Journal of Machine Learning Research 21 (2020) 1-50 Submitted 7/18; Revised 12/19; Published 6/20

Dual Iterative Hard Thresholding

Xiao-Tong Yuan XTYUAN@NUIST.EDU.CN
B-DAT Lab and CICAEET, Nanjing University of Information Science and Technology
Nanjing, 210044, China

Bo Liu KFLIUBO@GMAIL.COM
JD Finance America Corporation
Mountain View, CA 94043, USA

Lezi Wang WANGLEZI.BUPT@GMAIL.COM
Department of Computer Science, Rutgers University
Piscataway, NJ 08854, USA

Qingshan Liu QSLIU@NUIST.EDU.CN
B-DAT Lab and CICAEET, Nanjing University of Information Science and Technology
Nanjing, 210044, China

Dimitris N. Metaxas DNM@CS.RUTGERS.EDU
Department of Computer Science, Rutgers University
Piscataway, NJ 08854, USA

Editor: David Wipf

Abstract

Iterative Hard Thresholding (IHT) is a popular class of first-order greedy selection methods for loss minimization under cardinality constraint. The existing IHT-style algorithms, however, are proposed for minimizing the primal formulation. It is still an open issue to explore duality theory and algorithms for such a non-convex and NP-hard combinatorial optimization problem. To address this issue, we develop in this article a novel duality theory for ℓ_2-regularized empirical risk minimization under cardinality constraint, along with an IHT-style algorithm for dual optimization. Our sparse duality theory establishes a set of sufficient and/or necessary conditions under which the original non-convex problem can be equivalently or approximately solved in a concave dual formulation. In view of this theory, we propose the Dual IHT (DIHT) algorithm as a super-gradient ascent method to solve the non-smooth dual problem with provable guarantees on primal-dual gap convergence and sparsity recovery. Numerical results confirm our theoretical predictions and demonstrate the superiority of DIHT to the state-of-the-art primal IHT-style algorithms in model estimation accuracy and computational efficiency.[1]

Keywords: Iterative hard thresholding, Duality theory, Sparsity recovery, Non-convex optimization

1. Introduction

We consider the problem of learning sparse predictive models which has been extensively studied in high-dimensional statistical learning (Hastie et al., 2015; Bach et al., 2012). Given

1. A conference version of this article appeared at ICML 2017 (Liu et al., 2017). The first two authors contributed equally to this article.

a set of training samples $\{(x_i, y_i)\}_{i=1}^N$ in which $x_i \in \mathbb{R}^d$ 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 for learning the sparse representation of a linear predictive model (Bahmani et al., 2013):

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

Here $w \in \mathbb{R}^d$ is the model parameter vector, $l(w^\top x_i; y_i)$ is a convex function that measures the linear regression/prediction loss of w at data point (x_i, y_i), and $\lambda > 0$ is the regularization modulus. For example, the squared loss $l(u, v) = \frac{1}{2}(u - v)^2$ is used in linear regression and the hinge loss $l(u, v) = \max\{0, 1 - uv\}$ in support vector machines. The cardinality constraint $\|w\|_0 \leq k$ is imposed for improving learnability and interpretability of model when d is potentially much larger than N. Due to the presence of such a constraint, the problem (1) is simultaneously non-convex and NP-hard even for the quadratic loss function (Natarajan, 1995), hence is challenging for optimization. An alternative way to address this challenge is to use proper convex relaxation, e.g., ℓ_1-norm (Tibshirani, 1996) and k-support norm (Argyriou et al., 2012), as an alternative of the cardinality constraint. However, the convex relaxation based techniques tend to introduce bias for parameter estimation (Zhang and Huang, 2008).

In this article, we are interested in algorithms that directly minimize the non-convex formulation (1). Early efforts mainly lie in compressed sensing for signal recovery, which is a special case of (1) with squared loss (Donoho, 2006; Pati et al., 1993). Among others, a family of the so called Iterative Hard Thresholding (IHT) methods have gained particular interests and they have been witnessed to offer the fastest and most scalable solutions in many cases (Blumensath and Davies, 2009; Foucart, 2011). In recent years, IHT-style methods have been generalized to handle more general convex loss functions (Beck and Eldar, 2013; Jain et al., 2014; Yuan et al., 2018) as well as structured sparsity constraints (Jain et al., 2016). 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. In our problem setting, a plain IHT iteration is given by

$$w^{(t)} = \mathrm{H}_k\left(w^{(t-1)} - \eta \nabla P(w^{(t-1)})\right),$$

where $\mathrm{H}_k(\cdot)$ is the truncation operator that preserves the top k (in magnitude, with ties broken arbitrarily) entries of input and sets the remaining to be zero, and $\eta > 0$ is the learning rate. In practice, IHT-style algorithms have found their applications in deep neural networks compression (Jin et al., 2016), sparse signal demixing from noisy observations (Soltani and Hegde, 2017), and fluorescence molecular lifetime tomography (Cai et al., 2016), to name a few.

Although IHT-style methods have long been studied, so far this class of methods is only designed for optimizing the primal formulation (1). It still remains an open problem to investigate the feasibility of solving the original NP-hard/non-convex formulation in a dual space that might potentially generate new sparsity recovery theory and further improve computational efficiency. To fill this gap, inspired by the emerging success of dual

optimization methods in regularized learning problems (Shalev-Shwartz and Zhang, 2013b, 2016; Xiao, 2010; Tan et al., 2018), we establish in this article a sparse Lagrangian duality theory and propose an IHT-style algorithm along with its stochastic extension for efficient dual optimization.

1.1 Overview of Contribution

The core contribution of this work is two-fold in theory and algorithm. As the theoretical aspect of our contribution, we have established a novel strong sparse Lagrangian duality theory for the NP-hard and non-convex combinatorial optimization problem (1) which to the best of our knowledge has not been reported elsewhere in literature. A fundamental result is showing that the following condition is *sufficient and necessary* to guarantee a zero primal-dual gap between the primal non-convex problem and the dual concave problem:

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

where $\bar{w}$ is the k-sparse primal minimizer and $\bar{\alpha} \in [\partial l_1(\bar{w}^\top x_1), ..., \partial l_N(\bar{w}^\top x_N)]$. The strong sparse duality theory suggests a natural way for finding the global minimum of the sparsity-constrained minimization problem (1) via equivalently maximizing its dual problem as given in (6) which is concave. Although can be partially verified for some special models such as sparse linear regression and logistic regression, the preceding condition could more often than not be violated in practice, leading to non-zero primal-dual gap. In order to address this issue, we further develop an approximate sparse duality theory to cover the setting where sparse strong duality does not hold exactly.

On the algorithm side, we propose the Dual IHT (DIHT) algorithm as a super-gradient ascent 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 further proposed to handle large-scale learning problems. For both algorithms, we provide non-asymptotic convergence analysis on dual estimation error, primal-dual gap, and sparsity recovery as well. In contrast to the existing analysis for primal IHT-style algorithms, our analysis is not explicitly relying on the Restricted Isometry Property (RIP) conditions and thus would be less restrictive in real-life high-dimensional estimation. Numerical results on synthetic and benchmark data sets demonstrate that DIHT and its stochastic extension significantly outperform the state-of-the-art primal IHT-style algorithms in estimation accuracy and computational efficiency.

The main contributions of this article are highlighted in below:

- *Sparse Lagrangian duality theory*: we introduce a sparse saddle point theorem (Theorem 2), a sparse mini-max theorem (Theorem 5) and a sparse strong duality theorem (Theorem 8). Moreover, we establish an approximate duality theory (Theorem 13) as a complement to the strong sparse duality theory.

- *Dual iterative hard thresholding algorithm*: we propose 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

Notation	Definition
N	number of samples
d	number of features
k	the sparsity level hyper-parameter
λ	the regularization strength hyper-parameter
$P(w)$	the primal objective function
$\bar{w}$	the primal k-sparse minimizer given by $\bar{w} := \arg\min_{\|w\|_0 \leq k} P(w)$
$D(\alpha)$	the dual objective function
$\mathcal{F}$	the feasible set of dual variables
$\bar{\alpha}$	the dual maximizer given by $\bar{\alpha} := \arg\max_{\alpha \in \mathcal{F}} D(\alpha)$
$\epsilon_{PD}(w, \alpha)$	the sparse primal-dual gap defined by $\epsilon_{PD}(w, \alpha) := P(w) - D(\alpha)$
X	the data matrix of which the columns are N data samples
$\mathrm{H}_F(x)$	the truncation operator that restricts vector x to the index set F
$\mathrm{H}_k(x)$	the truncation operator that restricts x to its top k (in magnitude) entries
$\mathrm{supp}(x)$	the index set of non-zero entries of x
$\|x\|_0$	number of non-zero entries of x, i.e., $\|x\|_0 := \|\mathrm{supp}(x)\|$
$[x]_i$	the i-th entry of x
$\|x\|_\infty$	the largest (in magnitude) element of x, i.e., $\|x\|_\infty := \max_i \|[x]_i\|$
$x_{\min}$	the smallest non-zero element of x, i.e., $x_{\min} := \min_{i \in \mathrm{supp}(x)} \|[x]_i\|$
$\lambda_{\max}(A)$	the largest eigenvalue of matrix A
$\lambda_{\min}(A)$	the smallest eigenvalue of A
$\|A\|$	the spectral norm (the largest singular value) of A
A_F	the restriction of A with *rows* restricted to F

Table 1: Table of notation.

are Lipschitz smooth (see Theorem 15 and Theorem 19), and at linear rates if further assuming that the loss functions are strongly convex (see Theorem 17 and Theorem 22).

1.2 Notation and Organization

Notation. The key quantities and notations that commonly used in our analysis are summarized in Table 1.

Organization. The rest of this article is organized as follows: In Section 2 we briefly review the related literature. In Section 3 we develop a Lagrangian duality theory for sparsity-constrained minimization problems. The dual IHT algorithms along with convergence analysis are presented in Section 4. The numerical evaluation results are reported in Section 5. Finally, the concluding remarks are made in Section 6. All the technical proofs are deferred to the appendix sections.

2. Related Work

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 (Blumensath, 2013; Bahmani et al., 2013;

Yuan et al., 2014). In the work of Jain et al. (2014), several sparsity-level-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 has been studied to understand when the algorithm can exactly recover the support of a sparse signal from its compressed measurements (Yuan et al., 2016; Shen and Li, 2017a,b). A Nesterov's momentum based hard thresholding method was proposed by Khanna and Kyrillidis (2018) to further improve the efficiency of IHT. 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 via leveraging the finite-sum structure of learning problem (Nguyen et al., 2017; Li et al., 2016; Chen and Gu, 2016; Shen and Li, 2018; Zhou et al., 2018). For distributed learning with sparsity, an approximate Newton-type extension of IHT was developed that takes advantage of the stochastic nature of problem to improve communication efficiency. The generalization performance of IHT has recently been studied from the perspective of algorithmic stability (Yuan and Li, 2020).

Another related line of research is dual optimization which has gained considerable popularity in various machine learning tasks including kernel learning (Hsieh et al., 2008), online learning (Xiao, 2010), multi-task learning (Lapin et al., 2014) and graphical models learning (Mazumder and Hastie, 2012). In recent years, a number of stochastic dual coordinate ascent (SDCA) methods have been proposed for solving large-scale regularized loss minimization problems (Shalev-Shwartz and Zhang, 2013a,b, 2016). All these methods exhibit fast convergence rate in theory and highly competitive numerical performance in practice. Shalev-Shwartz (2016) also developed a dual free variant of SDCA that supports non-regularized objectives and non-convex individual loss functions. 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 (Chambolle and Pock, 2011) and generalized Dantzig selector (Lee et al., 2016). For large-scale machine learning, several stochastic and distributed variants were developed to make the primal-dual algorithms more computationally efficient and scalable (Zhang and Xiao, 2017; Yu et al., 2015; Tan et al., 2018; Xiao et al., 2019).

Our work lies at the intersection of the above two disciplines of research. Although dual optimization methods have long been studied in machine learning, it still remains largely unknown, in both theory and algorithm, how to apply dual methods to the non-convex and NP-hard sparse estimation problem (1) where the non-convexity arises from the cardinality constraint rather than the objective function. The main contribution of the present article is closing this gap by presenting a novel sparse Lagrangian duality theory and a dual IHT method with provable guarantees on sparsity recovery accuracy and efficiency.

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 (1) can be equivalently solved in a dual space. The results in this part lay a theoretical foundation for developing dual sparse estimation methods.

3.1 Sparse Strong Duality Theory

From here onward we abbreviate $l_i(u) = l(u, y_i)$. The convexity of $l(w^\top x_i, y_i)$ implies that $l_i(u)$ is also convex. Let $l_i^*(\alpha_i) = \max_u\{\alpha_i u - l_i(u)\}$ be the convex conjugate of $l_i(u)$ and $\mathcal{F}_i \subseteq \mathbb{R}$ be the feasible set of α_i. According to the standard expression of $l_i(u) = \max_{\alpha_i \in \mathcal{F}_i}\{\alpha_i u - l_i^*(\alpha_i)\}$, the problem (1) can be reformulated into the following mini-max formulation:

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

The following defined Lagrangian form will be useful in analysis:

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

where $\alpha = [\alpha_1, ..., \alpha_N] \in \mathcal{F} := \mathcal{F}_1 \times \cdots \times \mathcal{F}_N \subseteq \mathbb{R}^N$ 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 1 (Sparse saddle point) *Let $k > 0$ be an integer. A pair $(\bar{w}, \bar{\alpha}) \in \mathbb{R}^d \times \mathcal{F}$ is said to be a k-sparse saddle point for L if $\|\bar{w}\|_0 \le k$ and the following holds for all $\|w\|_0 \le k, \alpha \in \mathcal{F}$:*

$$L(\bar{w}, \alpha) \le L(\bar{w}, \bar{\alpha}) \le L(w, \bar{\alpha}). \tag{3}$$

Different from the conventional definition of saddle point, the k-sparse saddle point only requires that the inequality (3) holds for an arbitrary k-sparse vector w. The following result is a basic sparse saddle point theorem for L. Throughout the article, we will use $l'(\cdot)$ to denote a sub-gradient (or super-gradient) of a convex (or concave) function $l(\cdot)$, and use $\partial l(\cdot)$ to denote its sub-differential (or super-differential).

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

(a) $\bar{w}$ solves the primal problem in (1);

(b) $\bar{\alpha} \in [\partial l_1(\bar{w}^\top x_1), ..., \partial l_N(\bar{w}^\top x_N)]$;

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

Remark 3 *Theorem 2 shows that the conditions (a)$\sim$(c) are sufficient and necessary to guarantee the existence of a sparse saddle point for the Lagrangian form L. The condition (c) can be regarded as a Sparsity Constraint Qualification condition 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 straightforward to verify that the condition (c) in Theorem 2 is equivalent to*

$$\mathrm{H}_{\bar{F}}(P'(\bar{w})) = 0, \quad \bar{w}_{\min} \ge \frac{1}{\lambda}\|P'(\bar{w})\|_\infty. \tag{4}$$

To gain some intuition of the above condition, let us consider a simple example where $w \in \mathbb{R}^N$, $\{x_i = e_i\}$ *are the standard basis of* $\mathbb{R}^N$ *and* $P(w) = \frac{1}{2N}\sum_{i=1}^N (y_i - w_i)^2 + \frac{\lambda}{2}\|w\|^2$ *is quadratic. In this case, it is trivial to see that the primal minimizer is given by* $\bar{w} = \mathrm{H}_k\left(\frac{y}{1+\lambda N}\right)$ *and* $P'(\bar{w}) = (\lambda + \frac{1}{N})\bar{w} - \frac{1}{N}y$. *Denote* $[|y|]_{(j)}$ *the* j*-th largest entry of* $|y|$, *i.e.,* $[|y|]_{(1)} \geq [|y|]_{(2)} \geq ... \geq [|y|]_{(N)}$. *Then the above condition* (4) *is characterized by* $\frac{[|y|]_{(k)}}{1+\lambda N} \geq \frac{[|y|]_{(k+1)}}{\lambda N}$ *which basically 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 such a* λ.

The following sparse mini-max theorem guarantees that the min and max in formulation (2) can be safely switched if and only if there exists a sparse saddle point for $L(w, \alpha)$.

Theorem 5 (Sparse mini-max theorem) *The mini-max relationship*

$$\max_{\alpha\in\mathcal{F}} \min_{\|w\|_0\leq k} L(w,\alpha) = \min_{\|w\|_0\leq k} \max_{\alpha\in\mathcal{F}} L(w,\alpha) \tag{5}$$

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

The sparse mini-max result in Theorem 5 provides sufficient and necessary conditions under which one can safely exchange a min-max for a max-min, in the presence of non-convex cardinality constraint. The following corollary is a direct consequence of invoking Theorem 2 to Theorem 5.

Corollary 6 *The mini-max relationship*

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

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

The mini-max result in Theorem 5 can be used as a basis for establishing sparse duality theory. Indeed, we have already shown the following:

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

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.,

$$\max_{\alpha\in\mathcal{F}} D(\alpha) = \max_{\alpha\in\mathcal{F}} \min_{\|w\|_0\leq k} L(w,\alpha). \tag{6}$$

The following Proposition 7 shows that the dual objective function $D(\alpha)$ is concave and explicitly gives the expression of its super-differential.

Proposition 7 *The dual objective function* $D(\alpha)$ *is given by*

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

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

$$\partial D(\alpha) = \frac{1}{N}[w(\alpha)^\top x_1 - \partial l_1^*(\alpha_1), ..., w(\alpha)^\top x_N - \partial l_N^*(\alpha_N)].$$

Particularly, if $w(\alpha)$ *is unique at* α *with respect to the truncation operator* $\mathrm{H}_k(\cdot)$ *and* $\{l_i^*\}_{i=1,...,N}$ *are differentiable, then* $\partial D(\alpha)$ *is unique and it is the super-gradient of* $D(\alpha)$.

