\alpha_s}{4}\|w - w'\|^2 + \frac{1}{2}\langle \nabla_S f(w') + \nabla_S f(w), w' - w \rangle - \frac{1}{4\alpha_s}\left\|\nabla_S f(w) - \nabla_S f(w')\right\|^2. \end{aligned}$$

This is exactly the desired inequality in (4.7.1). By adding two copies of the inequality (4.7.1) with w and w' interchanged we arrive at

$$\left\|\nabla_S f(w) - \nabla_S f(w')\right\| \leq \alpha_s \|w - w'\|.$$

This indicates that $f(w)$ has α_s-RLG.

The "⇐" direction of the first part: Assume that $f(w)$ has α_s-RLG. From the Taylor's theorem we know that

$$f(w) - f(w') = \int_0^1 \langle \nabla_S f(w + t(w' - w)), w - w' \rangle dt.$$

Therefore

$$\begin{aligned}
&\left| f(w) - f(w') - \langle \nabla_S f(w'), w - w' \rangle \right| \\
=&\left| \int_0^1 \langle \nabla_S f(w') - \nabla_S f(w + t(w' - w)), w - w' \rangle dt \right| \\
\leq&\|w - w'\| \int_0^1 \left\| \nabla_S f(w) - \nabla_S f(w + t(w' - w)) \right\| dt \\
\leq&\alpha_s \|w - w'\|^2 \int_0^1 t dt = \frac{\alpha_s}{2} \|w - w'\|^2.
\end{aligned}$$

This completes the proof of the first part.

To prove the second part, we invoke the Taylor's theorem for vector-valued functions to get

$$\nabla_S f(w) - \nabla_S f(w') = \int_0^1 \nabla^2_{SS} f(w + t(w' - w))(w - w') dt.$$

Therefore

$$\begin{aligned}
&\left\| \nabla_S f(w) - \nabla_S f(w') - \nabla^2_{SS} f(w')(w - w') \right\| \\
=&\left\| \int_0^1 [\nabla^2_{SS} f(w') - \nabla^2_{SS} f(w + t(w' - w))](w - w') dt \right\| \\
\leq&\|w - w'\| \int_0^1 \left\| \nabla^2_{SS} f(w) - \nabla^2_{SS} f(w + t(w' - w)) \right\| dt \\
\leq&\beta_s \|w - w'\|^2 \int_0^1 t dt = \frac{\beta_s}{2} \|w - w'\|^2.
\end{aligned}$$

This proves the claim of the second part. □

The following lemma gives a necessary condition on spare minimizer.

Lemma 10. *If $f(w)$ is L_{2k}-smooth, then the following inequality holds for the global minimizer $w^\star = \arg\min_{\|w\|_0 \leq k} f(w)$:*

$$w^\star_{\min} \geq \frac{\|\nabla f(w^\star)\|_\infty}{L_{2k}}.$$

where $w^\star_{\min}$ denotes the minimum nonzero value in magnitude of $w^\star$.

Proof. Assume otherwise that $\vartheta^\star := \frac{L_{2k} w^\star_{\min}}{\|\nabla f(w^\star)\|_\infty} < 1$. Let us consider $\tilde{w}^\star = w^\star - \eta \nabla f(w^\star)$ with any $\eta \in (\vartheta^\star / L_{2k}, 1/L_{2k})$. Since f is L_{2k}-smooth, it follows that

$$\begin{aligned} f(\tilde{w}^\star_k) - f(w^\star) \leq & \langle \nabla f(w^\star_k), \tilde{w}^\star_k - w^\star \rangle + \frac{L_{2k}}{2} \|\tilde{w}^\star_k - w^\star\|^2 \\ \overset{\xi_1}{\leq} & -\frac{1}{2\eta} \|\tilde{w}^\star_k - w^\star\|^2 + \frac{L_{2k}}{2} \|\tilde{w}^\star_k - w^\star\|^2 \\ = & -\frac{1 - \eta L_{2k}}{2\eta} \|\tilde{w}^\star_k - w^\star\|^2, \end{aligned}$$

where ξ_1 follows from the fact that $\tilde{w}^\star_k$ is the best k-support approximation to $\tilde{w}^\star$ such that

$$\|\tilde{w}^\star_k - \tilde{w}^\star\|^2 = \|\tilde{w}^\star_k - w^\star + \eta \nabla f(w^\star)\|^2 \leq \|w^\star - w^\star + \eta \nabla f(w^\star)\|^2 = \|\eta \nabla f(w^\star)\|^2,$$

which implies $2\eta \langle \nabla f(w^\star), \tilde{w}^\star_k - w^\star \rangle \leq -\|\tilde{w}^\star_k - w^\star\|^2$. Since $\eta \in (\vartheta^\star / L_{2k}, 1/L_{2k})$ and $w^\star_{\min} = \frac{\vartheta^\star \|\nabla f(w^\star)\|_\infty}{L_{2k}} < \eta \|\nabla f(w^\star)\|_\infty$, we have $\tilde{w}^\star_k \neq w^\star$ and thus it follows from the above inequality that $f(\tilde{w}^\star_k) < f(w^\star)$. This contradicts the optimality of $w^\star$. □

The following lemma is key to our analysis.

Lemma 11. *Let $\bar{w}$ be a $\bar{k}$-sparse target vector with $\bar{k} \leq k$. Assume that each component $F_j(w)$ is μ_{3k}-strongly-convex and $\eta F(w) - F_j(w) - \frac{\gamma}{2}\|w\|^2$ has α_{3k}-RLG. Then*

$$\|w^{(t)} - \bar{w}\| \leq \frac{3.24 \alpha_{3k}}{\gamma + \mu_{3k}} \|w^{(t-1)} - \bar{w}\| + \frac{5.62 \eta \sqrt{k}}{\gamma + \mu_{3k}} \|\nabla F(\bar{w})\|_\infty + 2.3 \sqrt{\frac{\epsilon}{\gamma + \mu_{3k}}}.$$

Moreover, assume that each $F_j(w)$ has β_{3k}-RLH. Let $\bar{H}_j = \nabla^2 F_j(\bar{w})$ and $\bar{H} = \frac{1}{m}\sum_{j=1}^m \bar{H}_j$. Then

$$\begin{aligned} \|w^{(t)} - \bar{w}\| \leq & \frac{3.24(\gamma + \max_j \|\bar{H}_j - \eta \bar{H}\|)}{\gamma + \mu_{3k}} \|w^{(t-1)} - \bar{w}\| + \frac{1.62(1+\eta)\beta_{3k}}{\gamma + \mu_{3k}} \|w^{(t-1)} - \bar{w}\|^2 \\ & + \frac{5.62 \eta \sqrt{k}}{\gamma + \mu_{3k}} \|\nabla F(\bar{w})\|_\infty + 2.3 \sqrt{\frac{\epsilon}{\gamma + \mu_{3k}}}. \end{aligned}$$