In view of Theorem 2 and Theorem 5, we are in the position to 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 8 (Sparse strong duality theorem) *Let* $\bar{w} \in \mathbb{R}^d$ *be a* k*-sparse primal vector and* $\bar{\alpha} \in \mathcal{F}$ *be a dual vector. Then* $\bar{\alpha}$ *solves the dual problem in* (6)*, i.e.,* $D(\bar{\alpha}) \geq D(\alpha)$, $\forall \alpha \in \mathcal{F}$*, 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.*

We define the sparse primal-dual gap $\epsilon_{PD}(w, \alpha) := P(w) - D(\alpha)$. The main message conveyed by Theorem 8 is that the conditions (a)∼(c) in Theorem 2 are sufficient and necessary to guarantee a zero primal-dual gap at the sparse primal-dual pair $(\bar{w}, \bar{\alpha})$.

3.2 On 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 $X = [x_1, ..., x_N] \in \mathbb{R}^{d\times N}$ the data matrix which contains the N data samples as columns.

Theorem 9 *Assume that each* l_i^* *is differentiable and smooth, and each dual feasible set* $\mathcal{F}_i$ *is convex. Let* $\bar{\alpha} = \arg\max_\alpha D(\alpha)$ *be a dual maximizer. If* $w(\bar{\alpha}) = \mathrm{H}_k\left(-\frac{1}{\lambda N}\sum_{i=1}^N \bar{\alpha}_i x_i\right)$ *is unique at* $\bar{\alpha}$ *with respect to truncation operation, 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 10 *The dual sufficient condition given in Theorem 9 basically shows that under mild conditions, if* $w(\bar{\alpha})$ *constructed from a dual maximizer* $\bar{\alpha}$ *is unique with respect to the truncation operator* $\mathrm{H}_k(\cdot)$*, 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.*

To gain better intuition of Theorem 9, we discuss its implications for sparse linear regression and logistic regression models which are commonly used in statistical machine learning.

Example I: Sparse strong duality for linear regression. Consider the special case of the primal problem (1) with least square loss $l(w^\top x_i, y_i) = \frac{1}{2}(y_i - w^\top x_i)^2$. Let us write $l(w^\top x_i, y_i) = l_i(w^\top x_i)$ with $l_i(a) = \frac{1}{2}(a - y_i)^2$. It is standard to know that the convex conjugate of $l_i(a)$ is $l_i^*(\alpha_i) = \frac{\alpha_i^2}{2} + y_i\alpha_i$ and $\mathcal{F}_i = \mathbb{R}$. Obviously, l_i^* is differentiable and

$\mathcal{F}_i$ is convex. According to Theorem 9, if $w(\bar{\alpha}) = \mathrm{H}_k\left(-\frac{1}{\lambda N}\sum_{i=1}^N \bar{\alpha}_i x_i\right)$ is unique at the dual maximizer $\bar{\alpha}$ with respect to truncation operation, then $w(\bar{\alpha})$ is a primal minimizer of $P(w)$. To illustrate this claim, let us consider the same example as presented in Remark 4, of which the dual objective function is expressed as

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

Provided that $\frac{\lambda N [|y|]_{(k)}}{1+\lambda N} > [|y|]_{(k+1)}$, it can be readily verified 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 $w(\bar{\alpha}) = \mathrm{H}_k\left(-\frac{1}{\lambda N}\bar{\alpha}\right)$ is then by definition unique with respect to the truncation operator. According to the discussion in Remark 4, $w(\bar{\alpha})$ is exactly the primal minimizer. This verifies the validness of Theorem 9 on the considered example. On the other side, to see an example in which the uniqueness condition on $w(\bar{\alpha})$ can be violated, let us consider a special case $y = [2, 2, 1]$, $\lambda = 1$ and $k = 1$. From the discussion in Remark 4 we know that the primal sparse minimizer is $\bar{w} = [0.5, 0, 0]$ (or $\bar{w} = [0, 0.5, 0]$) with $P(\bar{w}) = 4/3$. In the meanwhile, it can be verified by exhaustive search that the maximal dual objective value is attained at $\bar{\alpha} = [-1.5, -1.5, -1]$ with $D(\bar{\alpha}) = 7/6$. Obviously, here $w(\bar{\alpha}) = \mathrm{H}_k\left(-\frac{1}{\lambda N}\bar{\alpha}\right)$ is not unique and the primal-dual gap is non-zero which indicates that strong duality fails in this case.

Example II: Sparse strong duality for logistic regression. In logistic regression model, given a k-sparse parameter vector $\bar{w}$, the relation between the random feature vector $x \in \mathbb{R}^d$ 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 $l(w^\top x_i, y_i) = l_i(w^\top x_i) = \log\left(1 + \exp(-y_i w^\top x_i)\right)$, where $l_i(a) = \log\left(1 + \exp(-a y_i)\right)$. In this case, we have $l_i^*(\alpha_i) = -\alpha_i y_i \log(-\alpha_i y_i) + (1 + \alpha_i y_i)\log(1 + \alpha_i y_i)$ with $\alpha_i y_i \in [-1, 0]$. Note that l_i^* is differentiable and $\mathcal{F}_i$ is convex. Therefore Theorem 9 implies that if $w(\bar{\alpha}) = \mathrm{H}_k\left(-\frac{1}{\lambda N}\sum_{i=1}^N \bar{\alpha}_i x_i\right)$ is unique at $\bar{\alpha}$ with respect to truncation operation, then $w(\bar{\alpha})$ is a primal minimizer of $P(w)$ satisfying $P(w(\bar{\alpha})) = D(\bar{\alpha})$.

3.3 Approximate Sparse Duality

The strong sparse duality theory developed in the previous subsection relies on certain sparse constraint qualification conditions imposed on the primal or dual optimizers as appeared in Theorem 2 and Theorem 9. Although these conditions can be partially verified for some special models such as sparse linear regression and logistic regression, they could still be restrictive and more often than not be violated in practice, leading to non-zero primal-dual gap at primal and dual optimizers. To cover the regime where sparse strong duality does not hold exactly, we further derive in the present subsection a set of primal-dual gap bounding results which only require the sparse duality holds in an approximate way. The starting point is to define the concept of *approximate sparse saddle point* which is fundamental to the subsequent analysis.

Definition 11 (Approximate sparse saddle point) *Let $k > 0$ be an integer and $\nu \geq 0$ be a scalar. A pair $(\tilde{w}, \tilde{\alpha}) \in \mathbb{R}^d \times \mathcal{F}$ is said to be a ν-approximate k-sparse saddle point for $L(w, \alpha)$ if $\|\tilde{w}\|_0 \leq k$ and the following is valid for all k-sparse vector w and $\alpha \in \mathcal{F}$:*

$$L(\tilde{w}, \alpha) \leq L(w, \tilde{\alpha}) + \nu. \tag{7}$$

Obviously, the above definition allows $L(\tilde{w}, \alpha) \leq L(\tilde{w}, \tilde{\alpha}) + \nu$ for any $\alpha \in \mathcal{F}$, and $L(\tilde{w}, \tilde{\alpha}) \leq L(w, \tilde{\alpha}) + \nu$ for any w with $\|w\|_0 \leq k$. Specially when $\nu = 0$, an approximate sparse saddle point reduces to an exact sparse saddle point. The approximate sparse saddle point can also be understood as an extension of the so called *approximate saddle point* in approximate duality theory (Scovel et al., 2007) to the setting of sparsity-constrained minimization. Based on the concept of approximate sparse saddle point, we can prove the following proposition of approximate sparse duality.

Proposition 12 *Let $\tilde{w} \in \mathbb{R}^d$ be a k-sparse primal vector and $\tilde{\alpha} \in \mathcal{F}$ be a dual vector. Then for any $\nu \geq 0$, the primal-dual gap satisfies*

$$P(\tilde{w}) - D(\tilde{\alpha}) \leq \nu$$

if and only if the pair $(\tilde{w}, \tilde{\alpha})$ admits a ν-approximate k-sparse saddle point for $L(w, \alpha)$.

Now we are going to bound the primal-dual gap under the condition of approximate sparse duality. Before presenting the main result, we need to define some notations. We say a univariate differentiable function $l(x)$ is μ-strongly convex and ℓ-smooth if $\forall x, y$,

$$\frac{\mu}{2}|x - y|^2 \leq l(y) - l(x) - \langle l'(x), y - x\rangle \leq \frac{\ell}{2}|x - y|^2.$$

The following defined sparse largest (smallest) eigenvalue of the empirical covariance matrix $\hat{\Sigma} = \frac{1}{N} X X^\top$ will be used in our analysis:

$$\gamma_s^+ = \max_{v \in \mathbb{R}^d} \left\{ v^\top \hat{\Sigma} v \mid \|v\|_0 \leq s, \|v\| = 1 \right\},$$
$$\gamma_s^- = \min_{v \in \mathbb{R}^d} \left\{ v^\top \hat{\Sigma} v \mid \|v\|_0 \leq s, \|v\| = 1 \right\}.$$

Let us re-express the primal objective function as

$$P(w) = f(w) + \frac{\lambda}{2}\|w\|^2, \quad \text{where } f(w) := \frac{1}{N}\sum_{i=1}^N l_i(w^\top x_i).$$

We show in the following theorem that the primal-dual optimizer pair $(\bar{w}, \bar{\alpha})$ admits an approximate sparse saddle point with approximation level controlled by the underlying statistical error of model.

Theorem 13 *Assume that the primal loss functions l_i are μ-strongly convex and ℓ-smooth. Let $\tilde{w}$ be any k-sparse vector satisfying $\tilde{w}_{\min} > \frac{\|\nabla f(\tilde{w})\|_\infty}{\ell \gamma_k^+}$. Assume that $\lambda \leq \frac{\mu\gamma_k^- \sqrt{k}\|\nabla f(\tilde{w})\|_\infty}{\ell\gamma_k^+ \|\tilde{w}\| - \sqrt{k}\|\nabla f(\tilde{w})\|_\infty}$. Then*

$$P(\bar{w}) \leq D(\bar{\alpha}) + \frac{k}{\lambda}\left(2 + \frac{\ell\gamma_k^+}{\mu\gamma_k^- + \lambda}\right)^2 \|\nabla f(\tilde{w})\|_\infty^2,$$

where $\bar{w}$ is the primal minimizer of $P(x)$ and $\bar{\alpha}$ is the dual maximizer of $D(\alpha)$.

Remark 14 *Specially, by setting* $\lambda = \frac{\mu\gamma_k^- \sqrt{k}\|\nabla f(\tilde{w})\|_\infty}{\ell\gamma_k^+ \|\tilde{w}\|} > 0$*, we get from the above theorem that*

$$P(\bar{w}) - D(\bar{\alpha}) \leq \frac{\ell\gamma_k^+ \|\tilde{w}\|\sqrt{k}}{\mu\gamma_k^-} \left(2 + \frac{\ell\gamma_k^+}{\mu\gamma_k^-}\right)^2 \|\nabla f(\tilde{w})\|_\infty.$$

This bound shows that the optimal primal-dual gap at $(\bar{w}, \bar{\alpha})$ *is controlled by the approximation level* $\nu = \mathcal{O}(\sqrt{k}\|\nabla f(\tilde{w})\|_\infty)$ *which usually represents the statistical estimation error of a nominal vector* $\tilde{w}$*. The smaller* $\|\nabla f(\tilde{w})\|_\infty$ *is, the more accurate approximation will be. This theorem, however, does not provide any guarantee on the sub-optimality of the primal solution* $w(\bar{\alpha}) = \mathrm{H}_k\left(-\frac{1}{\lambda N}\sum_{i=1}^N \bar{\alpha}_i x_i\right)$ *produced from the dual maximizer* $\bar{\alpha}$*. We leave this prima-dual connection issue as an open problem for future investigation. In any case, this prima-dual gap bound can be a useful tool for getting a rough idea of how well the dual formulation can capture the optimal primal objective value.*

In the following, we show the implications of Theorem 13 for the sparse linear regression and logistic regression models.

Approximate sparse duality for linear regression. Given a k-sparse parameter vector $\tilde{w}$, assume the samples are generated according to the linear model $y = \tilde{w}^\top x + \varepsilon$ where ε is a zero-mean Gaussian random noise variable with parameter σ. Let $l_i(w^\top x_i) = \frac{1}{2}(y_i - w^\top x_i)^2$ be the least square loss over data sample (x_i, y_i). In this example, we have $\ell = \mu = 1$. Suppose x_i are drawn from Gaussian distribution with covariance Σ. Then with overwhelming probability we have $\gamma_k^- \geq \lambda_{\min}(\Sigma) - \mathcal{O}(k\log(d)/N)$ and $\gamma_k^+ \leq \max_i \|x_i\|$; and $\|\nabla f(\tilde{w})\|_\infty = \mathcal{O}\left(\sigma\sqrt{\log(d)/N}\right)$. Thus according to Theorem 13, by setting the regularization parameter $\lambda = \mathcal{O}\left(\frac{\sigma\gamma_k^-}{\gamma_k^+\|\tilde{w}\|_\infty}\sqrt{\log(d)/N}\right)$, the primal-dual gap can be upper bounded with high probability as $P(\bar{w}) - D(\bar{\alpha}) = \mathcal{O}\left(\sigma\sqrt{k\log(d)/N}\right)$.

Approximate sparse duality for logistic regression. Suppose x_i are sub-Gaussian with parameter σ in logistic regression model. It is known that $\|\nabla f(\tilde{w})\|_\infty = \mathcal{O}\left(\sigma\sqrt{\log(d)/N}\right)$ hold with high probability (Yuan et al., 2018). Also it is well known that the binary logistic loss $l_i(u) = \log\left(1 + \exp(-y_i u)\right)$ is ℓ-smooth with $\ell = 1/4$. By assuming without loss of generality that $|u| \leq r$ we can verify that it is also μ-strongly convex with $\mu = \exp(r)/(1+\exp(r))^2$. Then according to the bound in Theorem 13, by setting the regularization parameter $\lambda = \mathcal{O}\left(\sigma\sqrt{\log(d)/N}\right)$, the primal-dual gap can be upper bounded with high probability as $P(\bar{w}) - D(\bar{\alpha}) = \mathcal{O}\left(\sigma\sqrt{k\log(d)/N}\right)$.

We remark that the sparse duality theory developed in this section suggests a natural way for finding the global minimum of the sparsity-constrained minimization problem in (1) via maximizing its dual problem in (6). Particularly in the case when the strong sparse duality holds, once the dual maximizer $\bar{\alpha}$ is estimated, the primal sparse minimizer $\bar{w}$ can then be recovered from it according to the prima-dual connection $\bar{w} = \mathrm{H}_k\left(-\frac{1}{\lambda N}\sum_{i=1}^N \bar{\alpha}_i x_i\right)$. Since the dual objective function $D(\alpha)$ is shown to be concave, its global maximum can be estimated using off-the-shelf convex/concave optimization methods. In the next section, we present a simple projected super-gradient method to solve the dual maximization problem with strong guarantees on convergence and sparsity recovery.

4. Algorithms

Let us now consider the dual maximization problem (6) which can be expressed as

$$\max_{\alpha \in \mathcal{F}} D(\alpha) = \frac{1}{N} \sum_{i=1}^{N} -l_i^*(\alpha_i) - \frac{\lambda}{2} \|w(\alpha)\|^2, \tag{8}$$

where $w(\alpha) = \mathrm{H}_k\left(-\frac{1}{\lambda N} \sum_{i=1}^{N} \alpha_i x_i\right)$. Generally speaking, $D(\alpha)$ is a non-smooth function because: 1) the conjugate function l_i^* of an arbitrary convex loss l_i is generally non-smooth and 2) the term $\|w(\alpha)\|^2$ is non-smooth with respect to α due to the truncation operation involved in computing $w(\alpha)$. We propose to adopt projected sub-gradient-type methods to solve the constrained non-smooth dual maximization problem in (6).

4.1 The DIHT Algorithm

The Dual Iterative Hard Thresholding (DIHT) algorithm, as outlined in Algorithm 1, is essentially a projected super-gradient method for maximizing $D(\alpha)$. Initialized with $w^{(0)} = 0$ and $\alpha^{(0)} = 0$, the procedure generates a sequence of prima-dual pairs $\{(w^{(t)}, \alpha^{(t)})\}_{t \geq 1}$. At the t-th iteration, the dual update step **S1** conducts the projected super-gradient ascent in (9) to update $\alpha^{(t)}$ from $\alpha^{(t-1)}$ and $w^{(t-1)}$. Then in the primal update step **S2**, the primal vector $w^{(t)}$ is constructed from $\alpha^{(t)}$ based on a k-sparse primal-dual connection operator (10).

Algorithm 1: Dual Iterative Hard Thresholding (DIHT)

Input : Training set $\{x_i, y_i\}_{i=1}^N$. Regularization strength λ. Sparsity level k.
Initialization $w^{(0)} = 0$, $\alpha_1^{(0)} = ... = \alpha_N^{(0)} = 0$.
for $t = 1, 2, ..., T$ **do**

`/* Dual projected super-gradient ascent */`

(**S1**) For all $i \in \{1, 2, ..., N\}$, update the dual variables $\alpha_i^{(t)}$ as

$$\alpha_i^{(t)} = \mathrm{P}_{\mathcal{F}_i}\left(\alpha_i^{(t-1)} + \eta^{(t-1)} g_i^{(t-1)}\right), \tag{9}$$

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

`/* Primal hard thresholding */`

(**S2**) Update the primal vector $w^{(t)}$ as:

$$w^{(t)} = \mathrm{H}_k\left(-\frac{1}{\lambda N} \sum_{i=1}^{N} \alpha_i^{(t)} x_i\right). \tag{10}$$

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

In the following we analyze the non-asymptotic convergence behavior of DIHT. We denote $\bar{w} = \arg\min_{\|w\|_0 \leq k} P(w)$ and use the abbreviation $\epsilon_{PD}^{(t)} := \epsilon_{PD}(w^{(t)}, \alpha^{(t)})$. In order

to avoid technical complications, we will limit optimization to bounded dual feasible sets $\mathcal{F}_i$ and derivatives $l_i^{*'}$, i.e., we will let $r = \max_{i,a\in\mathcal{F}_i} |a|$ and $\rho = \max_{i,a\in\mathcal{F}_i} |l_i^{*'}(a)|$. For example, such quantities exist when l_i and l_i^* are Lipschitz continuous (Shalev-Shwartz and Zhang, 2013b). We assume without loss of generality that $\|x_i\| \leq 1$. In the following theorem, we show that DIHT converges sub-linearly in dual parameter estimation error and primal-dual gap, and exact sparsity recovery can be guaranteed after sufficient iteration.

Theorem 15 *Assume the primal loss functions $l_i(\cdot)$ are $1/\mu$-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.** *Let $\bar{\alpha} = [l_1'(\bar{w}^\top x_1), ..., l_N'(\bar{w}^\top x_N)]$. Then the sequence $\{\alpha^{(t)}\}_{t\geq 1}$ generated by Algorithm 1 satisfies*

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

Moreover the primal-dual gap is bounded as

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

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

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

Remark 16 *To gain some intuition of the bounds in the theorem, if conventionally choosing the regularization parameter $\lambda \propto \frac{1}{\sqrt{N}}$, then $\frac{1}{\sqrt{N}}\|\alpha^{(t)} - \bar{\alpha}\| = \mathcal{O}\left(\frac{1}{\sqrt{t}}\right)$, $\epsilon_{PD}^{(t)} = \mathcal{O}\left(\frac{1}{\sqrt{t}}\right)$ and $supp(w^{(t)}) = supp(\bar{w})$ is guaranteed after $\mathcal{O}\left(\frac{1}{\bar{\epsilon}^2}\right)$ steps of iteration. Theorem 15 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.*

Next, we consider the case when l_i are simultaneously smooth and strongly convex, which can lead to improved linear rate of convergence as stated in the following theorem.

Theorem 17 *Suppose that the loss functions $l_i(\cdot)$ are $1/\mu$-smooth and $1/\ell$-strongly-convex. Assume that $\bar{\epsilon} := \bar{w}_{\min} - \frac{1}{\lambda}\|P'(\bar{w})\|_\infty > 0$. Let $\ell_D = \left(\frac{\sqrt{2}\|X\|}{\lambda N\sqrt{N}}\left(1 + \frac{4\|X\|\|\bar{\alpha}\|}{\lambda N\bar{\epsilon}}\right) + \frac{\sqrt{2}\ell}{N}\right)$ and $\mu_D = \frac{\mu}{N}$. Set the step-size as $\eta^{(t)} = \frac{\mu_D}{\ell_D^2}$.*

(a) **Parameter estimation error and primal-dual gap.** *Let $\bar{\alpha} = [l_1'(\bar{w}^\top x_1), ..., l_N'(\bar{w}^\top x_N)]$. Then the sequence $\{\alpha^{(t)}\}_{t\geq 1}$ generated by Algorithm 1 satisfies*

$$\|\alpha^{(t)} - \bar{\alpha}\|^2 \leq \|\bar{\alpha}\|^2\left(1 - \frac{\mu_D^2}{\ell_D^2}\right)^t.$$

Moreover, the primal-dual gap $\epsilon_{PD}^{(t)}$ is upper bounded as

$$\epsilon_{PD}^{(t)} \leq \ell_D\|\bar{\alpha}\|^2\left(\frac{\|X\|}{\lambda\mu\sqrt{N}}\left(1 + \frac{4\|X\|\|\bar{\alpha}\|}{\lambda N\bar{\epsilon}}\right) + 1\right)\left(1 - \frac{\mu_D^2}{\ell_D^2}\right)^t.$$

(b) **Sparsity recovery.** *The exact support recovery, i.e., $supp(w^{(t)}) = supp(\bar{w})$, occurs after*

$$t \geq \left\lceil \frac{\ell_D^2}{\mu_D^2} \log\left(\frac{4\|\bar{\alpha}\|^2 \|X\|^2}{\lambda^2 N^2 \bar{\epsilon}^2} \right) \right\rceil$$

rounds of iteration.

Remark 18 *Consider setting the regularization parameter as $\lambda \propto \frac{1}{\sqrt{N}}$. Then the contraction factor $\frac{\mu_D^2}{\ell_D^2}$ is of the order $\mathcal{O}\left(\frac{\mu^2}{(\|X\|+\ell)^2}\right)$, and $supp(w^{(t)}) = supp(\bar{w})$ can be guaranteed after $\mathcal{O}\left(\frac{(\|X\|+\ell)^2}{\mu^2} \log\left(\frac{1}{\bar{\epsilon}}\right)\right)$ steps of iteration using step-size $\eta^{(t)} = \mathcal{O}\left(\frac{N\mu}{(\|X\|+\ell)^2}\right)$. For the special case of linear regression with $l_i(u) = \frac{1}{2}(u - y_i)^2$ and $l_i^*(\alpha_i) = \frac{\alpha_i^2}{2} + y_i\alpha_i$, we have $\mu = \ell = 1$, hence the contraction factor is of the order $\mathcal{O}\left(\frac{1}{\|X\|}\right)$ and support recovery can be guaranteed after $t = \mathcal{O}\left(\|X\| \log\left(\frac{1}{\bar{\epsilon}}\right)\right)$ rounds of iteration with $\eta^{(t)} = \mathcal{O}\left(\frac{N}{\|X\|}\right)$.*

Regarding the 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 15 and Theorem 17 are directly applicable to the primal sub-optimality. We comment that under the sparse strong duality conditions, our convergence results on $\epsilon_P^{(t)}$ are not explicitly relying on the Restricted Isometry Property (RIP) (or restricted strong condition number condition) which is required in most existing analysis of primal IHT-style algorithms (Blumensath and Davies, 2009; Bahmani et al., 2013; Yuan et al., 2014). It was shown by Jain et al. (2014) that with proper relaxation of sparsity level, the estimation consistency of IHT-style algorithms can be guaranteed without imposing RIP-type conditions. In contrast to these prior analysis, our RIP-free results in Theorem 15 and Theorem 17 do not require the sparsity level k to be relaxed.

4.2 Stochastic DIHT

When a batch estimation of super-gradient $D'(\alpha)$ becomes expensive in large-scale applications, it is optional to consider the stochastic implementation of DIHT, namely SDIHT, as outlined in Algorithm 2. Different from the batch computation in Algorithm 1, the dual update step **S1** in Algorithm 2 randomly selects a block of samples (from a given block partition of samples) and update their corresponding dual variables according to (11). 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 where $m = 1$, SDIHT reduces to the batch DIHT. At the opposite extreme end where $m = N$, i.e., each block contains one sample, SDIHT becomes a stochastic coordinate-wise super-gradient method.

The dual update (11) in SDIHT is computationally more efficient than full DIHT as the former only needs to access a small subset of samples at a time. If the complexity of hard thresholding operation in primal update is not negligible as 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.

Algorithm 2: Stochastic Dual Iterative Hard Thresholding (SDIHT)

Input : Training set $\{x_i, y_i\}_{i=1}^N$. Regularization strength λ. Sparsity level k. A block disjoint partition $\{B_1, ..., B_m\}$ of the sample index set $[N]$.

Initialization $w^{(0)} = \tilde{w}^{(0)} = 0$, $\alpha_{B_1}^{(0)} = ... = \alpha_{B_m}^{(0)} = 0$.

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

`/* Stochastic blockwise dual projected super-gradient ascent */`

(**S1**) Uniformly randomly select a block index $i^{(t)} \in [m]$. For all $j \in B_{i^{(t)}}$ update $\alpha_j^{(t)}$ as

$$\alpha_j^{(t)} = \mathrm{P}_{\mathcal{F}_j}\left(\alpha_j^{(t-1)} + \eta^{(t-1)} g_j^{(t-1)}\right), \tag{11}$$

and set $\alpha_j^{(t)} = \alpha_j^{(t-1)}$, $\forall j \notin B_{i^{(t)}}$.

`/* Primal hard thresholding */`

(**S2**) Update the primal vector $w^{(t)}$ via the following operations:

– (**S2.1**) Update $\tilde{w}^{(t)}$ according to

$$\tilde{w}^{(t)} = \tilde{w}^{(t-1)} - \frac{1}{\lambda N} \sum_{j \in B_{i^{(t)}}} (\alpha_j^{(t)} - \alpha_j^{(t-1)}) x_j. \tag{12}$$

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

end

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

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

Theorem 19 *Assume that the primal loss functions $l_i(\cdot)$ are $1/\mu$-smooth and $\bar{\epsilon} := \bar{w}_{\min} - \frac{1}{\lambda}\|P'(\bar{w})\|_\infty > 0$. Set the step-size as $\eta^{(t)} = \frac{mN}{\mu(t+2)}$.*

(a) **Parameter estimation error and primal-dual gap**. *Let $\bar{\alpha} = [l_1'(\bar{w}^\top x_1), ..., l_N'(\bar{w}^\top x_N)]$. The sequence $\{\alpha^{(t)}\}_{t \geq 1}$ generated by Algorithm 2 satisfies*

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

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

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

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

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

rounds of iteration.

Remark 20 *Theorem 19 indicates that up to scaling factors, the expected iteration complexity of SDIHT is identical to that of DIHT. The additional scaling factors m or $\sqrt{m}$ appeared in the bounds essentially reflect a trade-off between the decreased per-iteration computational cost and the increased iteration complexity.*

Remark 21 *The part(b) of Theorem 19 states that $supp(w^{(t)}) = supp(\bar{w})$ occurs with high probability when t is sufficiently large. When this event occurs, SDITH (with $m = N$) reduces to a restricted version of SDCA (Shalev-Shwartz and Zhang, 2013b) 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 stage convergence behavior of SDIHT after exact support recovery is not of primal interest of this work.*

When the primal loss functions are smooth and strongly convex, we can further generalize the linear convergence rates in Theorem 17 from batch DIHT to SDIHT, as formally summarized in the following theorem.

Theorem 22 *Assume that the loss functions $l_i(\cdot)$ are $1/\mu$-smooth and $1/\ell$-strongly-convex. Assume that $\bar{\epsilon} := \bar{w}_{\min} - \frac{1}{\lambda}\|P'(\bar{w})\|_\infty > 0$. Let $\ell_D = \left(\frac{\sqrt{2}\|X\|}{\lambda N\sqrt{N}}\left(1 + \frac{4\|X\|\|\bar{\alpha}\|}{\lambda N\bar{\epsilon}}\right) + \frac{\sqrt{2}\ell}{N}\right)$ and $\mu_D = \frac{\mu}{N}$. Set the step-size as $\eta^{(t)} = \frac{\mu_D}{\ell_D^2}$.*

(a) **Parameter estimation error and primal-dual gap.** *Let $\bar{\alpha} = [l_1'(\bar{w}^\top x_1), ..., l_N'(\bar{w}^\top x_N)]$. The sequence $\{\alpha^{(t)}\}_{t\geq 1}$ generated by Algorithm 2 satisfies*

$$\mathbb{E}[\|\alpha^{(t)} - \bar{\alpha}\|^2] \leq \|\bar{\alpha}\|^2\left(1 - \frac{\mu_D^2}{m\ell_D^2}\right)^t.$$

Moreover, the primal-dual gap $\epsilon_{PD}^{(t)}$ is bounded in expectation as

$$\mathbb{E}[\epsilon_{PD}^{(t)}] \leq \ell_D\|\bar{\alpha}\|^2\left(\frac{\|X\|}{\lambda\mu\sqrt{N}}\left(1 + \frac{4\|X\|\|\bar{\alpha}\|}{\lambda N\bar{\epsilon}}\right) + 1\right)\left(1 - \frac{\mu_D^2}{m\ell_D^2}\right)^t.$$

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

$$t \geq \left\lceil \frac{m\ell_D^2}{\mu_D^2}\log\left(\frac{4\|\bar{\alpha}\|^2\|X\|^2}{\delta^2\lambda^2N^2\bar{\epsilon}^2}\right)\right\rceil$$

rounds of iteration.

Remark 23 *Like in Theorem 19, the scaling factor m appeared in the contraction factors represents a trade-off between the reduced per-iteration computational cost and the increased iteration complexity.*

4.3 Comparison against Primal IHT Methods

We now compare DIHT and SDIHT in theory with several representative primal IHT-style algorithms for sparse estimation. Here, we use primal ϵ-sub-optimality as the metric of performance. Specifically, we compare the considered algorithms in terms of RIP-type condition, sparsity level relaxation condition, and incremental first-order oracle (IFO)[2] complexity for achieving the primal sub-optimality $P(\tilde{w}) - P(\bar{w}) \leq \epsilon$, where $\tilde{w}$ is the k-sparse estimator and $\bar{w}$ is the target $\bar{k}$-sparse primal minimizer with $\bar{k} \leq k$. Table 2 summarizes the comparison results in the setting where the univariate loss functions $l_i(\cdot)$ are $1/\mu$-smooth and $1/\ell$-strongly-convex. In the following elaboration, we highlight the key observations that can be made from these results.

DIHT versus primal full gradient IHT methods. Since DIHT is a full gradient hard-thresholding method, we compare it with several popularly studied full gradient primal IHT algorithms including GraSP (Bahmani et al., 2013), IHT (Jain et al., 2014) and GraHTP (Yuan et al., 2018). The comparison results are summarized on the top panel of Table 2, from which we can observe that: 1) Different from the considered full gradient IHT algorithms which are either relying on RIP-type bounding conditions on κ_s or requiring relaxation $k = \Omega(\kappa_s^2 \bar{k})$ on sparsity level, DIHT is free of explicitly assuming RIP-type conditions and sparsity level relaxation; 2) In terms of IFO complexity, based on Theorem 17 we can verify that DIHT needs $\mathcal{O}\left(\left(\frac{N\ell^2}{\mu^2} + \frac{1}{\lambda^2}\left(1 + \frac{1}{\lambda N}\right)^2\right)\log\left(\frac{1}{\epsilon}\right)\right)$ IFO queries to achieve primal ϵ-sub-optimality. This should be superior to the considered primal algorithms when $\lambda \propto \frac{1}{\sqrt{N}}$ and $\kappa_s \gg \frac{\ell^2}{\mu^2}$ which is expected to be the case in ill-conditioned problems where κ_s scales up quickly with data size and sparsity level while ℓ, μ typically do not.

SDIHT versus primal stochastic gradient IHT methods. In order to improve computational efficiency, stochastic gradient IHT algorithms have recently been developed via leveraging the finite-sum structure of statistical learning problems. We further compare SDIHT against several state-of-the-art stochastic gradient IHT algorithms including StoIHT (Nguyen et al., 2017), SVR-GHT (Li et al., 2016) and HSG-HT (Zhou et al., 2018). At each iteration, StoIHT only evaluates gradient of one (or a mini-batch) randomly selected sample for variable update and hard thresholding. Although efficient in iteration, StoIHT can only be shown to converge to a sub-optimal statistical estimation accuracy which is inferior to that of the full-gradient methods. Another limitation of StoIHT is that it requires the restricted condition number κ_s to be not larger than $4/3$ which is hard to meet in realistic high-dimensional sparse estimation problems such as sparse linear regression (Jain et al., 2014). The SVR-GHT algorithm (Li et al., 2016) was developed as an adaptation of SVRG (Johnson and Zhang, 2013) to the iterative hard thresholding methods. Benefiting from the variance reduced technique, SVR-GHT can converge more stably and efficiently, while allowing arbitrary bounded restricted condition number at the cost of sparsity level relaxation. Lately, Zhou et al. (2018) proposed HSG-HT as a stochastic-deterministic hybrid stochastic gradient hard thresholding method that can be provably shown to have sample-size-independent IFO complexity. From the bottom panel of Table 2 we can see that in

2. For the primal objective $P(w)$ in problem (1), an IFO takes a data point (x_i, y_i) and returns the pair $\left\{l_i(w^\top x_i) + \frac{\lambda\|w\|^2}{2}, l_i'(w^\top x_i)x_i + \lambda w\right\}$. For the dual objective $D(\alpha)$ in problem (8), an IFO takes a point (x_i, y_i) and returns the pair $\left\{l_i^*(\alpha_i) + \frac{\lambda\|w(\alpha)\|^2}{2}, l_i^{*'}(\alpha_i) - x_i^\top w(\alpha)\right\}$.