Proof. For any $j \in [m]$, since $F_j(w)$ is μ_{3k}-strongly-convex, we have $P_j(w; w^{(t-1)} \mid \eta, \gamma)$ is $(\gamma + \mu_{3k})$-strongly-convex. Let $S^{(t)}_j = \mathrm{supp}(w^{(t)}_j)$, $S^{(t-1)} = \mathrm{supp}(w^{(t-1)})$ and $\bar{S} = \mathrm{supp}(\bar{w})$. Consider $S = S^{(t)}_j \cup S^{(t-1)} \cup \bar{S}$. Then

$$\begin{aligned} & P_j(w^{(t)}_j; w^{(t-1)} \mid \eta, \gamma) \\ \geq & P_j(\bar{w}; w^{(t-1)} \mid \eta, \gamma) + \langle \nabla P_j(\bar{w}; w^{(t-1)} \mid \eta, \gamma), w^{(t)}_j - \bar{w} \rangle + \frac{\gamma + \mu_{3k}}{2} \|w^{(t)}_j - \bar{w}\|^2 \\ = & P_j(\bar{w}; w^{(t-1)} \mid \eta, \gamma) + \langle \nabla_S P_j(\bar{w}; w^{(t-1)} \mid \eta, \gamma), w^{(t)}_j - \bar{w} \rangle + \frac{\gamma + \mu_{3k}}{2} \|w^{(t)}_j - \bar{w}\|^2 \\ \overset{\xi_1}{\geq} & P_j(w^{(t)}; w^{(t-1)} \mid \eta, \gamma) - \epsilon - \|\nabla_S P_j(\bar{w}; w^{(t-1)} \mid \eta, \gamma)\| \|w^{(t)}_j - \bar{w}\| + \frac{\gamma + \mu_{3k}}{2} \|w^{(t)}_j - \bar{w}\|^2, \end{aligned}$$

where "ξ_1" follows from the definition of $w^{(t)}$ as an approximate k-sparse minimizer of $P_j(w; w^{(t-1)} \mid \eta, \gamma)$ up to precision ϵ. By rearranging both sides of the above inequality with proper elementary calculation we get

$$
\begin{aligned}
&\|w_j^{(t)} - \bar{w}\| \\
\leq &\frac{2}{\gamma + \mu_{3k}} \|\nabla_S P_j(\bar{w}; w^{(t-1)} \mid \eta, \gamma)\| + \sqrt{\frac{2\epsilon}{\gamma + \mu_{3k}}} \\
= &\frac{2}{\gamma + \mu_{3k}} \|\eta \nabla_S F(w^{(t-1)}) - \nabla_S F_j(w^{(t-1)}) + \gamma(\bar{w} - w^{(t-1)}) + \nabla_S F_j(\bar{w})\| + \sqrt{\frac{2\epsilon}{\gamma + \mu_{3k}}} \\
= &\frac{2}{\gamma + \mu_{3k}} \|\eta \nabla_S F(w^{(t-1)}) - \eta \nabla_S F(\bar{w}) - (\nabla_S F_j(w^{(t-1)}) - \nabla_S F_j(\bar{w})) + \gamma(\bar{w} - w^{(t-1)}) \\
&+ \eta \nabla_S F(\bar{w})\| + \sqrt{\frac{2\epsilon}{\gamma + \mu_{3k}}} \\
\overset{\xi_1}{\leq} &\frac{2}{\gamma + \mu_{3k}} \left\| \left(\eta \nabla_S F(w^{(t-1)}) - \nabla_S F_j(w^{(t-1)}) - \gamma w^{(t-1)} \right) - (\eta \nabla_S F(\bar{w}) - \nabla_S F_j(\bar{w}) - \gamma \bar{w}) \right\| \\
&+ \frac{2\eta}{\gamma + \mu_{3k}} \|\nabla_S F(\bar{w})\| + \sqrt{\frac{2\epsilon}{\gamma + \mu_{3k}}} \\
\leq &\frac{2\alpha_{3k}}{\gamma + \mu_{3k}} \|w^{(t-1)} - \bar{w}\| + \frac{2\eta\sqrt{3k}}{\gamma + \mu_{3k}} \|\nabla F(\bar{w})\|_\infty + \sqrt{\frac{2\epsilon}{\gamma + \mu_{3k}}},
\end{aligned}
$$

where ζ_1 is according to the assumption that $\eta F(w) - F_j(w) - \frac{\gamma}{2}\|w\|^2$ has α_{3k}-RLG. From the definition of $w^{(t)} = \mathrm{H}_k\left(\frac{1}{m}\sum_{j=1}^m w_j^{(t)}\right)$ and by applying the first claim in Lemma 8 we have

$$
\begin{aligned}
\|w^{(t)} - \bar{w}\| &= 1.62 \left\| \frac{1}{m} \sum_{j=1}^m w_j^{(t)} - \bar{w} \right\| \\
&\leq \frac{1.62}{m} \sum_{j=1}^m \left\| w_j^{(t)} - \bar{w} \right\| \leq \frac{3.24\alpha_{3k}}{\gamma + \mu_{3k}} \|w^{(t-1)} - \bar{w}\| + \frac{5.62\eta\sqrt{k}}{\gamma + \mu_{3k}} \|\nabla F(\bar{w})\|_\infty + 2.3\sqrt{\frac{\epsilon}{\gamma + \mu_{3k}}}.
\end{aligned}
$$

This shows the validity of the first part.

Now we prove the second part. Similar to the above argument, we have

$$\begin{aligned}
\|w_j^{(t)} - \bar{w}\| \leq & \frac{2}{\gamma + \mu_{3k}} \|\nabla_S P_j(\bar{w}; w^{(t-1)} \mid \eta, \gamma)\| + \sqrt{\frac{2\epsilon}{\gamma + \mu_{3k}}} \\
\leq & \frac{2}{\gamma + \mu_{3k}} \left\| \gamma(w^{(t-1)} - \bar{w}) + \eta \nabla_S F(w^{(t-1)}) - \eta \nabla_S F(\bar{w}) - (\nabla_S F_j(w^{(t-1)}) - \nabla_S F_j(\bar{w})) \right\| \\
& + \frac{2\eta}{\gamma + \mu_{3k}} \|\nabla_S F(\bar{w})\| + \sqrt{\frac{2\epsilon}{\gamma + \mu_{3k}}} \\
\leq & \frac{2}{\gamma + \mu_{3k}} \left\| \gamma(w^{(t-1)} - \bar{w}) + \eta \nabla^2_{SS} F(\bar{w})(w^{(t-1)} - \bar{w}) - \nabla^2_{SS} F_j(\bar{w})(w^{(t-1)} - \bar{w}) \right\| \\
& + \frac{2\eta}{\gamma + \mu_{3k}} \|\nabla_S F(\bar{w})\| + \frac{2\eta}{\gamma + \mu_{3k}} \left\| \nabla_S F(w^{(t-1)}) - \nabla_S F(\bar{w}) - \nabla^2_{SS} F(\bar{w})(w^{(t-1)} - \bar{w}) \right\| \\
& + \frac{2}{\gamma + \mu_{3k}} \left\| \nabla_S F_j(w^{(t-1)}) - \nabla_S F_j(\bar{w}) - \nabla^2_{SS} F_j(\bar{w})(w^{(t-1)} - \bar{w}) \right\| + \sqrt{\frac{2\epsilon}{\gamma + \mu_{3k}}} \\
\leq & \frac{2}{\gamma + \mu_{3k}} \left(\gamma + \|\eta \nabla^2_{SS} F(\bar{w}) - \nabla^2_{SS} F_j(\bar{w})\| \right) \|w^{(t-1)} - \bar{w}\| + \frac{2\eta}{\gamma + \mu_{3k}} \|\nabla_S F(\bar{w})\| \\
& + \frac{(1+\eta)\beta_{3k}}{\gamma + \mu_{3k}} \|w^{(t-1)} - \bar{w}\|^2 + \sqrt{\frac{2\epsilon}{\gamma + \mu_{3k}}} \\
\leq & \frac{2(\gamma + \max_{j'} \|\bar{H}_{j'} - \eta \bar{H}\|)}{\gamma + \mu_{3k}} \|w^{(t-1)} - \bar{w}\| + \frac{(1+\eta)\beta_{3k}}{\gamma + \mu_{3k}} \|w^{(t-1)} - \bar{w}\|^2 \\
& + \frac{2\eta\sqrt{3k}}{\gamma + \mu_{3k}} \|\nabla F(\bar{w})\|_\infty + \sqrt{\frac{2\epsilon}{\gamma + \mu_{3k}}},
\end{aligned}$$