	Method	RIP-Free	Sparsity level k	IFO Complexity
Full gradient	GraSP	✗	$\bar{k}$	$\mathcal{O}\left(N\kappa_s \log\left(\frac{1}{\epsilon}\right)\right)$
	IHT	✓	$\Omega\left(\kappa_s^2\bar{k}\right)$	$\mathcal{O}\left(N\kappa_s \log\left(\frac{1}{\epsilon}\right)\right)$
	GraHTP	✓	$\Omega\left(\kappa_s^2\bar{k}\right)$	$\mathcal{O}\left(N\kappa_s \log\left(\frac{1}{\epsilon}\right)\right)$
	DIHT (**this work**)	✓	$\bar{k}$	$\mathcal{O}\left(\left(\frac{N\ell^2}{\mu^2} + \left(\frac{1}{\lambda} + \frac{1}{\lambda^2 N}\right)^2\right) \log\left(\frac{1}{\epsilon}\right)\right)$
Stochastic gradient	StoIHT	✗	$\bar{k}$	—
	SVR-GHT	✓	$\Omega\left(\kappa_s^2\bar{k}\right)$	$\mathcal{O}\left((N+\kappa_s) \log\left(\frac{1}{\epsilon}\right)\right)$
	HSG-HT	✓	$\Omega\left(\kappa_s^2\bar{k}\right)$	$\mathcal{O}\left(\frac{\kappa_s}{\epsilon}\right)$
	SDIHT (**this work**)	✓	$\bar{k}$	$\mathcal{O}\left(\left(\frac{N\ell^2}{\mu^2} + \left(\frac{1}{\lambda} + \frac{1}{\lambda^2 N}\right)^2\right) \log\left(\frac{1}{\epsilon}\right)\right)$

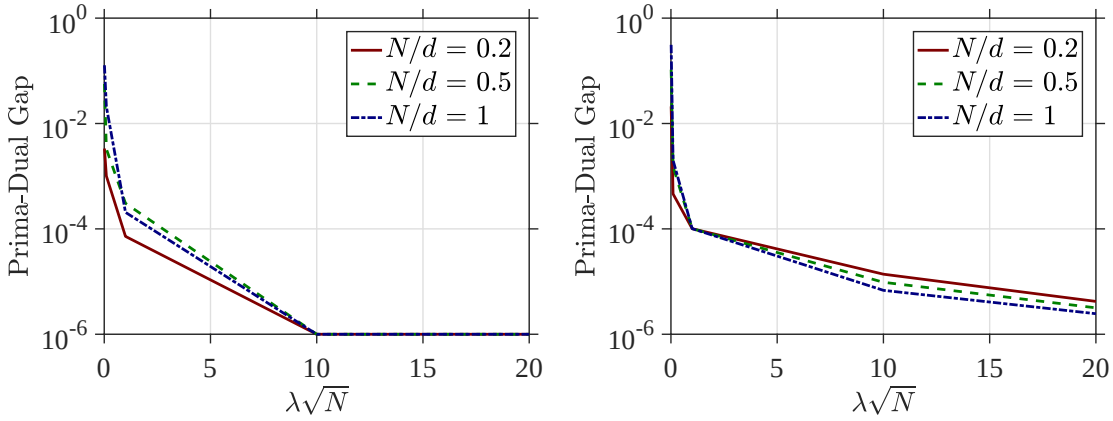
Table 2: Comparison of primal and dual IHT-style methods for achieving primal ϵ-sub-optimality in full gradient (top panel) and stochastic gradient (bottom panel) optimization settings. We denote κ_s as the restricted condition number of $P(w)$ with $s = \mathcal{O}(k + \bar{k})$. The loss functions $l_i(\cdot)$ are assumed to be $1/\mu$-smooth and $1/\ell$-strongly-convex. The mark "—" indicates that the related result is unknown in the corresponding reference.

comparison to the considered primal stochastic gradient IHT methods, SDIHT is the only one that is free of assuming RIP-type conditions while allowing the sparsity level to be unrelaxed for sparse estimation. From Theorem 22 we know that in expectation, SDIHT needs $\mathcal{O}\left(\left(\frac{N\ell^2}{\mu^2} + \frac{1}{\lambda^2}\left(1 + \frac{1}{\lambda N}\right)^2\right) \log\left(\frac{1}{\epsilon}\right)\right)$ IFO queries to achieve primal ϵ-sub-optimality, which is comparable to SVR-GHT when $\lambda \propto \frac{1}{\sqrt{N}}$ and superior to HSG-HT when the sample size N is dominated by $\frac{\kappa_s}{\epsilon}$.

To summarize the above comparison, when strong sparse duality holds, DIHT and SDIHT have provable guarantees on convergence without assuming RIP-type conditions and sparsity level relaxation conditions. The IFO complexity bounds of DIHT and SDIHT are superior or comparable to the best known results for primal IHT-style algorithms. As always, there is no free lunch here: DIHT and SDIHT are customized for solving the ℓ_2-norm regularized sparse learning problems while the primal IHT-style algorithms can be applied to more general sparse learning problems without needing to impose ℓ_2-norm penalty on the empirical risk.

5. Experiments

In this section we present numerical study for theory verification and algorithm evaluation. In the theory verification part, we conduct simulations on sparse linear regression problems to verify the strong/approximate sparse duality theorems established in Section 3. Then in the algorithm evaluation part, we run experiments on synthetic and real data sets to



(a) $d = 30$, $k = 5$ (b) $d = 500$, $k = 50$

Figure 1: Verification of strong sparse duality theory on linear regression problem: optimal primal-dual gap evolving curves as functions of regularization strength λ under different values of sample size N. For the sake of semi-log curve plotting, we set the primal-dual gap as 10^{-6} when the gap is exactly zero.

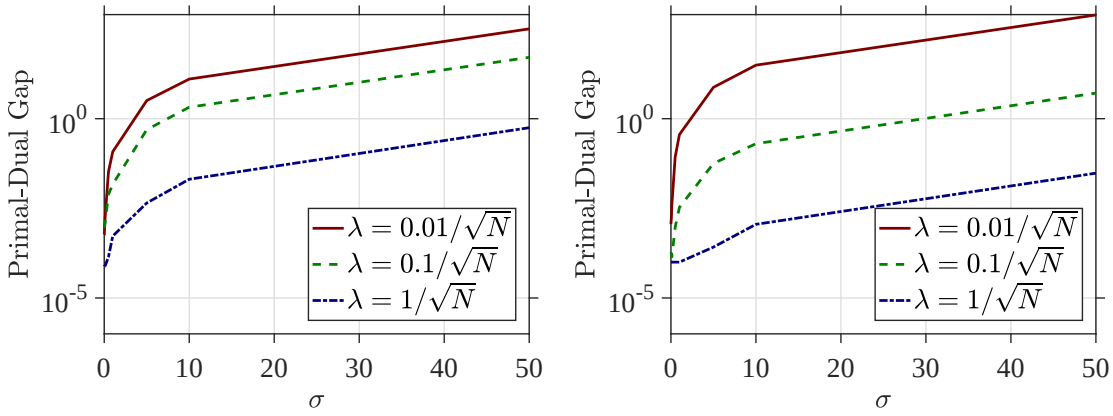
evaluate the numerical performance of DIHT and SDIHT when applied to sparse linear regression and hinge loss minimization tasks.

5.1 Theory Verification

For theory verification, we consider the sparse ridge regression model with quadratic loss function $l(y_i, w^\top x_i) = \frac{1}{2}(y_i - w^\top x_i)^2$. The feature points $\{x_i\}_{i=1}^N$ are sampled from standard multivariate normal distribution. The responses $\{y_i\}_{i=1}^N$ are generated according to a linear model $y_i = \tilde{w}^\top x_i + \varepsilon_i$ with a k-sparse parameter $\tilde{w} \in \mathbb{R}^d$ and random Gaussian noise $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. For this simulation study, we test with two baseline dimensionality-sparsity configurations $(d, k) \in \{(30, 5), (500, 50)\}$. For each configuration, we fix the parameter vector $\tilde{w}$ and study the effect of varying sample size N, regularization strength λ, and noise level σ on the optimal primal-dual gap between primal minimum and dual maximum.

5.1.1 Verification of Strong Sparse Duality Theory

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 test how the optimal primal-dual gap evolves under varying sample size N and regularization strength λ. To compute the optimal primal-dual gap, we need to find ways to estimate the primal and dual optimal values. For the configuration $(d, k) = (30, 5)$, the primal minimizer can be exactly determined via brute-force search among the optimal values over all the feasible index sets of cardinality k, and the dual maximizer is estimated via running the proposed DIHT algorithm until convergence. For $(d, k) = (500, 50)$, it becomes computationally prohibitive to compute the exact primal minimum. In this case, we just run DIHT on the dual problem until convergence and com-



(a) $d = 30$, $k = 5$ (b) $d = 500$, $k = 50$

Figure 2: Verification of approximate sparse duality theory on linear regression problem: optimal primal-dual gap evolving curves as functions of noise level σ under different regularization strength λ. Here we fix $N = d$.

pute the suboptimal primal-dual gap at the estimated dual maximizer. Figure 1 shows the (sub)optimal primal-dual gap evolving curves as functions of $\lambda\sqrt{N} \in \{10^{-2}, 10^{-1}, 1, 10, 20\}$ under different values of sample size with $N/d \in \{0.2, 0.5, 1\}$. From this group of curves we can make the following observations:

- For each curve with fixed N, the optimal primal-dual gap decreases as λ increases and the gap reaches zero when $\lambda\sqrt{N}$ 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.

- The primal-dual gap evolving curves are relatively insensitive to sample size N. This observation combined with the previous one indicates that sample size tends to have limited impact on the validness of the sparsity constraint qualification condition $\bar{w}_{\min} \geq \frac{1}{\lambda}\|P'(\bar{w})\|_\infty$.

5.1.2 Verification of Approximate Sparse Duality Theory

We further verify the approximate sparse strong duality theory stated in Theorem 13, which basically suggests that when $\bar{w}_{\min}$ is sufficiently large, by setting the regularization parameter $\lambda = \mathcal{O}\left(\sigma\sqrt{\log(d)/N}\right)$, the primal-dual gap can be upper bounded with high probability as $\epsilon_{PD} = \mathcal{O}\left(\sigma\sqrt{k\log(d)/N}\right)$. To confirm this result, fixing the sample size $N = d$, we studied how the optimal primal-dual gap evolves under varying noise level $\sigma \in [10^{-3}, 50]$ and $\lambda = \lambda_0/\sqrt{N}$ with $\lambda_0 \in \{10^{-2}, 10^{-1}, 1\}$. Figure 2 shows the optimal primal-dual gap evolving curves as functions of noise level σ under a variety of regularization strength λ. These results lead to the following observations:

- For each curve with fixed λ, the optimal primal-dual gap increases as σ increases. This confirms the implication of Theorem 13 in linear regression models that the optimal prima-dual gap of sparse linear regression model is controlled by the quantity $\sigma\sqrt{k\log d/N} \propto \sigma$.

- For a fixed σ, it can be observed that the optimal primal-dual gap approaches zero as λ increases. This again matches the prediction of Theorem 13 that the prima-dual gap bound is scaled inversely in λ.

5.2 Algorithm Evaluation

We now turn to evaluate the effectiveness and efficiency of DIHT and SDIHT for dual sparse optimization. We begin with a simulation study to confirm some theoretical properties of DIHT. Then we conduct a set of real-data experiments to demonstrate the computational efficiency of DIHT/SDIHT when applied to sparse hinge loss minimization problems.

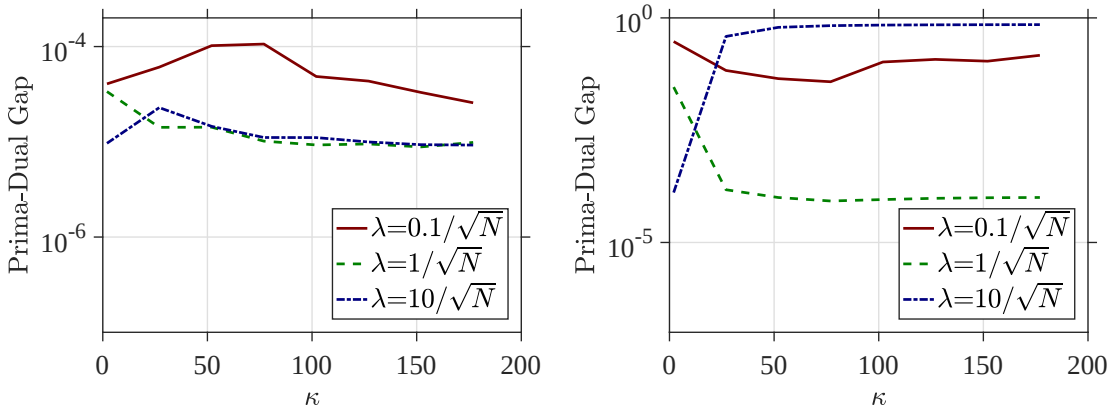
5.2.1 Simulation Study

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 (d, 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 DIHT under varying condition number. In this simulation, when λ is fixed and given a desirable condition number $\kappa > 1$, 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[3] such that the condition number of $\Sigma + \lambda I$ equals to κ. In this way of data generation, the condition number of the primal Hessian matrix $\frac{1}{N}XX^\top + \lambda I$ is close to κ. Keeping all other quantities fixed, we test how the optimal primal-dual gap output by DIHT evolves under varying $\kappa \in [1, 200]$ and regularization strength $\lambda = \lambda_0/\sqrt{N}$ for $\lambda_0 \in \{0.1, 1, 10\}$. Figure 3 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.

DIHT versus primal methods on ill-conditioned problems. We further run experiments to compare DIHT against primal IHT and HTP methods (Yuan et al., 2018; Jain et al., 2014) in high condition number setting. For this simulation study, we test with the dimensionality-sparsity configuration $(d, k) = (500, 50)$. To make the problem badly conditioned, we follow a protocol introduced by Jain et al. (2014) to select $k/2$ random coordinates from the support of nominal parameter vector $\tilde{w}$ and $k/2$ random coordinates outside its support and

3. We first generate a semi-positive definite matrix $\Sigma' \succeq 0$ such that $\lambda_{\min}(\Sigma') = 0$ and $\lambda_{\max}(\Sigma') = (2\kappa-1)\lambda$, and then we set $\Sigma = \Sigma' + \sigma I$ with $\sigma = \frac{\lambda\kappa}{\kappa-1}$. It is readily verifiable that the condition number of $\Sigma + \lambda I$ is given by $\frac{\lambda_{\max}(\Sigma)+\lambda}{\lambda_{\min}(\Sigma)+\lambda} = \frac{\lambda_{\max}(\Sigma')+\sigma+\lambda}{\lambda_{\min}(\Sigma')+\sigma+\lambda} = \frac{2\kappa\lambda+\kappa\lambda/(\kappa-1)}{\kappa\lambda/(\kappa-1)+\lambda} = \kappa$.



(a) $d = 30$, $k = 5$ (b) $d = 500$, $k = 50$

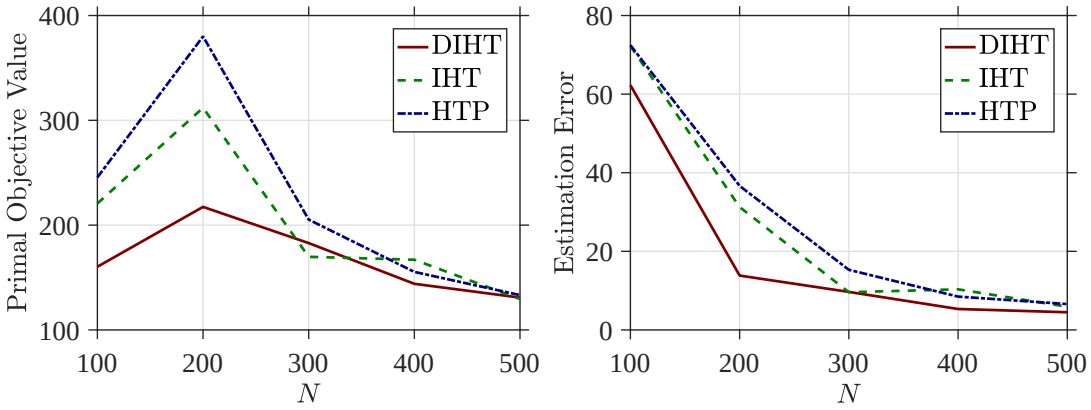
Figure 3: Convergence of DITH on linear regression problem under varying condition number: optimal primal-dual gap evolving curves as functions of condition number κ of the Hessian matrix $\frac{1}{N}XX^\top + \lambda I$, under different regularization strength λ.

constructed a covariance matrix with heavy correlations between these chosen coordinates. The condition number of the resulting matrix is around 50. Keeping all the other quantities fixed, we test how the primal objective value $P(w)$ and ℓ_2-norm parameter estimation error $\|w - \tilde{w}\|$ evolve under varying sample size $N \leq d$ and regularization strength $\lambda = \lambda_0/\sqrt{N}$ for $\lambda_0 \in \{1, 10\}$. The resulting curves are plot in Figure 4. It can be seen from these curves that in most cases DIHT is able to achieve more optimal primal objective values and smaller parameter estimation errors 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.

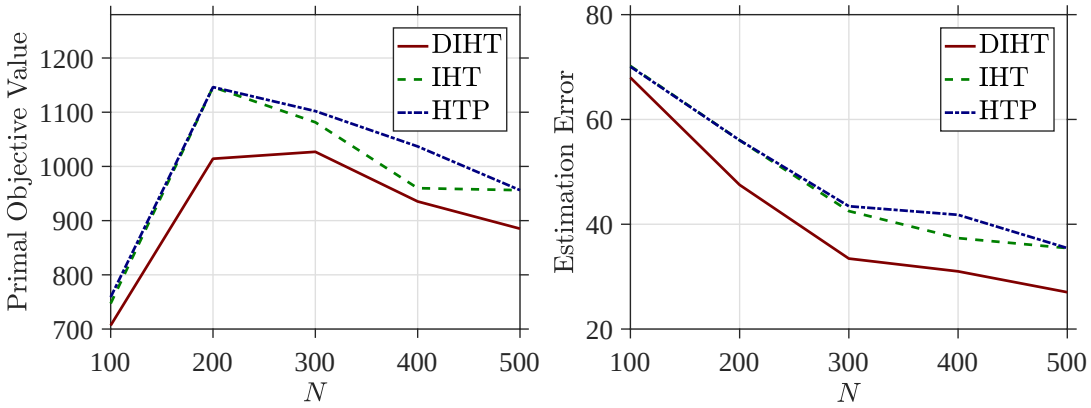
5.2.2 Real-data Experiment: Computational Efficiency Evaluation

For real data experiment, we mainly evaluate the computational efficiency of the proposed dual algorithms. We test with varying smoothed or non-smooth hinge loss functions which are commonly used by support vector machines. Two binary benchmark data sets from LibSVM data repository, RCV1 ($d = 47,236$) (Lewis et al., 2004) and News20 ($d = 1,355,191$) (Lang, 1995),[4] are used for algorithm efficiency evaluation and comparison. For the RCV1 data set, we select $N = 500,000$ ($N \gg d$) samples for model training and the rest $197,641$ samples for testing. For the News20 data set, we use $N = 15,000$ ($d \gg N$) samples for training and the left $4,996$ samples are used as test data.