Again, from the definition of $w^{(t)}$ and by applying the first claim in Lemma 8 we have

$$\begin{aligned}
& \left\| w^{(t)} - \bar{w} \right\| \\
\leq & \frac{3.24(\gamma + \max_j \|\bar{H}_j - \eta \bar{H}\|)}{\gamma + \mu_{3k}} \|w^{(t-1)} - \bar{w}\| + \frac{1.62(1+\eta)\beta_{3k}}{\gamma + \mu_{3k}} \|w^{(t-1)} - \bar{w}\|^2 \\
& + \frac{5.62\eta\sqrt{k}}{\gamma + \mu_{3k}} \|\nabla F(\bar{w})\|_\infty + 2.3\sqrt{\frac{\epsilon}{\gamma + \mu_{3k}}}.
\end{aligned}$$

This proves the second part. □

4.7.2 Proof of Theorem 8 and Corollary 5

Proof of Theorem 8. Since the local objectives F_j are quadratic, we can simply set $\beta_s = 0$ for all cardinality s. By assumption $F_j(w)$ is μ_{3k}-strongly-convex. Then by applying the second part of Lemma 11 with $\beta_{3k} = 0$, $\gamma = 0$ and $\epsilon \leq \frac{k\eta^2 \|\nabla F(\bar{w})\|_\infty^2}{5.29\mu_{3k}}$ we get

$$\|w^{(t)} - \bar{w}\| \leq \frac{3.24 \max_j \|\bar{H}_j - \eta \bar{H}\|}{\mu_{3k}} \|w^{(t-1)} - \bar{w}\| + \frac{6.62\eta\sqrt{k}}{\mu_{3k}} \|\nabla F(\bar{w})\|_\infty.$$

It can be checked that the factor $\frac{3.24\max_j \|\bar{H}_j - \eta\bar{H}\|}{\mu_{3k}} \leq \theta < 1$. By recursively applying the above inequality we arrive at

$$\|w^{(t)} - \bar{w}\| \leq \theta^t \|w^{(0)} - \bar{w}\| + \frac{6.62\eta\sqrt{k}\|\nabla F(\bar{w})\|_\infty}{(1-\theta)\mu_{3k}}.$$

Based on the inequality $1 - x \leq \exp(-x)$ we need

$$t \geq \frac{1}{1-\theta} \log \frac{(1-\theta)\mu_{3k}\|w^{(0)} - \bar{w}\|}{\eta\sqrt{k}\|\nabla F(\bar{w})\|_\infty}$$

steps of iteration to achieve the precision of $\|w^{(t)} - \bar{w}\| \leq \frac{7.62\eta\sqrt{k}\|\nabla F(\bar{w})\|_\infty}{(1-\theta)\mu_{3k}}$. This proves the desired complexity bound. □

Proof of Corollary 5. From the definition of θ and Lemma 7 we get $\max_j \|H_j - H\| \leq \frac{\theta\mu_{3k}}{3.24}$ holds with probability at least $1 - \delta$. Since $n > \frac{336L^2\log(mp/\delta)}{\mu_{3k}^2}$, we have $\theta \in (0, 1)$. The desired bound is then directly implied by Theorem 8. □

Implications for distributed sparse linear regression. Given a $\bar{k}$-sparse parameter vector $\bar{w}$, assume the samples are generated according to the linear model $y = \bar{w}^\top x + \varepsilon$ where ε is a zero-mean Gaussian random noise variable with parameter σ. Assume the data samples $\{D_j = \{x_{ji}, y_{ji}\}_{i=1}^n\}_{j=1}^m$ are distributed over m machines and let $F_j(w) = \frac{1}{2n}\sum_{i=1}^n \|y_{ji} - w^\top x_{ji}\|^2$, $j \in [m]$ be the least square loss over D_j and $F(w) = \frac{1}{m}\sum_{j=1}^m F_j(w)$ be the average of local loss. This example belongs to the quadratic case for which the performance of DINPS is analyzed in §4.3.2. Suppose x_{ji} are drawn from Gaussian distribution with covariance Σ. Then it holds with high probability that $F_j(w)$ has restricted strong-convexity constant $\mu_{3k} \geq \lambda_{\min}(\Sigma) - \mathcal{O}(k\log p/n)$ and smoothness constant $L \leq \max_{j,i}\|x_{ji}\|$; and $\|\nabla F(\bar{w})\|_\infty = \mathcal{O}\left(\sigma\sqrt{\log p/(mn)}\right)$ and $\|\nabla F_j(\bar{w})\|_\infty = \mathcal{O}\left(\sigma\sqrt{\log p/n}\right)$. Consider the local initialization strategy of $w^{(0)} \approx \arg\min_{\|w\|_0 \leq k} F_1(w)$. Then according to the bound in (4.3.1), if the sample size $n = \mathcal{O}\left(\frac{L^2\log(mp)}{\mu_{3k}^2}\right)$ is sufficiently large, DINPS needs $\mathcal{O}(\log m)$ rounds of iteration/communication to reach the statistical error level $\mathcal{O}\left(\sigma\sqrt{k\log p/(mn)}\right)$.

4.7.3 Proof of Theorem 9 and Corollary 6

Proof of Theorem 9. We first claim that $\|w^{(t)} - \bar{w}\| \leq \frac{\mu_{3k}\theta}{3.24(1+\eta)\beta_{3k}}$ holds for all $t \geq 0$. This can be shown by induction. Based on the theorem assumptions the claim holds

for $t = 0$. Now suppose that $\|w^{(t-1)} - \bar{w}\| \leq \frac{\mu_{3k}\theta}{3.24(1+\eta)\beta_{3k}}$ for some $t \geq 1$. Since $\gamma = 0$, according to Lemma 11 we have

$$\begin{aligned}
\|w^{(t)} - \bar{w}\| &\leq \frac{3.24 \max_j \|\bar{H}_j - \eta\bar{H}\|}{\mu_{3k}}\|w^{(t-1)} - \bar{w}\| + \frac{1.62(1+\eta)\beta_{3k}}{\mu_{3k}}\|w^{(t-1)} - \bar{w}\|^2 \\
&\quad + \frac{5.62\eta\sqrt{k}}{\mu_{3k}}\|\nabla F(\bar{w})\|_\infty + 2.3\sqrt{\frac{\epsilon}{\mu_{3k}}} \\
&\overset{\zeta_1}{\leq} \frac{\theta}{2}\|w^{(t-1)} - \bar{w}\| + \frac{1.62(1+\eta)\beta_{3k}}{\mu_{3k}}\|w^{(t-1)} - \bar{w}\|^2 + \frac{6.62\eta\sqrt{k}}{\mu_{3k}}\|\nabla F(\bar{w})\|_\infty \\
&\leq \theta\|w^{(t-1)} - \bar{w}\| + \frac{6.62\eta\sqrt{k}}{\mu_{3k}}\|\nabla F(\bar{w})\|_\infty \\
&\overset{\zeta_2}{\leq} \frac{\mu_{3k}\theta^2}{3.24(1+\eta)\beta_{3k}} + \frac{\mu_{3k}\theta(1-\theta)}{3.24(1+\eta)\beta_{3k}} = \frac{\mu_{3k}\theta}{3.24(1+\eta)\beta_{3k}},
\end{aligned}$$

where "ζ_1" follows from the assumptions on $\max_j \|\bar{H}_j - \eta\bar{H}\|$ and ϵ, and "ζ_2" follows from the condition of $\|\nabla F(\bar{w})\|_\infty \leq \frac{(1-\theta)\theta\mu_{3k}^2}{21.45\eta(1+\eta)\beta_{3k}\sqrt{k}}$. Thus by induction $\|w^{(t)} - \bar{w}\| \leq \frac{\mu_{3k}\theta}{3.24(1+\eta)\beta_{3k}}$ holds for all $t \geq 1$. Then it follows from the inequality below "ζ_1" of the above we get that for all $t \geq 0$,

$$\|w^{(t)} - \bar{w}\| \leq \theta\|w^{(t-1)} - \bar{w}\| + \frac{6.62\eta\sqrt{k}}{\mu_{3k}}\|\nabla F(\bar{w})\|_\infty.$$

By recursively applying the above inequality we get

$$\|w^{(t)} - \bar{w}\| \leq \theta^t\|w^{(0)} - \bar{w}\| + \frac{6.62\eta\sqrt{k}}{(1-\theta)\mu_{3k}}\|\nabla F(\bar{w})\|_\infty.$$

Based on the inequality $1 - x \leq \exp(-x)$ we need

$$t \geq \frac{1}{1-\theta}\log\left(\frac{(1-\theta)\mu_{3k}\|w^{(0)} - \bar{w}\|}{\eta\sqrt{k}\|\nabla F(\bar{w})\|_\infty}\right)$$

steps of iteration to achieve $\|w^{(t)} - \bar{w}\| \leq \frac{7.62\eta\sqrt{k}\|\nabla F(\bar{w})\|_\infty}{(1-\theta)\mu_{3k}}$. This proves the desired complexity bound. □

Based on the results in Theorem 9 we can easily prove Corollary 6.