4. These data sets are available at `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html`.



(a) $\lambda = 1/\sqrt{N}$



(b) $\lambda = 10/\sqrt{N}$

Figure 4: DIHT versus primal IHT-style methods on badly conditioned linear regression problem: primal objective value (left panel) and parameter estimation error (right panel) evolving curves as functions of sample size N under regularization strength (a) $\lambda = 1/\sqrt{N}$ and (b) $\lambda = 10/\sqrt{N}$, respectively.

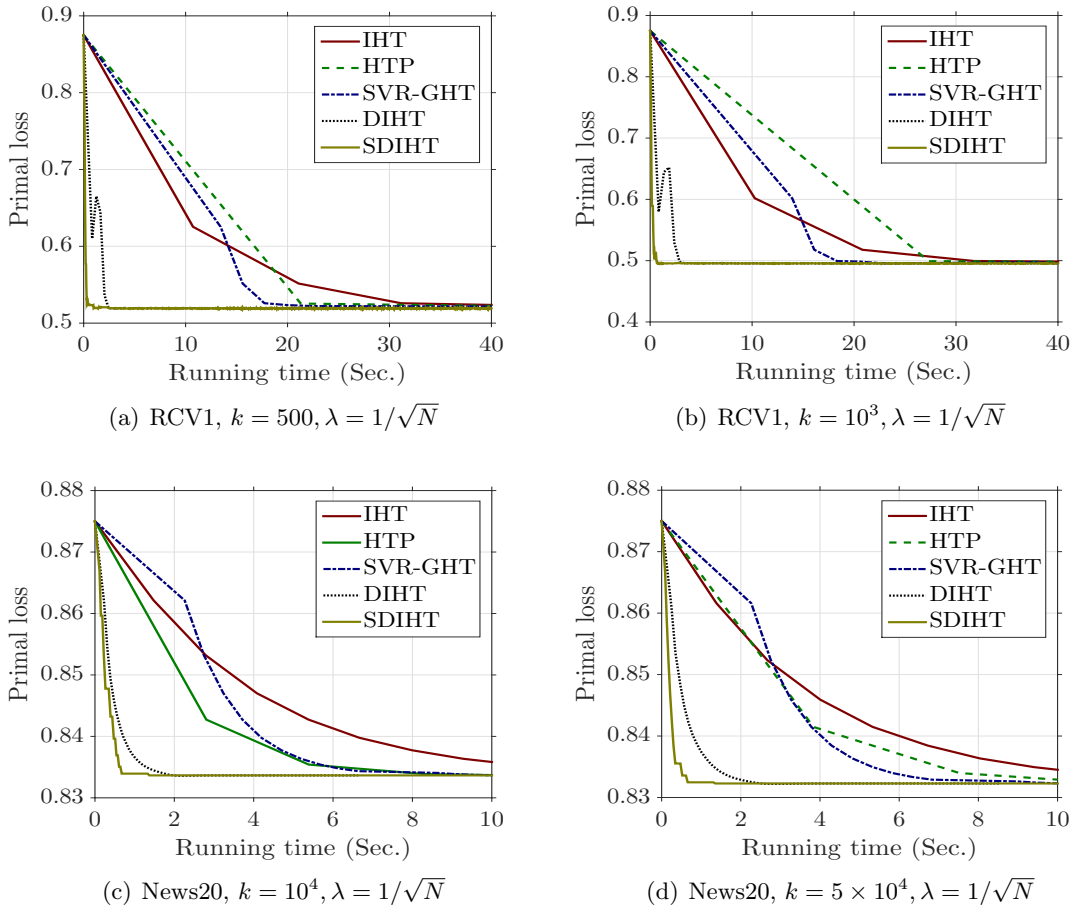
Experiment with smoothed hinge loss. We first consider the sparse learning model (1) with the following smoothed hinge loss function

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

Its convex conjugate is given by

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

We set $\gamma = 0.25$ throughout our experiment. The computational efficiency of DIHT and SDIHT is evaluated by comparing their wall-clock running time against three primal baseline



(a) RCV1, $k = 500, \lambda = 1/\sqrt{N}$

(b) RCV1, $k = 10^3, \lambda = 1/\sqrt{N}$

(c) News20, $k = 10^4, \lambda = 1/\sqrt{N}$

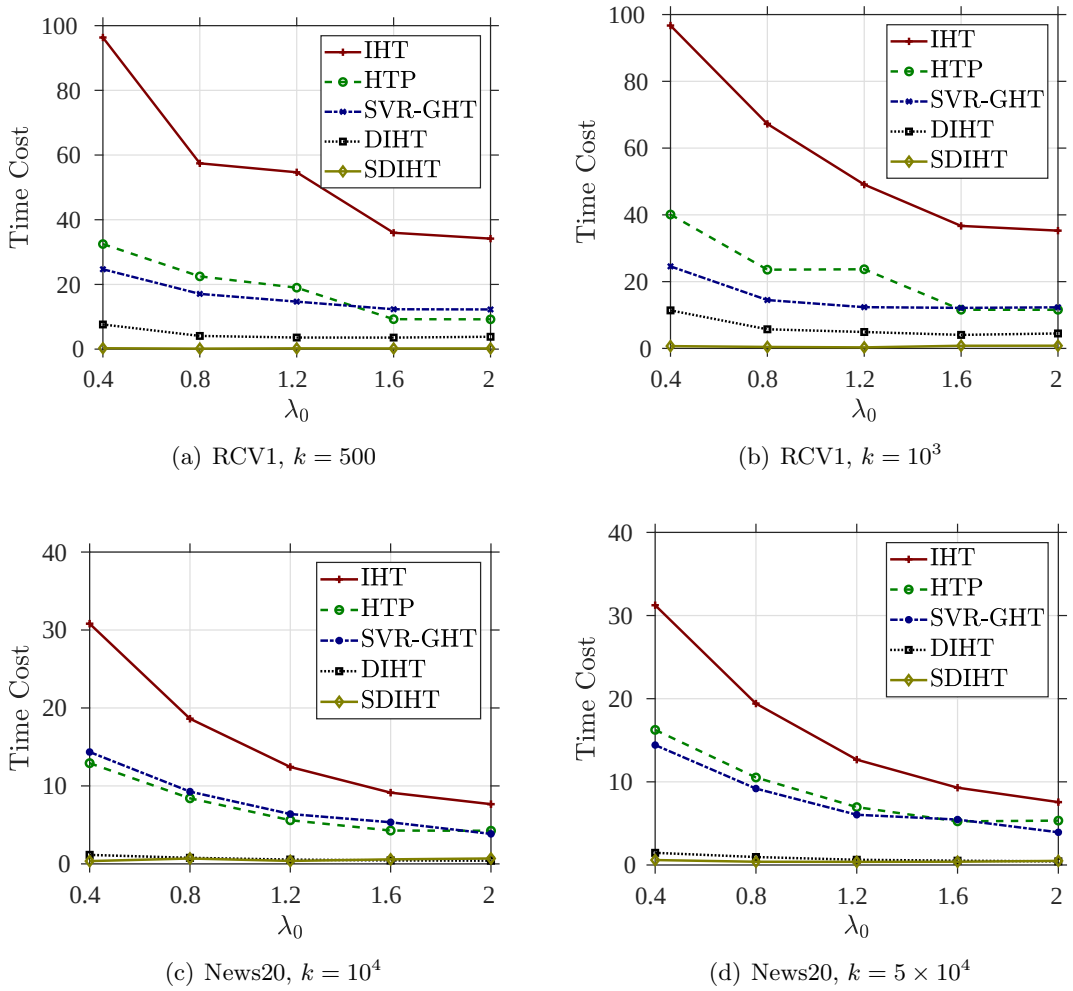
(d) News20, $k = 5 \times 10^4, \lambda = 1/\sqrt{N}$

Figure 5: Real-data experiment with smooth hinge loss: Primal loss evolving curves as functions of running time (in second).

algorithms: IHT, HTP, and SVR-GHT (Li et al., 2016) which is a stochastic variance reduced variant of IHT. The learning rates of all the considered algorithms are tuned via grid search. For the two stochastic algorithms SDIHT and SVR-GHT, the training data is uniformly randomly divided into mini-batches with batch size 10 .

Figure 5 shows the primal loss evolving curves with respect to wall-clock running time under $\lambda = 1/\sqrt{N}$ and varying sparsity level k. It can be seen from these results that under all the considered configurations of λ and k, DIHT and SDIHT outperform the considered primal IHT algorithms in minimizing the primal objective value. In the meanwhile, it can be seen that SDIHT is more efficient than DIHT which matches the consensus that stochastic dual coordinate methods often outperform their batch counterparts (Hsieh et al., 2008; Shalev-Shwartz and Zhang, 2013b).

We further compare the computational efficiency of the considered methods in terms of the training time needed to reach comparable test accuracy. We set the desirable test error as 0.08 for RCV1 and 0.24 for News20. Figure 6 shows the time cost comparison



(a) RCV1, $k = 500$

(b) RCV1, $k = 10^3$

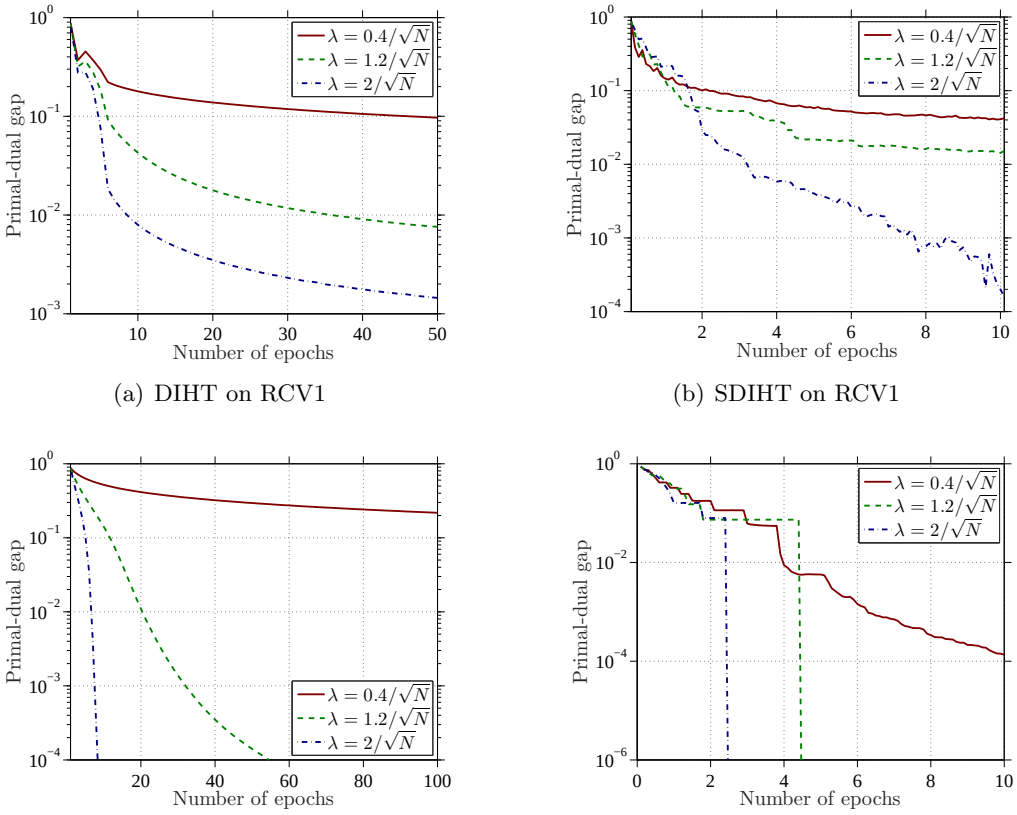
(c) News20, $k = 10^4$

(d) News20, $k = 5 \times 10^4$

Figure 6: Real-data experiment with smoothed hinge loss: Running time (in second) comparison of the considered algorithms to reach comparable test accuracy.

under varying regularization parameter $\lambda = \lambda_0/\sqrt{N}$ and sparsity level k. From this group of curves we can observe that DIHT and SDIHT are significantly more efficient than the considered primal IHT algorithms to reach comparable generalization performance on the test set. Also, we can see that SDIHT is consistently more efficient than DIHT.

Moreover, to evaluate the primal-dual convergence behavior of DIHT and SDIHT, we plot in Figure 7 their primal-dual gap evolving curves with respect to the number of epochs processing, under sparsity level $k = 10^3$ for RCV1 and $k = 5 \times 10^4$ for News20. The regularization parameters are set to be $\lambda = \lambda_0/\sqrt{N}, \lambda_0 = \{0.4, 1.2, 2\}$, respectively. The results again showcase the superior efficiency of SDIHT over DIHT as the former uses much fewer epoches of processing to reach comparable primal-dual gaps to the latter.



(a) DIHT on RCV1

(b) SDIHT on RCV1

(c) DIHT on News20

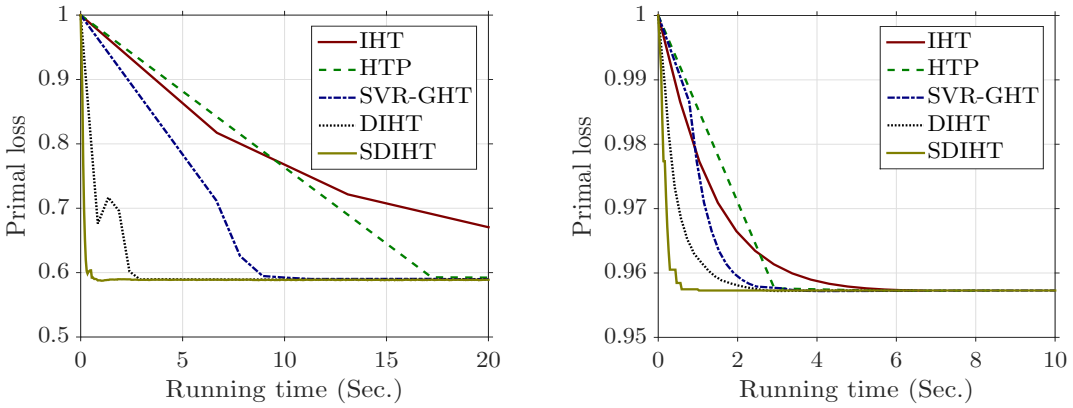
(d) SDIHT on News20

Figure 7: Real-data experiment with smoothed hinge loss: The primal-dual gap evolving curves of DIHT and SDIHT. We test with the sparsity level $k = 10^3$ for RCV1 and $k = 5 \times 10^4$ for News20.

Experiment with non-smooth hinge loss. Finally, we test the efficiency of the proposed algorithms when applied to the support vector machines with vanilla hinge loss function $l(w^\top x_i, y_i) = \max(0, 1 - y_i w^\top x_i)$. It is standard to know that

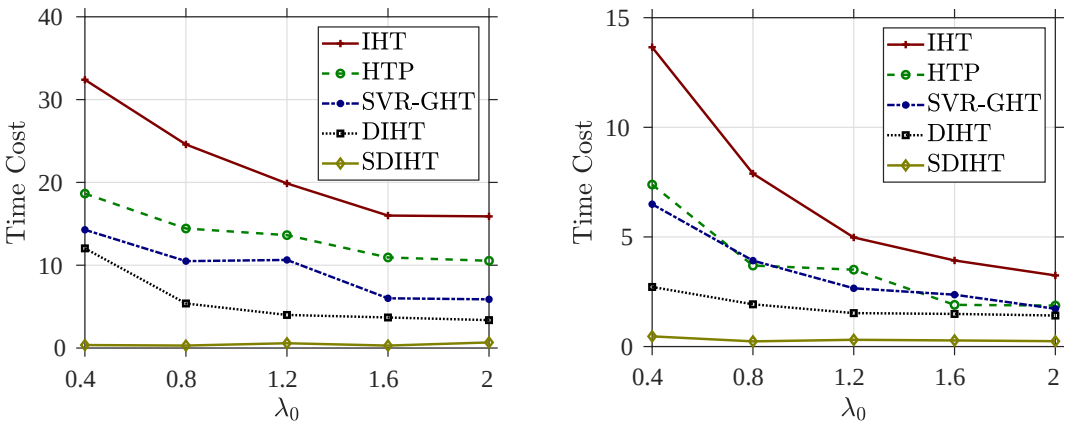
$$l^*(\alpha_i) = \begin{cases} y_i \alpha_i & \text{if } \ y_i \alpha_i \in [-1, 0] \\ +\infty & \text{otherwise} \end{cases} .$$

We follow the same experiment protocol as in the previous smoothed case to compare the considered primal and dual IHT algorithms on the two benchmark data sets. In this non-smooth case, 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. In Figure 8, we plot the primal loss evolving curves with respect to running time under $\lambda = 1/\sqrt{N}$. The computational time curves of the considered algorithms to reach comparable test errors (0.074 for RCV1 and 0.23 for News20) are shown in Figure 9. These two groups of results demonstrate the remarkable efficiency advantage of DIHT and SDIHT over the considered primal IHT algorithms even



(a) RCV1, $k = 10^3, \lambda = 1/\sqrt{N}$ (b) News20, $k = 5 \times 10^4, \lambda = 1/\sqrt{N}$

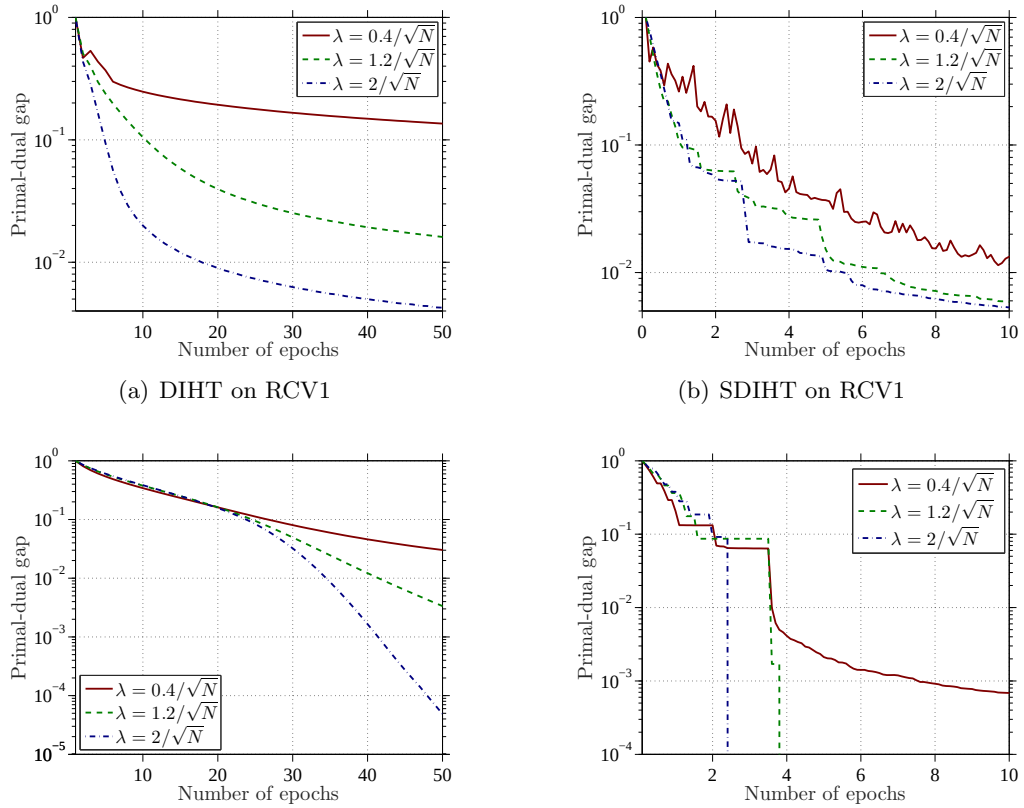
Figure 8: Real-data experiment with non-smooth hinge loss: Primal loss evolving curves as functions of running time (in second).