Proof of Corollary 6. From the definition of θ and Lemma 7 we get that $\max_j \|H_j - H\| \leq \frac{\theta\mu_{3k}}{6.48}$ holds with probability at least $1 - \delta$. Since $n > \frac{1344L^2\log(mp/\delta)}{\mu_{3k}^2}$, we have $\theta \in (0, 1)$. By invoking Theorem 9 we get the desired result. □

4.7.4 Proof of Proposition 2

Proof. Consider an index set S with cardinality $|S| \leq s$ and all w, w' with $\text{supp}(w) \cup \text{supp}(w') \subseteq S$. Since $\sigma(z)$ is Lipschitz continuous with constant 1, we have that

$$\begin{aligned}&|\sigma(2y_i w^\top x_i) - \sigma(2y_i w'^\top x_i)|\\ \leq&|2(w-w')^\top y_i x_i| \leq 2\|[x_i]_S\|\|w-w'\| \leq 2r_s\|w-w'\|.\end{aligned}$$

Using this above inequality and the fact that $\sigma(z) \leq 1$ we obtain

$$\begin{aligned}&|\sigma(2v_i w^\top u_i)(1-\sigma(2v_i w^\top u_i)) - \sigma(2v_i w'^\top u_i)(1-\sigma(2v_i w'^\top u_i))|\\ \leq&|\sigma(2v_i w^\top u_i) - \sigma(2v_i w'^\top u_i)|(1+\sigma(2v_i w^\top u_i) + \sigma(2v_i w'^\top u_i))\\ \leq&3|\sigma(2v_i w^\top u_i) - \sigma(2v_i w'^\top u_i)| \leq 6r_s\|w-w'\|.\end{aligned}$$

This yields $\|\Lambda(w) - \Lambda(w')\| \leq 24r_s\|w-w'\|$. Therefore,

$$\begin{aligned}\left\|\nabla^2_{SS} f(w) - \nabla^2_{SS} f(w')\right\| \leq&\frac{1}{n}\|X^n_S\|^2\|\Lambda(w) - \Lambda(w')\|\\ \overset{\zeta_1}{\leq}&24r_s\left\|\frac{1}{n}X^n_S(X^n_S)^\top\right\|\|w-w'\|\\ \leq&24r_s\rho^{\max}_s(\Sigma_n)\|w-w'\|,\end{aligned}$$

where the "ζ_1" follows from the standard matrix norm equality $\|A\|^2 = \|AA^\top\|$. This proves the desired result. □

4.7.5 Proof of Theorem 10

Proof. Recall the definition of $P_j(w; w^{(t-1)} \mid \eta, \gamma)$:

$$P_j(w; w^{(t-1)} \mid \eta, \gamma) = \langle \eta\nabla F(w^{(t-1)}) - \nabla F_j(w^{(t-1)}), w - w^{(t-1)}\rangle + \frac{\gamma}{2}\|w-w^{(t-1)}\|^2 + F_j(w).$$

For any sparsity level s, $P_j(w; w^{(t-1)} \mid \eta, \gamma)$ is $(\gamma + L_s)$-smooth and $(\gamma + \mu_s)$-strongly-convex. Let $S = S^{(t)}_j \cup \bar{S}$. Since $w^{(t)}_j = \arg\min_{\|w\|_0 \leq k} P_j(w; w^{(t-1)} \mid \eta, \gamma)$ (note that $\epsilon = 0$ as assumed), from Lemma 10 we have

$$w^{(t)}_{j,\min} \geq \frac{\|\nabla P_j(w^{(t)}_j; w^{(t-1)} \mid \eta, \gamma)\|_\infty}{\gamma + L_{2k}}. \tag{B.1}$$

Then based on the strong convexity of $P_j(w; w^{(t-1)} \mid \eta, \gamma)$ we can derive

$$\begin{aligned}
\|w_j^{(t)} - \bar{w}\| \leq & \frac{\|\nabla_S P_j(w_j^{(t)}; w^{(t-1)} \mid \eta, \gamma) - \nabla_S P_j(\bar{w}; w^{(t-1)} \mid \eta, \gamma)\|}{\gamma + \mu_{2k}} \\
\overset{\xi_1}{\leq} & \frac{\|\nabla_{\bar{S} \setminus S_j^{(t)}} P_j(w_j^{(t)}; w^{(t-1)} \mid \eta, \gamma)\| + \|\nabla_S P_j(\bar{w}; w^{(t-1)} \mid \eta, \gamma)\|}{\gamma + \mu_{2k}} \\
\leq & \frac{\sqrt{\bar{k}} \|\nabla P_j(w_j^{(t)}; w^{(t-1)} \mid \eta, \gamma)\|_\infty + \|\nabla_S P_j(\bar{w}; w^{(t-1)} \mid \eta, \gamma)\|}{\gamma + \mu_{2k}} \\
\overset{\xi_2}{\leq} & \frac{\sqrt{\bar{k}}(\gamma + L_{2k}) w_{j,\min}^{(t)} + \|\nabla_S P_j(\bar{w}; w^{(t-1)} \mid \eta, \gamma)\|}{\gamma + \mu_{2k}} \\
\overset{\xi_3}{\leq} & \frac{\sqrt{\bar{k}}(\gamma + L_{2k}) \|w_j^{(t)} - \bar{w}\|}{\sqrt{k - \bar{k}}(\gamma + \mu_{2k})} + \frac{\|\nabla_S P_j(\bar{w}; w^{(t-1)} \mid \eta, \gamma)\|}{\gamma + \mu_{2k}},
\end{aligned}$$

where "ξ_1" follows from the optimality of $w^{(t)}$ on its own support such that $\nabla_{S^{(t)}} P_j(w_j^{(t)}; w^{(t-1)} \mid \eta, \gamma) = 0$, "$\xi_2$" is according to (B.1) and "ξ_3" is based on the fact of $\|w_j^{(t)} - \bar{w}\| \geq \sqrt{k - \bar{k}} w_{j,\min}^{(t)}$. Since the condition on k implies $\frac{\sqrt{\bar{k}}(\gamma + L_{2k})}{\sqrt{k-\bar{k}}(\gamma + \mu_{2k})} < 1$, by properly rearranging both sides of the above inequality and noting $\mu_{3k} \leq \mu_{2k}$ and $L_{3k} \geq L_{2k}$,

$$\|w_j^{(t)} - \bar{w}\| \leq \frac{\sqrt{k - \bar{k}} \|\nabla_S P_j(\bar{w}; w^{(t-1)} \mid \eta, \gamma)\|}{\sqrt{k - \bar{k}}(\gamma + \mu_{3k}) - \sqrt{\bar{k}}(\gamma + L_{3k})}.$$

Let $Q(w) = \eta F(w) - F_j(w) - \frac{\gamma}{2}\|w\|^2$. Then for any w, w' with $\|w - w'\|_0 \leq 3k$, we have

$$\frac{\eta\mu_{3k} - L_{3k} - \gamma}{2}\|w - w'\|^2 \leq Q(w) - Q(w') - \langle \nabla Q(w'), w - w' \rangle \leq \frac{\eta L_{3k} - \mu_{3k} - \gamma}{2}\|w - w'\|^2.$$

By setting $\eta = \frac{L_{3k}}{\mu_{3k}}$ and $\gamma = \frac{L_{3k}^2 - \mu_{3k}^2}{2\mu_{3k}}$ in the above we get

$$-\frac{L_{3k}^2 - \mu_{3k}^2}{4\mu_{3k}} \leq Q(w) - Q(w') - \langle \nabla Q(w'), w - w' \rangle \leq \frac{L_{3k}^2 - \mu_{3k}^2}{4\mu_{3k}}\|w - w'\|^2.$$

Then according to Lemma 9 we know that $Q(w)$ has $\left(\frac{L_{3k}^2 - \mu_{3k}^2}{2\mu_{3k}}\right)$-RLG. Let $S' = S_j^{(t)} \cup S^{(t-1)} \cup \bar{S}$. Then based on the above we can show that

$$\begin{aligned}
& \|\nabla_S P(\bar{w}; w^{(t-1)} \mid \eta, \gamma)\| \\
\leq & \|\nabla_{S'} P_j(\bar{w}; w^{(t-1)} \mid \eta, \gamma)\| \\
= & \left\|\eta \nabla_{S'} F(w^{(t-1)}) - \nabla_{S'} F_j(w^{(t-1)}) - \gamma w^{(t-1)} + \nabla_{S'} F_j(\bar{w}) + \gamma \bar{w}\right\| \\
\leq & \left\|\left(\eta \nabla_{S'} F(w^{(t-1)}) - \nabla_{S'} F_j(w^{(t-1)}) - \gamma w^{(t-1)}\right) - \left(\eta \nabla_{S'} F(\bar{w}) - \nabla_{S'} F_j(\bar{w}) - \gamma \bar{w}\right)\right\| \\
& + \eta \|\nabla_{S'} F(\bar{w})\| \\
\leq & \frac{L_{3k}^2 - \mu_{3k}^2}{2\mu_{3k}}\|w^{(t-1)} - \bar{w}\| + \frac{L_{3k}\|\nabla_{S'} F(\bar{w})\|}{\mu_{3k}}.
\end{aligned}$$

From the preceding two inequalities we get

$$\|w_j^{(t)} - \bar{w}\| \leq \frac{(L_{3k}^2 - \mu_{3k}^2)\sqrt{k-\bar{k}}}{(L_{3k}^2 + \mu_{3k}^2)\sqrt{k-\bar{k}} - \sqrt{\bar{k}}(L_{3k}^2 - \mu_{3k}^2 + 2\mu_{3k}L_{3k})}\|w^{(t-1)} - \bar{w}\| + \frac{2L_{3k}\sqrt{k-\bar{k}}}{(L_{3k}^2 + \mu_{3k}^2)\sqrt{k-\bar{k}} - \sqrt{\bar{k}}(L_{3k}^2 - \mu_{3k}^2 + 2\mu_{3k}L_{3k})}\|\nabla_{S'}F(\bar{w})\|.$$

From the definition of $w^{(t)}$ and by applying the second part of Lemma 8 we obtain

$$\|w^{(t)} - \bar{w}\| \leq \sqrt{1 + 2\sqrt{\frac{\bar{k}}{k-\bar{k}}}}\frac{1}{m}\sum_{j=1}^{m}\|w_j^{(t)} - \bar{w}\| \leq \theta\|w^{(t-1)} - \bar{w}\| + \frac{2\theta L_{3k}}{L_{3k}^2 - \mu_{3k}^2}\|\nabla_{S'}F(\bar{w})\|,$$

where $\theta = \frac{(L_{3k}^2-\mu_{3k}^2)(\sqrt{k-\bar{k}}+\sqrt{\bar{k}})}{(L_{3k}^2+\mu_{3k}^2)\sqrt{k-\bar{k}}-\sqrt{\bar{k}}(L_{3k}^2-\mu_{3k}^2+2\mu_{3k}L_{3k})}$. Given $k > \left(1 + \left(\frac{L_{3k}^2+\mu_{3k}L_{3k}-\mu_{3k}^2}{\mu_{3k}^2}\right)^2\right)\bar{k}$, we have $\theta \in (0,1)$. Then from the above recursion and $|S'| \leq 3k$ we obtain

$$\|w^{(t)} - \bar{w}\| \leq \theta^t\|w^{(0)} - \bar{w}\| + \frac{3.47\theta L_{3k}\sqrt{k}}{(1-\theta)(L_{3k}^2 - \mu_{3k}^2)}\|\nabla F(\bar{w})\|_\infty.$$

We need

$$t \geq \frac{1}{1-\theta}\log\frac{(1-\theta)(L_{3k}^2 - \mu_{3k}^2)\|w^{(0)} - \bar{w}\|}{\theta L_{3k}\sqrt{k}\|\nabla F(\bar{w})\|_\infty}$$

steps of iteration to achieve $\|w^{(t)} - \bar{w}\| \leq \frac{4.47\theta L_{3k}\sqrt{k}}{(1-\theta)(L_{3k}^2-\mu_{3k}^2)}\|\nabla F(\bar{w})\|_\infty$. This proves the desired complexity bound. □

4.7.6 Proof of Theorem 11

Proof. Since $F_j(w)$ are μ_{3k}-strongly-convex and L_{3k}-smooth, $F(w) = \frac{1}{m}\sum_{j=1}^m F_j(w)$ is also μ_{3k}-strongly-convex and L_{3k}-smooth. Let $Q(w) = \eta F(w) - F_j(w) - \frac{\gamma}{2}\|w\|^2$. Then for any w, w' with $\|w - w'\|_0 \leq 3k$, we have

$$\frac{\eta\mu_{3k} - L_{3k} - \gamma}{2}\|w - w'\|^2 \leq Q(w) - Q(w') - \langle\nabla Q(w'), w - w'\rangle \leq \frac{\eta L_{3k} - \mu_{3k} - \gamma}{2}\|w - w'\|^2.$$

By setting $\eta = \frac{L_{3k}}{\mu_{3k}}$ and $\gamma = \frac{L_{3k}^2-\mu_{3k}^2}{2\mu_{3k}}$ in the above we get

$$-\frac{L_{3k}^2 - \mu_{3k}^2}{4\mu_{3k}} \leq Q(w) - Q(w') - \langle\nabla Q(w'), w - w'\rangle \leq \frac{L_{3k}^2 - \mu_{3k}^2}{4\mu_{3k}}\|w - w'\|^2.$$

Then according to Lemma 9 we know that $Q(w)$ has $\left(\frac{L_{3k}^2-\mu_{3k}^2}{2\mu_{3k}}\right)$-RLG. By applying the first part of Lemma 11 we further obtain

$$\begin{aligned}\|w^{(t)} - \bar{w}\| &\leq \frac{3.24(L_{3k}^2 - \mu_{3k}^2)}{L_{3k}^2 + \mu_{3k}^2}\|w^{(t-1)} - \bar{w}\| + \frac{11.24L_{3k}\sqrt{k}}{L_{3k}^2 + \mu_{3k}^2}\|\nabla F(\bar{w})\|_\infty + 2.3\sqrt{\frac{2\mu_{3k}\epsilon}{L_{3k}^2 + \mu_{3k}^2}} \\ &\leq \theta\|w^{(t-1)} - \bar{w}\| + \frac{12.24L_{3k}\sqrt{k}}{L_{3k}^2 + \mu_{3k}^2}\|\nabla F(\bar{w})\|_\infty,\end{aligned}$$

where the last inequality is due to the assumption $L_{3k} < 1.37\mu_{3k}$ (which also implies $\theta = \frac{3.24(L_{3k}^2 - \mu_{3k}^2)}{L_{3k}^2 + \mu_{3k}^2} < 1$) and the assumption on the precision level ϵ. The desired result then follows by recursively applying the above inequality. □