(a) RCV1, $k = 10^3$ (b) News20, $k = 5 \times 10^4$

Figure 9: Real-data experiment with non-smooth hinge loss: Running time (in second) comparison of the considered algorithms to reach comparable test accuracy.

when the loss function is non-smooth. The prima-dual gap evolving curves of DIHT and SDIHT under a variety of $\lambda = \lambda_0/\sqrt{N}$ are illustrated in Figure 10, from which we can observe that when using non-smooth hinge loss function, SDIHT is still more efficient than DIHT in closing the primal-dual gap.



(a) DIHT on RCV1

(b) SDIHT on RCV1

(c) DIHT on News20

(d) SDIHT on News20

Figure 10: Real-data experiment with non-smooth hinge loss: The primal-dual gap evolving curves of DIHT and SDIHT. We test with the sparsity level $k = 10^3$ for RCV1 and $k = 5 \times 10^4$ for News20.

6. Conclusion and Future Work

In this article, we investigated duality theory and optimization algorithms for solving the sparsity-constrained empirical risk minimization problem which has been widely applied in sparse learning. As a 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 scenarios where sparse strong duality would be violated, we further developed an approximate sparse duality theory that upper bounds the prima-dual gap at the level of statistical estimation error of model. Our theory opens the gate to solve the original NP-hard and 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 algorithms we have proved sub-linear primal-dual gap convergence rates when the loss is smooth, and improved

linear rates of convergence when the loss is also strongly convex. 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 generic 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 approximate duality theory (Theorem 13) only gives a duality gap bound between the (unknown) primal minimizer $\bar{w}$ and the dual maximizer $\bar{\alpha}$. From the perspective of primal solution quality certification, it would be more informative to have results on the duality gap between $\bar{\alpha}$ and the primal vector $w(\bar{\alpha})$ produced from $\bar{\alpha}$. Or third, our convergence results in Theorem 19 and 22 merely indicate that SDIHT 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 (Shalev-Shwartz and Zhang, 2013b). In our opinion, a main challenge here we are facing with is the non-smoothness of the dual objective $D(\alpha)$, which prevents us from directly extending the analysis of SDCA to SDIHT. We need to develop 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 distributed machine learning (Jaggi et al., 2014; Wang et al., 2017; Liu et al., 2019).

Acknowledgements

The authors sincerely thank the anonymous reviewers for their constructive comments on this work. Xiao-Tong Yuan is supported in part by National Major Project of China for New Generation of AI under Grant No.2018AAA0100400 and in part by Natural Science Foundation of China (NSFC) under Grant No.61876090 and No.61936005. Qingshan Liu is supported by NSFC under Grant No.61532009 and No.61825601.

Appendix A. Proofs of Results in Section 3

In this section, we present the proofs of the main results stated in Section 3.

A.1 Proof of Theorem 2

Proof The "⇐" direction: If the pair $(\bar{w}, \bar{\alpha})$ is a sparse saddle point for L, then from the definition of conjugate convexity and inequality (3) 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 $\alpha \in \mathcal{F}$

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

combining the preceding two inequalities yields

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

Therefore $P(\bar{w}) = \min_{\|w\|_0 \leq k} P(w)$, i.e., $\bar{w}$ solves the problem in (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 $\bar{\alpha}_i \in \partial l_i(\bar{w}^\top x_i)$. Thus the necessary condition (b) holds. Note that

$$L(w, \bar{\alpha}) = \frac{\lambda}{2} \left\| w + \frac{1}{\lambda N} \sum_{i=1}^{N} \bar{\alpha}_i x_i \right\|^2 - \frac{1}{N} \sum_{i=1}^{N} l_i^*(\bar{\alpha}_i) - \frac{1}{2\lambda N^2} \left(\sum_{i=1}^{N} \bar{\alpha}_i x_i \right)^2. \tag{13}$$

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

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

This validates the necessary condition (c).

The "⇒" direction: Conversely, let us assume that $\bar{w}$ is a k-sparse solution to the problem (1) (i.e., conditio(a)) and let $\bar{\alpha}_i \in \partial l_i(\bar{w}^\top x_i)$ (i.e., condition (b)). Again from the maximizing argument property of convex conjugate we know that $l_i(\bar{w}^\top x_i) = \bar{\alpha}_i \bar{w}^\top x_i - l_i^*(\bar{\alpha}_i)$. This leads to the following:

$$L(\bar{w}, \alpha) \leq P(\bar{w}) = \max_{\alpha \in \mathcal{F}} L(\bar{w}, \alpha) = L(\bar{w}, \bar{\alpha}). \tag{14}$$

The sufficient condition (c) guarantees that $\bar{F}$ contains the top k (in absolute value) entries of $-\frac{1}{\lambda N} \sum_{i=1}^{N} \bar{\alpha}_i x_i$. Then based on the expression in (13) we can see that the following holds for any k-sparse vector w

$$L(\bar{w}, \bar{\alpha}) \leq L(w, \bar{\alpha}). \tag{15}$$

By combining the inequalities (14) and (15) we obtain 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. ■

A.2 Proof of Theorem 5

Proof The "⇒" direction: 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

$$\max_{\alpha \in \mathcal{F}} \min_{\|w\|_0 \leq k} L(w, \alpha) \leq \min_{\|w\|_0 \leq k} \max_{\alpha \in \mathcal{F}} L(w, \alpha). \tag{16}$$

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

$$\begin{aligned} \min_{\|w\|_0 \leq k} \max_{\alpha \in \mathcal{F}} L(w, \alpha) &\leq \max_{\alpha \in \mathcal{F}} L(\bar{w}, \alpha) \\ &\leq L(\bar{w}, \bar{\alpha}) \leq \min_{\|w\|_0 \leq k} L(w, \bar{\alpha}) \leq \max_{\alpha \in \mathcal{F}} \min_{\|w\|_0 \leq k} L(w, \alpha). \end{aligned} \tag{17}$$

In view of (16) and (17) we have that the equality in (5) must hold.

The "⇐" direction: Assume that the equality in (5) 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 (5). 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. ■

A.3 Proof of Proposition 7

Proof Recall that

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

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

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

Thus we have

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

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 $g(\alpha'') \in \frac{1}{N}[w(\alpha'')^\top x_1 - \partial l_1^*(\alpha_1''), ..., w(\alpha'')^\top x_N - \partial l_N^*(\alpha_N'')]$. 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

$$D(\alpha') = L(w(\alpha'), \alpha') \leq L(w(\alpha''), \alpha') \leq L(w(\alpha''), \alpha'') + \left\langle g(\alpha''), \alpha' - \alpha'' \right\rangle.$$

This implies that $D(\alpha)$ is a concave function and its super-differential is given by

$$\partial D(\alpha) = \frac{1}{N}[w(\alpha)^\top x_1 - \partial l_1^*(\alpha_1), ..., w(\alpha)^\top x_N - \partial l_N^*(\alpha_N)].$$

If we further assume that $w(\alpha)$ is unique and $\{l_i^*\}_{i=1,...,N}$ are differentiable at any α, then $\partial D(\alpha) = \frac{1}{N}[w(\alpha)^\top x_1 - \partial l_1^*(\alpha_1), ..., w(\alpha)^\top x_N - \partial l_N^*(\alpha_N)]$ becomes unique, which implies that $\partial D(\alpha)$ is the unique super-gradient of $D(\alpha)$. ■

A.4 Proof of Theorem 8

Proof The "⇒" direction: 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 (6). Furthermore, Theorem 5 guarantees the following

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

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

The "⇐" direction: Assume that $\bar{\alpha}$ solves the dual problem in (6) 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 for "$\Leftarrow$" in the proof of Theorem 5 and Corollary 6 we get that the conditions (a)$\sim$(c) in Theorem 2 should be satisfied for $(\bar{w}, \bar{\alpha})$. ■

A.5 Proof of Theorem 9

Proof Recall the dual objective function is $D(\alpha) = \frac{1}{N}\sum_{i=1}^{N} -l_i^*(\alpha_i) - \frac{\lambda}{2}\|w(\alpha)\|^2$. Since $w(\bar{\alpha})$ is unique and each l_i^* is differentiable, according to Proposition 7 it is true that the super-gradient of $D(\alpha)$ at $\bar{\alpha}$ is given by $D'(\bar{\alpha}) = \frac{1}{N}[w(\bar{\alpha})^\top x_1 - l_1^{*'}(\bar{\alpha}_1), ..., w(\bar{\alpha})^\top x_N - l_N^{*'}(\bar{\alpha}_N)]$. Under the conditions in the theorem, we are going to show that for sufficiently small η, the following must hold:

$$\bar{\alpha}_i = \mathrm{P}_{\mathcal{F}_i}\left(\bar{\alpha}_i + \eta \bar{g}_i\right), \tag{18}$$

where $\bar{g}_i = \frac{1}{N}(w(\bar{\alpha})^\top x_i - l_i^{*'}(\bar{\alpha}_i))$ 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 $\tilde{w}(\alpha) = -\frac{1}{\lambda N}\sum_{i=1}^{N} \alpha_i x_i$. For a vector $x \in \mathbb{R}^d$, denote $[x]_{(j)}$ the j-th largest entry (in absolute value) of x, i.e., $|[x]_{(1)}| \geq |[x]_{(2)}| \geq ... \geq |[x]_{(d)}|$. 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 $\mathcal{B}(\bar{\alpha}) = \left\{\alpha \in \mathbb{R}^N : \|\alpha - \bar{\alpha}\| \leq \frac{\lambda N \bar{\epsilon}}{2\|X\|}\right\}$.

We prove the equation (18) by contradiction. Note that for any $\alpha \in \mathcal{B}(\bar{\alpha})$ we have

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

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 $\mathrm{supp}(w(\alpha')) = \bar{F}$. Let us assume $\alpha' \neq \bar{\alpha}$. Since l_i^* is ℓ-smooth, we have

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

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 (18) must hold for sufficiently small η.

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

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

Since (18) holds and $\mathcal{F}_i$ is convex, it must hold that either $\bar{g}_i = \frac{1}{N}(x_i^\top w(\bar{\alpha}) - l_i^{*'}(\bar{\alpha}_i)) = 0$ 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 $\frac{1}{N}\left(\alpha_i w^\top x_i - l_i^*(\alpha_i)\right)$. In any case, we always have that $\bar{\alpha}_i$ is a maximizer of $\frac{1}{N}\left(\alpha_i x_i^\top w(\bar{\alpha}) - l_i^*(\alpha_i)\right)$ 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 8 that $w(\bar{\alpha})$ admits a primal k-sparse minimizer. ■

A.6 Proof of Proposition 12

Proof The "$\Leftarrow$" direction: If the pair $(\tilde{w}, \tilde{\alpha})$ is a ν-approximate k-sparse saddle point for $L(w, \alpha)$, then from Definition 11 we can derive

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

The "$\Rightarrow$" direction: Conversely, let us assume that $P(\tilde{w}) - D(\tilde{\alpha}) \leq \nu$. Then

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

which implies that for any $\|w\|_0 \leq k$ and $\alpha \in \mathcal{F}$,

$$L(\tilde{w}, \alpha) \leq L(w, \tilde{\alpha}) + \nu.$$

Then by definition $(\tilde{w}, \tilde{\alpha})$ is a ν-approximate k-approximate saddle point of the Lagrangian $L(w, \alpha)$. This concludes the proof. ■

A.7 Proof of Theorem 13

We first introduce the following key lemma which bounds the approximation level of certain approximate sparse saddle point $(\tilde{w}, \tilde{\alpha})$ with the primal vector $\tilde{w}$ being optimal on its own supporting set.

Lemma 24 *Assume that the primal loss functions $l_i(\cdot)$ are differentiable. Let $\tilde{w} \in \mathbb{R}^d$ be a k-sparse primal vector and $\tilde{\alpha} = [l_1'(\tilde{w}^\top x_1), ..., l_N'(\tilde{w}^\top x_N)]$. Let $\tilde{F} = supp(\tilde{w})$. Assume that*

$\mathrm{H}_{\tilde{F}}(\nabla P(\tilde{w})) = 0$. *Then* $(\tilde{w}, \tilde{\alpha})$ *is a* ν*-approximate* k*-sparse saddle point of the Lagrangian* $L(w, \alpha)$ *with approximation level*

$$\nu = \frac{1}{\lambda}\|\mathrm{H}_k(\nabla f(\tilde{w}))\|^2.$$

Proof From the definition of $\tilde{\alpha}$ we know that $L(\tilde{w}, \alpha) \leq L(\tilde{w}, \tilde{\alpha})$. Recall the following formulation of $L(w, \tilde{\alpha})$:

$$L(w, \tilde{\alpha}) = \frac{\lambda}{2}\left\|w + \frac{1}{\lambda N}\sum_{i=1}^{N}\tilde{\alpha}_i x_i\right\|^2 + C = \frac{\lambda}{2}\left\|w + \frac{1}{\lambda}\nabla f(\tilde{w})\right\|^2 + C,$$

where the term C is not dependent on w. Since $\mathrm{H}_{\tilde{F}}(\nabla P(\tilde{w})) = 0$, we must have $\tilde{w} = \mathrm{H}_{\tilde{F}}(-\frac{1}{\lambda}\nabla f(\tilde{w}))$ which implies $L(\tilde{w}, \tilde{\alpha}) = \frac{1}{2\lambda}\left\|\mathrm{H}_{\tilde{F}^c}(\nabla f(\tilde{w}))\right\|^2 + C$. Let $\bar{w} = \arg\min_{\|w\|_0 \leq k} L(w, \tilde{\alpha})$ and $\bar{F} = \mathrm{supp}(\bar{w})$. Then from the first-order optimality of $\bar{w}$ we must have $L(\bar{w}, \tilde{\alpha}) = \frac{1}{2\lambda}\left\|\mathrm{H}_{\bar{F}^c}(\nabla f(\tilde{w}))\right\|^2 + C$. Therefore

$$\begin{aligned} L(\tilde{w}, \tilde{\alpha}) - L(w, \tilde{\alpha}) &\leq L(\tilde{w}, \tilde{\alpha}) - L(\bar{w}, \tilde{\alpha}) \\ &\leq \frac{1}{2\lambda}\left(\left\|\mathrm{H}_{\tilde{F}}(\nabla f(\tilde{w}))\right\|^2 + \|\mathrm{H}_{\bar{F}}(\nabla f(\tilde{w}))\|^2\right) \leq \frac{1}{\lambda}\|\mathrm{H}_k(\nabla f(\tilde{w}))\|^2. \end{aligned}$$

By combining the above arguments we get $L(\tilde{w}, \alpha) \leq L(\tilde{w}, \tilde{\alpha}) \leq L(w, \tilde{\alpha}) + \frac{1}{\lambda}\|\mathrm{H}_k(\nabla f(\tilde{w}))\|^2$, which by definition indicates that $(\tilde{w}, \tilde{\alpha})$ admits a ν-approximate sparse saddle point with $\nu = \frac{1}{\lambda}\|\mathrm{H}_k(\nabla f(\tilde{w}))\|^2$. ■

Next we introduce the concepts of smoothness and restricted strong convexity which are conventionally used in IHT-style methods (Jain et al., 2014; Yuan et al., 2018).

Definition 25 (Restricted strong convexity and smoothness) *For any integer* $s > 0$*, we say a function* $f(w) : \mathbb{R}^d \mapsto \mathbb{R}$ *is restricted* μ_s*-strongly convex for some* $m_s > 0$ *if*

$$f(w) - f(w') - \langle \nabla f(w'), w - w' \rangle \geq \frac{\mu_s}{2}\|w - w'\|^2, \quad \forall \|w - w'\|_0 \leq s. \tag{19}$$

Moreover, we say $f(w)$ *is* ℓ*-smooth for some* $\ell > 0$ *if*

$$f(w) - f(w') - \langle \nabla f(w'), w - w' \rangle \leq \frac{\ell}{2}\|w - w'\|^2, \quad \forall w, w' \in \mathbb{R}^d.$$

The ratio number ℓ/μ_s, which measures the curvature of the loss function over sparse subspaces, will be referred to as *restricted condition number*. The following is a simple lemma that summarizes some standard properties of smoothness and restricted strong convexity.