Remark 12. *As we can see from this result that the contraction factor θ has polynomial dependence on the restricted condition number L_{3k}/μ_{3k}. Nevertheless, as we require $L_{3k}/\mu_{3k} < 1.37$, the factor θ is still reasonably well-controlled.*

4.7.7 Proof of Theorem 12

Proof. Recall that we update $w^{(t)} = w_1^{(t)}$ in this non-convex setting. Then the assumption $\|\nabla P_1(w_1^{(t)}; w^{(t-1)} \mid \eta, \gamma)\| \leq \epsilon$ implies

$$\|\nabla F_1(w^{(t)}) + \eta \nabla F(w^{(t-1)}) - \nabla F_1(w^{(t-1)}) + \gamma(w^{(t)} - w^{(t-1)})\| \leq \epsilon. \tag{C.1}$$

Since $F(w)$ is L_{2k}-smooth,

$$\begin{aligned}
F(w^{(t)}) &\leq F(w^{(t-1)}) + \langle \nabla F(w^{(t-1)}), w^{(t)} - w^{(t-1)} \rangle + \frac{L_{2k}}{2}\|w^{(t)} - w^{(t-1)}\|^2 \\
&= F(w^{(t-1)}) - \frac{1}{\eta}\langle \nabla F_1(w^{(t)}) - \nabla F_1(w^{(t-1)}) + \gamma(w^{(t)} - w^{(t-1)}), w^{(t)} - w^{(t-1)} \rangle \\
&+ \frac{L_{2k}}{2}\|w^{(t)} - w^{(t-1)}\|^2 + \frac{1}{\eta}\langle \nabla F_1(w^{(t)}) \\
&+ \eta \nabla F(w^{(t-1)}) - \nabla F_1(w^{(t-1)}) + \gamma(w^{(t)} - w^{(t-1)}), w^{(t)} - w^{(t-1)} \rangle \\
&\leq F(w^{(t-1)}) - \frac{2\gamma - (\eta+1)L_{2k}}{2\eta}\|w^{(t)} - w^{(t-1)}\|^2 + \frac{\epsilon}{\eta}\|w^{(t)} - w^{(t-1)}\|,
\end{aligned}$$

which implies

$$\frac{2\gamma - (\eta+1)L_{2k}}{2\eta}\|w^{(t)} - w^{(t-1)}\|^2 - \frac{\epsilon}{\eta}\|w^{(t)} - w^{(t-1)}\| \leq F(w^{(t-1)}) - F(w^{(t)}).$$

By adding the both sides of the above from index 1 to t we obtain

$$\begin{aligned}
&\min_{\tau=1,\ldots,t} \frac{2\gamma - (\eta+1)L_{2k}}{2\eta}\|w^{(\tau)} - w^{(\tau-1)}\|^2 - \frac{\epsilon}{\eta}\|w^{(\tau)} - w^{(\tau-1)}\| \\
\leq &\frac{1}{t}\sum_{\tau=1}^{t} \frac{2\gamma - (\eta+1)L_{2k}}{2\eta}\|w^{(\tau)} - w^{(\tau-1)}\|^2 - \frac{\epsilon}{\eta}\|w^{(\tau)} - w^{(\tau-1)}\| \\
\leq &\frac{1}{t}(F(w^{(0)}) - F(w^{(t)})) \leq \frac{1}{t}(F(w^{(0)}) - F(w^*)).
\end{aligned}$$

From the above and the basic fact that $ax^2 - bx - c < 0$ implies $x^2 \leq \frac{2b^2}{a^2} + \frac{2c}{a}$ for $a, b, c > 0$, we can verify

$$\min_{\tau=1,\ldots,t} \|w^{(\tau)} - w^{(\tau-1)}\|^2 \leq \frac{8\epsilon^2}{(\gamma - (\eta+1)L_{2k})^2} + \frac{4\eta(F(w^{(0)}) - F(w^*))}{(\gamma - (\eta+1)L_{2k})t}.$$

Then based on (C.1) and triangle inequality we can show that

$$\begin{aligned}\|\nabla F(w^{(t-1)})\|^2 &\leq \left(\frac{1}{\eta}\|\nabla F_1(w^{(t)}) - \nabla F_1(w^{(t-1)}) + \gamma(w^{(t)} - w^{(t-1)})\| + \epsilon\right)^2 \\ &\leq \frac{2(L_{2k}+\gamma)^2}{\eta^2}\|w^{(t)} - w^{(t-1)}\|^2 + 2\epsilon^2.\end{aligned}$$

By combining the preceding two inequalities we get

$$\min_{\tau=1,\ldots,t} \|\nabla F(w^{(\tau)})\|^2 \leq \left(\frac{16(L_{2k}+\gamma)^2}{\eta^2(\gamma - (\eta+1)L_{2k})^2} + 2\right)\epsilon^2 + \left(\frac{8(L_{2k}+\gamma)^2(F(w^{(0)}) - F(w^*))}{\eta(\gamma - (\eta+1)L_{2k})}\right)\frac{1}{t}.$$

The desired bound then follows from the setting of $\gamma = (\eta+2)L_{2k}$ and elementary calculus. □

Chapter 5

Conclusion and Future Work

5.1 Conclusion

Training sparse machine learning model has been shown to be an effective way of alleviating model overfitting, improving model interpretability, reducing computational cost in prediction and model storage space. Designing efficient sparse model learning algorithm keeps on getting extensive attention in machine learning research. The main theme of this thesis is to introduce my research on designing efficient sparse model learning algorithm for some widely used general learning objectives. In this thesis some popular optimization algorithm frameworks are covered, including Frank-Wolfe method, dual method and Newton-type method. For each designed algorithm the convergence is analyzed and numerical experiment is conducted to verify the superior algorithm efficiency over state of the art baseline methods.

In chapter 2, we introduced the proposed fully corrective Frank-Wolfe-type algorithm for solving the k-support-norm regularized sparse model learning problem. The proximal gradient algorithm is conventionally applied as the optimization framework to solve such a non-differentiable problem. The proximal gradient algorithm is featured by solving the proximal operator in each iteration. Motivated by the Frank-Wolfe algorithm which is originally designed for constrained model learning, the proposed k-FCFW algorithm reformulates the regularized minimization into a constrained minimization problem. In each iteration of k-FCFW, the major computation is searching top k entries in magnitude of the gradient. Under proper conditions, the algorithm analysis establishes the linear convergence rate for the proposed k-FCFW algorithm.

In chapter 3, we consider the ℓ_2-norm regularized empirical risk minimization problem with model parameter ℓ_0-norm constraint. Because of the model parameter ℓ_0-norm

constraint, the problem is in general non-convex and NP-hard. We first propose the duality theory of this sparse model learning problem. The Dual Iterative Hard Thresholding algorithm as well as its stochastic variant are proposed based on the sparse duality theory. Algorithm convergence analysis is conducted when proper conditions are satisfied. The designed algorithm is shown to be more efficient than primal domain optimization algorithms for SVM-type model learning tasks.

In chapter 4, we turn to study solving the ℓ_0-constrained empirical risk minimization problem in distributed computing environment. When training samples are distributed on multiple machines, special optimization algorithm design is needed to learn the global optimal model based on all training samples. In distributed optimization algorithm design, one important consideration is to keep low communication cost between machines. The proposed Distributed In-exact Newton-type Pursuit (DINPS) algorithm involves solving a local learning objective inexactly on each machine, communicating gradient and model parameter between worker machines and master machine. Algorithm analysis demonstrates that for a general class of convex functions with Lipschitze continuous Hessian, the method converges linearly with contraction factor scales inversely with data size; whilst the communication complexity required to reach desirable statistical accuracy scales logarithmically with the number of machines for some popular statistical learning models.