Lemma 26 *Assume that* $f(w)$ *is* μ_s*-strongly convex and* ℓ*-smooth. For any index set* F *with cardinality* $|F| \leq s$ *and any* x, y *with* $supp(w) \cup supp(w') \subseteq F$*, it holds that*

$$\|\mathrm{H}_F(\nabla f(w)) - \mathrm{H}_F(\nabla f(w'))\| \geq \mu_s\|w - w'\|, \quad \|\nabla f(w) - \nabla f(w')\| \leq \ell\|w - w'\|.$$

Proof By adding two copies of the inequality (19) with w and w' interchanged we get

$$(w-w')^\top(\nabla f(w)-\nabla f(w')) \geq \mu_s\|w-w'\|^2,$$

which in turn according to Cauchy-Schwartz inequality leads to $\|\mathrm{H}_F(\nabla f(w))-\mathrm{H}_F(\nabla f(w'))\| \geq \mu_s\|w-w'\|$. The inequality $\|\nabla f(w)-\nabla f(w')\| \leq \ell\|w-w'\|$ is standard (see, e.g., Nesterov, 2004, Theorem 2.1.5). ■

The following is another key lemma to the proof of Theorem 13.

Lemma 27 *Assume that the loss functions l_i are μ-strongly convex and ℓ-smooth. Let $\tilde{w}$ be an arbitrary k-sparse vector and $\tilde{F} = supp(\tilde{w})$. Let $\tilde{w}^\star = \arg\min_{supp(w)=\tilde{F}} P(w)$. Then*

$$\|\mathrm{H}_k(\nabla f(\tilde{w}^\star))\| \leq \left(1+\frac{\ell\gamma_k^+}{\mu\gamma_k^- + \lambda}\right)\|\mathrm{H}_k(\nabla f(\tilde{w}))\| + \frac{\lambda\ell\gamma_k^+}{\mu\gamma_k^- + \lambda}\|\tilde{w}\|.$$

Proof It can be verified that $f(w)$ is $\mu\gamma_k^-$-strongly convex and thus $P(w)$ is $(\mu\gamma_k^- + \lambda)$-strongly convex. From Lemma 26 we know that

$$\begin{aligned}\|\tilde{w}-\tilde{w}^\star\| &\leq \frac{1}{\mu\gamma_k^- + \lambda}\|\mathrm{H}_{\tilde{F}}(\nabla P(\tilde{w})) - \mathrm{H}_{\tilde{F}}(\nabla P(\tilde{w}^\star))\| \\ &\overset{\zeta_1}{=} \frac{1}{\mu\gamma_k^- + \lambda}\|\mathrm{H}_{\tilde{F}}(\nabla P(\tilde{w}))\| \\ &\leq \frac{1}{\mu\gamma_k^- + \lambda}\|\mathrm{H}_{\tilde{F}}(\nabla f(\tilde{w}))\| + \frac{\lambda}{\mu\gamma_k^- + \lambda}\|\tilde{w}\|,\end{aligned}$$

where "ζ_1" is due to the first-order optimality condition of $\tilde{w}^\star$ over $\tilde{F}$. Also, it can be verified that $f(w)$ is $\ell\gamma_k^+$-smooth. Then it follows from Lemma 26 and the above inequality that

$$\|\nabla f(\tilde{w}) - \nabla f(\tilde{w}^\star)\| \leq \ell\gamma_k^+\|\tilde{w}-\tilde{w}^\star\| \leq \frac{\ell\gamma_k^+}{\mu\gamma_k^- + \lambda}\|\mathrm{H}_{\tilde{F}}(\nabla f(\tilde{w}))\| + \frac{\lambda\ell\gamma_k^+}{\mu\gamma_k^- + \lambda}\|\tilde{w}\|.$$

Thus,

$$\begin{aligned}\|\mathrm{H}_k(\nabla f(\tilde{w}^\star))\| &\leq \|\nabla f(\tilde{w}) - \nabla f(\tilde{w}^\star)\| + \|\mathrm{H}_k(\nabla f(\tilde{w}))\| \\ &\leq \frac{\ell\gamma_k^+}{\mu\gamma_k^- + \lambda}\|\mathrm{H}_{\tilde{F}}(\nabla f(\tilde{w}))\| + \frac{\lambda\ell\gamma_k^+}{\mu\gamma_k^- + \lambda}\|\tilde{w}\| + \|\mathrm{H}_k(\nabla f(\tilde{w}))\| \\ &\leq \left(1+\frac{\ell\gamma_k^+}{\mu\gamma_k^- + \lambda}\right)\|\mathrm{H}_k(\nabla f(\tilde{w}))\| + \frac{\lambda\ell\gamma_k^+}{\mu\gamma_k^- + \lambda}\|\tilde{w}\|.\end{aligned}$$

This proves the desired bound. ■

We are now in the position to prove the main result in Theorem 13.

Proof [of Theorem 13] Let us consider

$$\tilde{w}^\star = \operatorname*{arg\,min}_{\mathrm{supp}(w)=\mathrm{supp}(\tilde{w})} P(w), \quad \tilde{\alpha}^\star = [l_1'(x_1^\top\tilde{w}^\star), ..., l_N'(x_N^\top\tilde{w}^\star)].$$

By applying Lemma 24 we can show that $(\tilde{w}^\star, \tilde{\alpha}^\star)$ is a $\tilde{\nu}^\star$-approximate sparse saddle point with relaxation level

$$\tilde{\nu}^\star = \frac{1}{\lambda}\|\mathrm{H}_k(\nabla f(\tilde{w}^\star))\|^2.$$

We now bound the quantity $\|\mathrm{H}_k(\nabla f(\tilde{w}^\star))\|$ in $\tilde{\nu}^\star$ using $\|\mathrm{H}_k(\nabla f(\tilde{w}))\|$. Since $\tilde{w}_{\min} > \frac{\|\nabla f(\tilde{w})\|_\infty}{\ell\gamma_k^+}$ and $\lambda \leq \frac{\mu\gamma_k^-\sqrt{k}\|\nabla f(\tilde{w})\|_\infty}{\ell\gamma_k^+\|\tilde{w}\| - \sqrt{k}\|\nabla f(\tilde{w})\|_\infty}$, according to Lemma 27 we can show that

$$\begin{aligned}
\|\mathrm{H}_k(\nabla f(\tilde{w}^\star))\| &\leq \left(1 + \frac{\ell\gamma_k^+}{\mu\gamma_k^- + \lambda}\right)\|\mathrm{H}_k(\nabla f(\tilde{w}))\| + \frac{\lambda\ell\gamma_k^+}{\mu\gamma_k^- + \lambda}\|\tilde{w}\| \\
&\leq \left(1 + \frac{\ell\gamma_k^+}{\mu\gamma_k^- + \lambda}\right)\sqrt{k}\|\nabla f(\tilde{w})\|_\infty + \frac{\lambda\ell\gamma_k^+}{\mu\gamma_k^- + \lambda}\|\tilde{w}\| \\
&\leq \left(2 + \frac{\ell\gamma_k^+}{\mu\gamma_k^- + \lambda}\right)\sqrt{k}\|\nabla f(\tilde{w})\|_\infty.
\end{aligned}$$

Therefore

$$\tilde{\nu}^\star \leq \frac{k}{\lambda}\left(2 + \frac{\ell\gamma_k^+}{\mu\gamma_k^- + \lambda}\right)^2 \|\nabla f(\tilde{w})\|_\infty^2.$$

Finally, from Proposition 12 we obtain

$$P(\tilde{w}^\star) - D(\tilde{\alpha}^\star) \leq \tilde{\nu}^\star \leq \frac{k}{\lambda}\left(2 + \frac{\ell\gamma_k^+}{\mu\gamma_k^- + \lambda}\right)^2 \|\nabla f(\tilde{w})\|_\infty^2.$$

The desired bound then follows immediately from the fact $P(\bar{w}) - D(\bar{\alpha}) \leq P(\tilde{w}^\star) - D(\tilde{\alpha}^\star)$. ■

Appendix B. Proofs of Results in Section 4

In this section, we present the technical proofs of the main results stated in Section 4.

B.1 Proof of Theorem 15

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 losses $\{l_i\}_{i=1}^N$ are Lipschitz smooth.

Lemma 28 *Assume that the primal loss functions $\{l_i(\cdot)\}_{i=1}^N$ 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 $D(\alpha) = \frac{1}{N}\sum_{i=1}^N -l_i^*(\alpha_i) - \frac{\lambda}{2}\|w(\alpha)\|^2$. Let us consider two arbitrary dual variables $\alpha', \alpha'' \in \mathcal{F}$. The assumption of l_i being $1/\mu$-smooth implies that its convex

conjugate function l_i^* is μ-strongly-convex. Let $F'' = \text{supp}(w(\alpha''))$. Then

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

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

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

which 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 29 *For any dual vector $\alpha \in \mathcal{F}$ and the related primal vector*

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

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

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

Proof It can be directly verified according to the definitions of $P(w)$ and $D(\alpha)$ that

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

which is the desired expression. ■

Based on Lemma 29, we can further derive the following lemma which establishes a bound on the primal-dual gap.

Lemma 30 *Consider a primal-dual pair (w, α) satisfying $w = \mathrm{H}_k\left(-\frac{1}{\lambda N}\sum_{i=1}^N \alpha_i x_i\right)$. Then the following inequality holds for any $g(\alpha) \in \partial D(\alpha)$ and $\beta \in [\partial l_1(w^\top x_1), ..., \partial l_N(w^\top x_N)]$:*

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

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

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

and

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

By summing both sides of above two equalities we get

$$\begin{aligned} l_i(w^\top x_i) + l_i^*(\alpha_i) =& w^\top x_i l_i'(w^\top x_i) + \alpha_i l_i^{*'}(\alpha_i) - (l_i(l_i^{*'}(\alpha_i)) + l_i^*(l_i'(w^\top x_i))) \\ \overset{\zeta_1}{\leq}& w^\top x_i l_i'(w^\top x_i) + \alpha_i l_i^{*'}(\alpha_i) - l_i^{*'}(\alpha_i) l_i'(w^\top x_i), \end{aligned} \tag{20}$$

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

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

where "ζ_2" follows from (20) and "ζ_3" follows from Lemma 29. 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 31 *Assume that $\{l_i\}_{i=1,...,N}$ are differentiable and $\bar{\epsilon} := \bar{w}_{\min} - \frac{1}{\lambda}\|P'(\bar{w})\|_\infty > 0$. Let $\bar{\alpha} = [l_1'(\bar{w}^\top x_1), ..., l_N'(\bar{w}^\top x_N)]$.*

(a) If $\|\alpha - \bar{\alpha}\| \leq \frac{\lambda N \bar{\epsilon}}{2\|X\|}$, then supp$(w(\alpha))$ = supp$(\bar{w})$ and

$$\|w(\alpha) - \bar{w}\| \leq \frac{\|X\|}{\lambda N}\|\alpha - \bar{\alpha}\|.$$

(b) If $\|\alpha - \bar{\alpha}\| > \frac{\lambda N \bar{\epsilon}}{2\|X\|}$, then

$$\|w(\alpha) - \bar{w}\| \leq \frac{\|X\|}{\lambda N}\left(1 + \frac{4\|X\|\|\bar{\alpha}\|}{\lambda N \bar{\epsilon}}\right)\|\alpha - \bar{\alpha}\|.$$

Proof *Part(a):* For any $\alpha \in \mathcal{F}$, let us define

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

Consider $\bar{F} = \text{supp}(\bar{w})$. Given $\bar{\epsilon} > 0$, it is known from Theorem 8 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 $\|\alpha - \bar{\alpha}\| \leq \frac{\lambda N\bar{\epsilon}}{2\|X\|}$, we can show that

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

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

$$\begin{aligned}\|w(\alpha) - w(\bar{\alpha})\| &= \|\text{H}_{\bar{F}}(\tilde{w}(\alpha)) - \text{H}_{\bar{F}}(\tilde{w}(\bar{\alpha}))\| \\ &\leq \|\tilde{w}(\alpha) - \tilde{w}(\bar{\alpha})\| = \frac{1}{\lambda N}\|X(\alpha - \bar{\alpha})\| \leq \frac{\|X\|}{\lambda N}\|\alpha - \bar{\alpha}\|.\end{aligned}$$

This proves the desired bound in Part(a).

Part(b): Next let us consider the case $\|\alpha - \bar{\alpha}\| > \frac{\lambda N\bar{\epsilon}}{2\|X\|}$. From the expression of $w(\alpha)$ we can verify that $\|w(\alpha)\| \leq \frac{1}{\lambda N}\|X\alpha\| \leq \frac{1}{\lambda N}\|X\|\|\alpha\|$. Then we have

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

This completes the proof. ■

We are now ready to prove the main result in Theorem 15.

Proof [of Theorem 15] *Part(a):* Let us consider $g^{(t)} \in \partial D(\alpha^{(t)})$ with $g_i^{(t)} = \frac{1}{N}(x_i^\top w^{(t)} - l_i^{*\prime}(\alpha_i^{(t)}))$. From the expression of $w^{(t)}$ we can verify

$$\|w^{(t)}\| \leq \frac{1}{\lambda N}\|X\alpha^{(t)}\| \leq \frac{\|X\|\|\alpha^{(t)}\|}{\lambda N} \leq \frac{r\|X\|}{\lambda\sqrt{N}}.$$

Since $\|x_i\| \leq 1$, it can be verified that

$$\|g^{(t)}\| \leq \frac{r\|X\| + \lambda\sqrt{N}\rho}{\lambda N}. \tag{21}$$

Let $\bar{g} \in \partial D(\bar{\alpha})$ with $\bar{g}_i = \frac{1}{N}(x_i^\top w(\bar{\alpha}) - l_i^{*'}(\bar{\alpha}_i))$. 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 $l^{*'}(l'(a)) = a$ we can derive $g_i^{(t)} = \frac{1}{N}(x_i^\top \bar{w} - l_i^{*'}(l_i'(x_i^\top \bar{w}))) = \frac{1}{N}(x_i^\top \bar{w} - x_i^\top \bar{w}) = 0$, 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 28 we know that $(h^{(t)})^2 \leq N v^{(t)}/\mu$. Then

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

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

$$(h^{(t)})^2 \leq \left(1 - \frac{2}{t+1}\right)(h^{(t-1)})^2 + \frac{(r\|X\| + \lambda\sqrt{N}\rho)^2}{\lambda^2\mu^2(t+1)^2}. \tag{22}$$

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 $c_0 = \frac{(r\|X\| + \lambda\sqrt{N}\rho)^2}{\lambda^2\mu^2}$. The base-case $t = 1$ follows immediately from (22). Now considering $t \geq 2$, the bound in (22) 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 $\beta^{(t)} := [l_1'(x_1^\top w^{(t)}), ..., l_N'(x_N^\top w^{(t)})]$. According to Lemma 30 we have

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

From the smoothness of l_i and Lemma 31 we get

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

where in the first "$\leq$" we have used the assumption $\|x_i\| \leq 1$. By combining the above with the bound in (21) we obtain

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

This completes the proof of Part(a).

Part(b): Let us consider $\epsilon_0 = \frac{\lambda N\bar{\epsilon}}{2\|X\|}$. 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 31 that $\text{supp}(w^{(t)}) = \text{supp}(\bar{w})$. This proves the claim of Part(b). ■

B.2 Proof of Theorem 17

Proof *Part(a):* Let us consider $g^{(t)} \in \partial D(\alpha^{(t)})$ with $g_i^{(t)} = \frac{1}{N}(x_i^\top w^{(t)} - l_i^{*'}(\alpha_i^{(t)}))$, and $\bar{g} \in \partial D(\bar{\alpha})$ with $\bar{g}_i = \frac{1}{N}(x_i^\top w(\bar{\alpha}) - l_i^{*'}(\bar{\alpha}_i))$. Since $\bar{\epsilon} = \bar{w}_{\min} - \frac{1}{\lambda}\|P'(\bar{w})\|_\infty > 0$, based on the proof of the part(a) of Theorem 15 we can verify that $\bar{w} = w(\bar{\alpha})$ and $\bar{g} = 0$. The $1/\ell$-strong-convexity of l_i implies the ℓ-smoothness of l_i^*. Then we can show that

$$\begin{aligned}
\|g^{(t)} - \bar{g}\| &= \frac{1}{N}\sqrt{\sum_{i=1}^N \left(x_i^\top(w^{(t)} - w(\bar{\alpha})) - l_i^{*'}(\alpha_i^{(t)}) + l_i^{*'}(\bar{\alpha}_i)\right)^2} \\
&\leq \frac{\sqrt{2}}{N}\sqrt{\sum_{i=1}^N \left(x_i^\top(w^{(t)} - w(\bar{\alpha}))\right)^2 + \sum_{i=1}^N \left(l_i^{*'}(\alpha_i^{(t)}) - l_i^{*'}(\bar{\alpha}_i)\right)^2} \\
&\leq \frac{\sqrt{2}}{N}\sqrt{\sum_{i=1}^N \left(x_i^\top(w^{(t)} - w(\bar{\alpha}))\right)^2} + \frac{\sqrt{2}}{N}\sqrt{\sum_{i=1}^N \left(l_i^{*'}(\alpha_i^{(t)}) - l_i^{*'}(\bar{\alpha}_i)\right)^2} \\
&\overset{\zeta_1}{\leq} \sqrt{\frac{2}{N}}\|w^{(t)} - w(\bar{\alpha})\| + \frac{\sqrt{2}\ell}{N}\|\alpha^{(t)} - \bar{\alpha}\| \\
&\overset{\zeta_2}{\leq} \left(\frac{\sqrt{2}\|X\|}{\lambda N\sqrt{N}}\left(1 + \frac{4\|X\|\|\bar{\alpha}\|}{\lambda N\bar{\epsilon}}\right) + \frac{\sqrt{2}\ell}{N}\right)\|\alpha^{(t)} - \bar{\alpha}\| \\
&= \ell_D\|\alpha^{(t)} - \bar{\alpha}\|,
\end{aligned}$$

where in "ζ_1" we have used $\|x_i\| \leq 1$ and $|l_i^{*'}(\alpha_i^{(t)}) - l_i^{*'}(\bar{\alpha}_i)| \leq \ell|\alpha_i^{(t)} - \bar{\alpha}_i|$, "$\zeta_2$" follows from Lemma 31. Now let $h^{(t)} = \|\alpha^{(t)} - \bar{\alpha}\|$ and $v^{(t)} = \langle g^{(t)} - \bar{g}, \bar{\alpha} - \alpha^{(t)}\rangle$. From Lemma 28 we

know that $(h^{(t)})^2 \leq Nv^{(t)}/\mu = v^{(t)}/\mu_D$. Then

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

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

$$(h^{(t)})^2 \leq \left(1 - \frac{\mu_D^2}{\ell_D^2}\right)(h^{(t-1)})^2.$$

By recursively applying the above inequality we obtain that

$$(h^{(t)})^2 \leq \left(1 - \frac{\mu_D^2}{\ell_D^2}\right)^t (h^{(0)})^2 = \left(1 - \frac{\mu_D^2}{\ell_D^2}\right)^t \|\bar{\alpha}\|^2,$$

where we have used $\alpha^{(0)} = 0$. This proves the claim in Part(a).