5.2 Future Work

Some interesting future directions are listed but are not limited to the following topics:

- Because of the varying numerical properties of machine learning model and training data, optimization algorithm design for each specific problem is needed, for the sake of model learning efficiency. This direction is worthwhile to be explored as the sample scale and model complexity keep on increasing and efficient model training is highly demanded.

- The wide application of non-convex models in various areas, with deep neural network as a representative, proposes the requirement of studying the theory

and methodology in non-convex optimization. Some algorithms are developed to solve the non-convex model training, such as forward-backward optimization in deep neural network learning, but the understanding is insufficient. Compared to convex model optimization, there are more technical challenges in non-convex optimization, such as local minima, saddle points and flat regions. Some recent research efforts in this direction include [20, 1, 39].

- The third research direction is to further explore the application of sparse model learning. Some current application examples of sparse model include feature learning [54], MR image reconstruction [64] and data clustering [23]. The properties of sparse model is highly desirable in many machine learning applications, therefore we believe more applications will appear.

Bibliography

[1] Z. Allen-Zhu and E. Hazan. Variance reduction for faster non-convex optimization. In *International Conference on Machine Learning*, 2016.

[2] A. Argyriou, R. Foygel, and N. Srebro. Sparse prediction with the k-support norm. In *Advances in Neural Information Processing Systems*, 2012.

[3] S. Bahmani, B. Raj, and P. Boufounos. Greedy sparsity-constrained optimization. *Journal of Machine Learning Research*, 14:807–841, 2013.

[4] A. Beck and Y. C. Eldar. Sparsity constrained nonlinear optimization: Optimality conditions and algorithms. *SIAM Journal on Optimization*, 23(3):1480–1509, 2013.

[5] A. Beck and M. 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.

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

[7] E. Belilovsky, A. Argyriou, G. Varoquaux, and M. Blaschko. Convex relaxations of penalties for sparse correlated variables with bounded total variation. *Machine Learning*, 100(2-3):533–553, 2015.

[8] E. Belilovsky, K. Gkirtzou, M. Misyrlis, A. B. Konova, J. Honorio, N. Alia-Klein, R. Z. Goldstein, D. Samaras, and M. 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.

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

[10] T. Blumensath. Compressed sensing with nonlinear observations and related nonlinear optimization problems. *IEEE Transactions on Information Theory*, 59(6):3466–3474, 2013.

[11] T. Blumensath and M. E. Davies. Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis*, 27(3):265–274, 2009.

[12] H. D. Bondell and B. J. Reich. Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar. *Biometrics*, 64(1):115–123, 2008.

[13] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

[14] D. Cai, X. He, J. Han, and T. S. Huang. Graph regularized nonnegative matrix factorization for data representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(8):1548–1560, 2011.

[15] E. J. Candes. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, 346(9-10):589–592, 2008.

[16] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011.

[17] S. Chatterjee, S. Chen, and A. Banerjee. Generalized dantzig selector: Application to the k-support norm. In *Advances in Neural Information Processing Systems*, 2014.

[18] J. Chen and Q. Gu. Accelerated stochastic block coordinate gradient descent for sparsity constrained nonconvex optimization. In *Conference on Uncertainty in Artificial Intelligence*, 2016.

[19] W. Dai and O. Milenkovic. Subspace pursuit for compressive sensing signal reconstruction. *IEEE Transactions on Information Theory*, 55(5):2230–2249, 2009.

[20] Y. N. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, and Y. Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Advances in Neural Information Processing Systems*, 2014.

[21] D. L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

[22] M. Dudik, Z. Harchaoui, and J. Malick. Lifted coordinate descent for learning with trace-norm regularization. In *International Conference on Artificial Intelligence and Statistics*, 2012.

[23] E. Elhamifar and R. Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2765–2781, 2013.

[24] A. Eriksson, T. Thanh Pham, T.-J. Chin, and I. Reid. The k-support norm and convex envelopes of cardinality and rank. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[25] S. Foucart. Hard thresholding pursuit: an algorithm for compressive sensing. *SIAM Journal on Numerical Analysis*, 49(6):2543–2563, 2011.

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

[27] R. M. Freund and P. Grigas. New analysis and results for the frank–wolfe method. *Mathematical Programming*, pages 1–32, 2014.

[28] D. Garber and E. Hazan. Faster rates for the Frank-Wolfe method over strongly-convex sets. In *Internatioal Conference on Machine Learning*, 2014.

[29] E. Grave, G. R. Obozinski, and F. R. Bach. Trace lasso: a trace norm regularization for correlated designs. In *Advances in Neural Information Processing Systems*, 2011.

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

[31] S. Han, J. Pool, J. Tran, and W. Dally. Learning both weights and connections for efficient neural network. In *Advances in Neural Information Processing Systems*, 2015.

[32] Z. Harchaoui, A. Juditsky, and A. Nemirovski. Conditional gradient algorithms for norm-regularized smooth convex optimization. *Mathematical Programming*, pages 1–38, 2014.

[33] E. Hazan. Sparse approximate solutions to semidefinite programs. In *LATIN*, pages 306–316, 2008.

[34] C.-J. Hsieh, K.-W. Chang, C.-J. Lin, S. S. Keerthi, and S. Sundararajan. A dual coordinate descent method for large-scale linear SVM. In *International Conference on Machine Learning*, 2008.

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

[36] M. Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *International Conference on Machine Learning*, 2013.

[37] M. Jaggi, V. Smith, M. Takác, J. Terhorst, S. Krishnan, T. Hofmann, and M. I. Jordan. Communication-efficient distributed dual coordinate ascent. In *Advances in Neural Information Processing Systems*, 2014.

[38] M. Jaggi and M. Sulovský. A simple algorithm for nuclear norm regularized problem. In *International Conference on Machine Learning*, 2010.

[39] P. Jain, P. Kar, et al. Non-convex optimization for machine learning. *Foundations and Trends® in Machine Learning*, 10(3-4):142–336, 2017.

[40] P. Jain, N. Rao, and I. Dhillon. Structured sparse regression via greedy hard-thresholding. *Advances in Neural Information Processing Systems*, 2016.

[41] P. Jain, A. Tewari, and P. Kar. On iterative hard thresholding methods for high-dimensional m-estimation. In *Advances in Neural Information Processing Systems*, 2014.

[42] X. Jin, X.-T. Yuan, J. Feng, and S. Yan. Training skinny deep neural networks with

iterative hard thresholding methods. *arXiv preprint arXiv:1607.05423*, 2016.

[43] R. Khanna and A. Kyrillidis. IHT dies hard: Provable accelerated iterative hard thresholding. In *International Conference on Artificial Intelligence and Statistics*, 2018.

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

[45] P. Kumar and E. A. Yildirim. A linearly convergent linear-time first-order algorithm for support vector classification with a core set result. *INFORMS Journal on Computing*, 23(3):377–391, 2011.

[46] S. Lacoste-Julien and M. Jaggi. An affine invariant linear convergence analysis for Frank-Wolfe algorithms. *NIPS Workshop on Greedy Algorithms, Frank-Wolfe and Friends*, 2013.

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

[48] S. Lacoste-Julien, M. Jaggi, M. Schmidt, and P. Pletscher. Block-coordinate Frank-Wolfe optimization for structural SVMs. *International Conference on Machine Learning*, 2013.

[49] H. Lai, Y. Pan, C. Lu, Y. Tang, and S. Yan. Efficient k-support matrix pursuit. In *European Conference on Computer Vision*. 2014.

[50] G. Lan. The complexity of large-scale convex programming under a linear optimization oracle. *arXiv preprint arXiv:1309.5550*, 2013.

[51] K. Lang. Newsweeder: Learning to filter netnews. In *International Conference on Machine Learning*, 1995.

[52] M. Lapin, B. Schiele, and M. Hein. Scalable multitask representation learning for scene classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[53] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[54] H. Lee, A. Battle, R. Raina, and A. Y. Ng. Efficient sparse coding algorithms. In *Advances in Neural Information Processing Systems*, 2007.

[55] S. Lee, D. Brzyski, and M. Bogdan. Fast saddle-point algorithm for generalized dantzig selector and fdr control with ordered ℓ_1-norm. In *International Conference on Artificial Intelligence and Statistics*, 2016.

[56] E. S. Levitin and B. T. Polyak. Constrained minimization methods. *USSR Computational Mathematics and Mathematical Physics*, 6(5):787–823, 1966.

[57] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li. RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397, 2004.

[58] M. Li, D. G. Andersen, A. J. Smola, and K. Yu. Communication efficient distributed machine learning with the parameter server. In *Advances in Neural Information Processing Systems*, 2014.

[59] X. Li, T. Zhao, R. Arora, H. Liu, and J. Haupt. Stochastic variance reduced optimization for nonconvex sparse learning. In *International Conference on Machine Learning*, 2016.

[60] B. Liu, X.-T. Yuan, L. Z. Wang, Q. S. Liu, J. Z. Huang, and D. N. Metaxas. Distributed inexact Newton-type pursuit for non-convex sparse learning. In *International Conference on Artificial Intelligence and Statistics*, 2019, to appear.

[61] B. Liu, X.-T. Yuan, L. Z. Wang, Q. S. Liu, and D. N. Metaxas. Dual iterative hard thresholding: From non-convex sparse minimization to non-smooth concave maximization. In *International Conference on Machine Learning*, 2017.

[62] B. Liu, X.-T. Yuan, S. T. Zhang, Q. S. Liu, and D. N. Metaxas. Efficient k-support-norm regularized minimization via fully corrective Frank-Wolfe method. In *International Joint Conference on Artificial Intelligence*, 2016.

[63] A. Lorbert, D. Eis, V. Kostina, D. M. Blei, and P. J. Ramadge. Exploiting covariate similarity in sparse regression via the pairwise elastic net. In *International Conference on Artificial Intelligence and Statistics*, 2010.

[64] M. Lustig, D. Donoho, and J. M. Pauly. Sparse MRI: The application of compressed sensing for rapid MR imaging. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 58(6):1182–1195, 2007.

[65] S. Mallat and Z. Zhang. Matching pursuit with time-frequency dictionaries. *IEEE Transactions on Single Processing*, 41(12):3397–3415, 1993.

[66] R. Mazumder and T. Hastie. The graphical lasso: New insights and alternatives. *Electronic Journal of Statistics*, 6:2125, 2012.

[67] A. M. McDonald, M. Pontil, and D. Stamos. Spectral k-support norm regularization. In *Advances in Neural Information Processing Systems*, 2014.

[68] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, et al. Communication-efficient learning of deep networks from decentralized data. In *International Conference on Artificial Intelligence and Statistics*, 2017.

[69] R. Ñanculef, E. Frandi, C. Sartori, and H. Allende. A novel Frank–Wolfe algorithm. analysis and applications to large-scale SVM training. *Information Sciences*, 285:66–99, 2014.

[70] N. Nguyen, D. Needell, and T. Woolf. Linear convergence of stochastic iterative greedy algorithms with sparse constraints. *IEEE Transactions on Information Theory*,

63(11):6869–6895, 2017.

[71] N. Parikh and S. Boyd. Proximal algorithms. *Foundations and Trends® in Optimization*, 1(3):123–231, 2013.

[72] S. Patterson, Y. C. Eldar, and I. Keidar. Distributed compressed sensing for static and time-varying networks. *IEEE Transactions on Signal Processing*, 62(19):4931–4946, 2014.

[73] N. Rao, P. Shah, and S. Wright. Forward-backward greedy algorithms for atomic norm regularization. *IEEE Transactions on Single Processing*, 63(21):5798–5811, 2015.

[74] S. J. Reddi, A. Hefny, S. Sra, B. Poczos, and A. Smola. Stochastic variance reduction for nonconvex optimization. In *International Conference on Machine Learning*, 2016.

[75] S. J. Reddi, J. Konečnỳ, P. Richtárik, B. Póczós, and A. Smola. AIDE: Fast and communication efficient distributed optimization. *arXiv preprint arXiv:1608.06879*, 2016.

[76] J. Ren, X. Li, and J. Haupt. Communication-efficient algorithm for distributed sparse learning via two-way truncation. *IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing*, 2017.

[77] M. W. Schmidt, E. Berg, M. P. Friedlander, and K. P. Murphy. Optimizing costly functions with simple constraints: A limited-memory projected quasi-Newton algorithm. In *International Conference on Artificial Intelligence and Statistics*, 2009.

[78] S. Shalev-Shwartz. SDCA without duality, regularization, and individual convexity. In *International Conference on Machine Learning*, 2016.

[79] S. Shalev-Shwartz, A. Gonen, and O. Shamir. Large-scale convex minimization with a low-rank constraint. In *International Conference on Machine Learning*, 2011.

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

[81] S. Shalev-Shwartz and T. Zhang. Accelerated mini-batch stochastic dual coordinate ascent. In *Advances in Neural Information Processing Systems*, 2013.

[82] S. Shalev-Shwartz and T. Zhang. Stochastic dual coordinate ascent methods for regularized loss. *The Journal of Machine Learning Research*, 14(1):567–599, 2013.

[83] S. Shalev-Shwartz and T. Zhang. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. *Mathematical Programming*, 155(1-2):105–145, 2016.

[84] O. Shamir, N. Srebro, and T. Zhang. Communication-efficient distributed optimization using an approximate Newton-type method. In *International Conference on Machine Learning*, 2014.

[85] J. Shen and P. Li. On the iteration complexity of support recovery via hard thresholding pursuit. In *International Conference on Machine Learning*, 2017.

[86] J. Shen and P. Li. Partial hard thresholding: Towards a principled analysis of support

recovery. In *Advances in Neural Information Processing Systems*, 2017.

[87] J. Shen and P. Li. A tight bound of hard thresholding. *Journal of Machine Learning Research*, 18:7650–7691, 2018.

[88] J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

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

[90] J. Wang, M. Kolar, N. Srebro, and T. Zhang. Efficient distributed learning with sparsity. In *International Conference on Machine Learning*, 2017.

[91] W. Wen, C. Wu, Y. Wang, Y. Chen, and H. Li. Learning structured sparsity in deep neural networks. In *Advances in Neural Information Processing Systems*, 2016.

[92] A. W. Yu, Q. Lin, and T. Yang. Doubly stochastic primal-dual coordinate method for empirical risk minimization and bilinear saddle-point problem. *arXiv preprint arXiv:1508.03390*, 2015.

[93] X.-T. Yuan, P. Li, and T. Zhang. Gradient hard thresholding pursuit for sparsity-constrained optimization. In *International Conference on Machine Learning*, 2014.

[94] X.-T. Yuan, P. Li, and T. Zhang. Exact recovery of hard thresholding pursuit. In *Advances in Neural Information Processing Systems*, 2016.

[95] X.-T. Yuan, B. Liu, L. Z. Wang, Q. S. Liu, and D. N. Metaxas. Dual iterative hard thresholding. *Journal of Machine Learning Research*, under review. The first two authors contributed equally to this article.

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

[97] C.-H. Zhang and J. Huang. The sparsity and bias of the Lasso selection in high-dimensional linear regression. *The Annals of Statistics*, 36(4):1567–1594, 2008.

[98] Y. Zhang and L. Xiao. DiSCO: Distributed optimization for self-concordant empirical loss. In *International Conference on Machine Learning*, 2015.

[99] Y. Zhang and L. Xiao. Stochastic primal-dual coordinate method for regularized empirical risk minimization. *The Journal of Machine Learning Research*, 18(1):2939–2980, 2017.

[100] Y. Zhou, R. Jin, and S. C. Hoi. Exclusive lasso for multi-task feature selection. In *International Conference on Artificial Intelligence and Statistics*, 2010.

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