Now let $\beta^{(t)} := [l_1'(x_1^\top w^{(t)}), ..., l_N'(x_N^\top w^{(t)})]$. According to Lemma 30 we have

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

where we have used $\bar{g} = 0$. From the smoothness of l_i and Lemma 31 we obtain

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

Since we have already shown in Part(a) that $\|g^{(t)} - \bar{g}\| \leq \ell_D\|\alpha^{(t)} - \bar{\alpha}\|$, the following is valid immediately:

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

which then implies the desired bound on primal-dual gap. This concludes the proof of Part(a).

Part(b): Let us consider $\epsilon_0 = \frac{\lambda N\bar{\epsilon}}{2\|X\|}$. Based on the fact $(1-a)^t \leq \exp(-at)$ for $a \in (0,1)$ and the result in Part(a) we can show

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

after $t \geq \frac{\ell_D^2}{\mu_D^2}\log\left(\frac{4\|\bar{\alpha}\|^2\|X\|^2}{\lambda^2 N^2\bar{\epsilon}^2}\right)$. In this case, it is known from Lemma 31 that $\mathrm{supp}(w^{(t)}) = \mathrm{supp}(\bar{w})$. This proves the claim of Part(b). ■

B.3 Proof of Theorem 19

Proof *Part(a):* The proof argument largely mimics that of Theorem 15. Here we still provide a detailed proof for the sake of completeness. Let $h^{(t)} = \|\alpha^{(t)} - \bar{\alpha}\|$ and $v^{(t)} = \langle g^{(t)} - \bar{g}, \bar{\alpha} - \alpha^{(t)} \rangle$. From Lemma 28 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

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

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

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

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

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

By induction, this recursive inequality leads to

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

Moreover, similar to the argument in the proof of Theorem 15 we obtain

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

where in the first inequality we have used $\mathbb{E}[\|\alpha^{(t)} - \bar{\alpha}\|] \leq \sqrt{\mathbb{E}[\|\alpha^{(t)} - \bar{\alpha}\|^2]}$ This proves the results in Part(a).

Part(b): Let us consider $\epsilon_0 = \frac{\lambda N \bar{\epsilon}}{2\|X\|}$. From Part(a) we obtain $\mathbb{E}\left[\|\alpha^{(t)} - \bar{\alpha}\|\right] \leq \delta\epsilon_0$ after $t \geq t_1 = \left\lceil \frac{m(r\|X\|+\lambda\sqrt{N}\rho)^2}{\lambda^2\mu^2\delta^2\epsilon_0^2} \right\rceil$. 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 31 shows that $\|\alpha^{(t)} - \bar{\alpha}\| \leq \epsilon_0$ implies $\mathrm{supp}(w^{(t)}) = \mathrm{supp}(\bar{w})$. Therefore when $t \geq t_1$, the event $\mathrm{supp}(w^{(t)}) = \mathrm{supp}(\bar{w})$ occurs with probability at least $1 - \delta$. This proves the desired result in Part(b). ■

B.4 Proof of Theorem 22

Proof *Part(a):* Based on the proof of Theorem 17 we know that $\bar{g} = 0$ and $\|g^{(t)} - \bar{g}\| \leq \ell_D\|\alpha^{(t)} - \bar{\alpha}\|$. Let $h^{(t)} = \|\alpha^{(t)} - \bar{\alpha}\|$ and $v^{(t)} = \langle g^{(t)} - \bar{g}, \bar{\alpha} - \alpha^{(t)}\rangle$. From Lemma 28 we know that $(h^{(t)})^2 \leq v^{(t)}/\mu_D$. 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 have

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

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

$$\begin{aligned}
\mathbb{E}\left[(h^{(t)})^2 \mid \alpha^{(t-1)}\right] \leq& (h^{(t-1)})^2 - \frac{1}{m}\sum_{i=1}^{m} 2\eta^{(t-1)} v^{(t-1)}_{B_i} + \frac{1}{m}\sum_{i=1}^{m} (\eta^{(t-1)})^2 \|g^{(t-1)}_{B_i} - \bar{g}_{B_i}\|^2 \\
=& (h^{(t-1)})^2 - \frac{2\eta^{(t-1)}}{m} v^{(t-1)} + \frac{(\eta^{(t-1)})^2}{m}\|g^{(t-1)} - \bar{g}\|^2 \\
\leq& (h^{(t-1)})^2 - \frac{2\eta^{(t-1)}\mu_D}{m}(h^{(t-1)})^2 + \frac{(\eta^{(t-1)})^2\ell_D^2}{m}(h^{(t-1)})^2 \\
=& \left(1 - \frac{2\eta^{(t-1)}\mu_D}{m} + \frac{(\eta^{(t-1)})^2\ell_D^2}{m}\right)(h^{(t-1)})^2.
\end{aligned}$$

Let $\eta^{(t)} = \frac{\mu_D}{\ell_D^2}$. Then we obtain

$$\mathbb{E}\left[(h^{(t)})^2 \mid \alpha^{(t-1)}\right] \leq \left(1 - \frac{\mu_D^2}{m\ell_D^2}\right)(h^{(t-1)})^2.$$

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

$$\mathbb{E}\left[(h^{(t)})^2\right] \leq \left(1 - \frac{\mu_D^2}{m\ell_D^2}\right)\mathbb{E}\left[(h^{(t-1)})^2\right].$$

This recursive inequality leads to

$$\mathbb{E}\left[(h^{(t)})^2\right] \leq \left(1 - \frac{\mu_D^2}{m\ell_D^2}\right)^t \|\bar{\alpha}\|^2,$$

where we have used $\alpha^{(0)} = 0$.

Following the similar arguments in the proof of Part(a) of Theorem 17 we can show that

$$\begin{aligned}\mathbb{E}[\epsilon_{PD}^{(t)}] &\leq \ell_D \left(\frac{\|X\|}{\lambda\mu\sqrt{N}}\left(1 + \frac{4\|X\|\|\bar{\alpha}\|}{\lambda N\bar{\epsilon}}\right) + 1\right)\mathbb{E}[\|\alpha^{(t)} - \bar{\alpha}\|^2] \\ &\leq \ell_D \left(\frac{\|X\|}{\lambda\mu\sqrt{N}}\left(1 + \frac{4\|X\|\|\bar{\alpha}\|}{\lambda N\bar{\epsilon}}\right) + 1\right)\left(1 - \frac{\mu_D^2}{m\ell_D^2}\right)^t \|\bar{\alpha}\|^2,\end{aligned}$$

which is the desired bound in Part(a).

Part(b): Let us consider $\epsilon_0 = \frac{\lambda N\bar{\epsilon}}{2\|X\|}$. From Part(a) we obtain $\mathbb{E}\left[\|\alpha^{(t)} - \bar{\alpha}\|\right] \leq \delta\epsilon_0$ after $t \geq \frac{m\ell_D^2}{\mu_D^2}\log\left(\frac{4\|\bar{\alpha}\|^2\|X\|^2}{\delta^2\lambda^2N^2\bar{\epsilon}^2}\right)$. Then from the Markov inequality we know that $\|\alpha^{(t)} - \bar{\alpha}\| \leq \mathbb{E}\left[\|\alpha^{(t)} - \bar{\alpha}\|\right]/\delta \leq \epsilon_0$ holds with probability at least $1 - \delta$. Lemma 31 shows that $\|\alpha^{(t)} - \bar{\alpha}\| \leq \epsilon_0$ implies $\mathrm{supp}(w^{(t)}) = \mathrm{supp}(\bar{w})$. Therefore in this case, the event $\mathrm{supp}(w^{(t)}) = \mathrm{supp}(\bar{w})$ occurs with probability at least $1 - \delta$. ■

References

Andreas Argyriou, Rina Foygel, and Nathan Srebro. Sparse prediction with the k-support norm. In *Advances in Neural Information Processing Systems*, pages 1457–1465, 2012.

Francis Bach, Rodolphe Jenatton, Julien Mairal, and Guillaume Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends® in Machine Learning*, 4(1): 1–106, 2012.

Sohail Bahmani, Bhiksha Raj, and Petros T Boufounos. Greedy sparsity-constrained optimization. *Journal of Machine Learning Research*, 14:807–841, 2013.

Amir Beck and Yonina C Eldar. Sparsity constrained nonlinear optimization: Optimality conditions and algorithms. *SIAM Journal on Optimization*, 23(3):1480–1509, 2013.

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

Thomas Blumensath and Mike E Davies. Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis*, 27(3):265–274, 2009.

Chuangjian Cai, Lin Zhang, Wenjuan Cai, Dong Zhang, Yanlu Lv, and Jianwen Luo. Nonlinear greedy sparsity-constrained algorithm for direct reconstruction of fluorescence molecular lifetime tomography. *Biomedical optics express*, 7(4):1210–1226, 2016.

Antonin Chambolle and Thomas 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.

Jinghui Chen and Quanquan Gu. Accelerated stochastic block coordinate gradient descent for sparsity constrained nonconvex optimization. In *Conference on Uncertainty in Artificial Intelligence*, 2016.

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

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

Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC press, 2015.

Cho-Jui Hsieh, Kai-Wei Chang, Chih-Jen Lin, S Sathiya Keerthi, and Sellamanickam Sundararajan. A dual coordinate descent method for large-scale linear svm. In *International Conference on Machine Learning*, pages 408–415, 2008.

Martin Jaggi, Virginia Smith, Martin Takác, Jonathan Terhorst, Sanjay Krishnan, Thomas Hofmann, and Michael I Jordan. Communication-efficient distributed dual coordinate ascent. In *Advances in Neural Information Processing Systems*, pages 3068–3076, 2014.

Prateek Jain, Ambuj Tewari, and Purushottam Kar. On iterative hard thresholding methods for high-dimensional m-estimation. In *Advances in Neural Information Processing Systems*, pages 685–693, 2014.

Prateek Jain, Nikhil Rao, and Inderjit S Dhillon. Structured sparse regression via greedy hard thresholding. In *Advances in Neural Information Processing Systems*, pages 1516–1524, 2016.

Xiaojie Jin, Xiaotong Yuan, Jiashi Feng, and Shuicheng Yan. Training skinny deep neural networks with iterative hard thresholding methods. *arXiv preprint arXiv:1607.05423*, 2016.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, pages 315–323, 2013.

Rajiv Khanna and Anastasios Kyrillidis. IHT dies hard: Provable accelerated iterative hard thresholding. In *International Conference on Artificial Intelligence and Statistics*, pages 188–198, 2018.

Ken Lang. Newsweeder: Learning to filter netnews. In *International Conference on Machine Learning*, pages 331–339, 1995.

Maksim Lapin, Bernt Schiele, and Matthias Hein. Scalable multitask representation learning for scene classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1434–1441, 2014.

Sangkyun Lee, Damian Brzyski, and Malgorzata Bogdan. Fast saddle-point algorithm for generalized dantzig selector and fdr control with ordered L1-norm. In *International Conference on Artificial Intelligence and Statistics*, pages 780–789, 2016.

David D Lewis, Yiming Yang, Tony G Rose, and Fan Li. Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397, 2004.

Xingguo Li, Tuo Zhao, Raman Arora, Han Liu, and Jarvis Haupt. Stochastic variance reduced optimization for nonconvex sparse learning. In *International Conference on Machine Learning*, pages 917–925, 2016.

Bo Liu, Xiao-Tong Yuan, Lezi Wang, Qingshan Liu, and Dimitris N Metaxas. Dual iterative hard thresholding: From non-convex sparse minimization to non-smooth concave maximization. In *International Conference on Machine Learning*, pages 2179–2187, 2017.

Bo Liu, Xiao-Tong Yuan, Lezi Wang, Qingshan Liu, Junzhou Huang, and Dimitris N Metaxas. Distributed inexact newton-type pursuit for non-convex sparse learning. In *International Conference on Artificial Intelligence and Statistics*, pages 343–352, 2019.

Rahul Mazumder and Trevor Hastie. The graphical lasso: New insights and alternatives. *Electronic Journal of Statistics*, 6:2125, 2012.

Balas Kausik Natarajan. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24(2):227–234, 1995.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer Academic Publishers, 2004.

Nam Nguyen, Deanna Needell, and Tina Woolf. Linear convergence of stochastic iterative greedy algorithms with sparse constraints. *IEEE Transactions on Information Theory*, 63(11):6869–6895, 2017.

Yagyensh Chandra Pati, Ramin Rezaiifar, and Perinkulam Sambamurthy Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Annual Asilomar Conference on Signals, Systems, and Computers*, pages 40–44, 1993.

Clint Scovel, Don Hush, and Ingo Steinwart. Approximate duality. *Journal of Optimization Theory and Applications*, 135(3):429–443, 2007.

Shai Shalev-Shwartz. SDCA without duality, regularization, and individual convexity. In *International Conference on Machine Learning*, pages 747–754, 2016.

Shai Shalev-Shwartz and Tong Zhang. Accelerated mini-batch stochastic dual coordinate ascent. In *Advances in Neural Information Processing Systems*, pages 378–385, 2013a.

Shai Shalev-Shwartz and Tong Zhang. Stochastic dual coordinate ascent methods for regularized loss. *Journal of Machine Learning Research*, 14(1):567–599, 2013b.

Shai Shalev-Shwartz and Tong Zhang. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. *Mathematical Programming*, 155(1-2):105–145, 2016.

Jie Shen and Ping Li. On the iteration complexity of support recovery via hard thresholding pursuit. In *International Conference on Machine Learning*, pages 3115–3124, 2017a.

Jie Shen and Ping Li. Partial hard thresholding: Towards a principled analysis of support recovery. In *Advances in Neural Information Processing Systems*, pages 3127–3137, 2017b.

Jie Shen and Ping Li. A tight bound of hard thresholding. *Journal of Machine Learning Research*, 18(208):1–42, 2018.

Mohammadreza Soltani and Chinmay Hegde. Fast algorithms for demixing sparse signals from nonlinear observations. *IEEE Transactions on Signal Processing*, 65(16):4209–4222, 2017.

Conghui Tan, Tong Zhang, Shiqian Ma, and Ji Liu. Stochastic primal-dual method for empirical risk minimization with o (1) per-iteration complexity. In *Advances in Neural Information Processing Systems*, pages 8376–8385, 2018.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

Jialei Wang, Mladen Kolar, Nathan Srebro, and Tong Zhang. Efficient distributed learning with sparsity. In *International Conference on Machine Learning*, pages 3636–3645, 2017.

Lin Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11:2543–2596, 2010.

Lin Xiao, Adams Wei Yu, Qihang Lin, and Weizhu Chen. Dscovr: Randomized primal-dual block coordinate algorithms for asynchronous distributed optimization. *Journal of Machine Learning Research*, 20:1–58, 2019.

Adams Wei Yu, Qihang Lin, and Tianbao Yang. Doubly stochastic primal-dual coordinate method for empirical risk minimization and bilinear saddle-point problem. *arXiv preprint arXiv:1508.03390*, 2015.

Xiao-Tong Yuan and Ping Li. Generalization bounds for high-dimensional m-estimation under sparsity constraint. *arXiv preprint arXiv:2001.07212*, 2020.

Xiao-Tong Yuan, Ping Li, and Tong Zhang. Gradient hard thresholding pursuit for sparsity-constrained optimization. In *International Conference on Machine Learning*, pages 127–135, 2014.

Xiao-Tong Yuan, Ping Li, and Tong Zhang. Exact recovery of hard thresholding pursuit. In *Advances in Neural Information Processing Systems*, pages 3558–3566, 2016.

Xiao-Tong Yuan, Ping Li, and Tong Zhang. Gradient hard thresholding pursuit. *Journal of Machine Learning Research*, 18(166):1–43, 2018.

Cun-Hui Zhang and Jian Huang. The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics*, pages 1567–1594, 2008.

Yuchen Zhang and Lin Xiao. Stochastic primal-dual coordinate method for regularized empirical risk minimization. *Journal of Machine Learning Research*, 18(1):2939–2980, 2017.

Pan Zhou, Xiao-Tong Yuan, and Jiashi Feng. Efficient stochastic gradient hard thresholding. In *Advances in Neural Information Processing Systems*, pages 1988–1997, 2018.