Dual Iterative Hard Thresholding: From Non-convex Sparse Minimization to Non-smooth Concave Maximization

Bo Liu [1] Xiao-Tong Yuan [2] Lezi Wang [1] Qingshan Liu [2] Dimitris N. Metaxas [1]

Abstract

Iterative Hard Thresholding (IHT) is a class of projected gradient descent methods for optimizing sparsity-constrained minimization models, with the best known efficiency and scalability in practice. As far as we know, the existing IHT-style methods are designed for sparse minimization in primal form. It remains open to explore duality theory and algorithms in such a non-convex and NP-hard problem setting. In this paper, we bridge this gap by establishing a duality theory for sparsity-constrained minimization with ℓ_2-regularized loss function and proposing an IHT-style algorithm for dual maximization. Our sparse duality theory provides a set of sufficient and necessary conditions under which the original NP-hard/non-convex problem can be equivalently solved in a dual formulation. The proposed dual IHT algorithm is a super-gradient method for maximizing the non-smooth dual objective. An interesting finding is that the sparse recovery performance of dual IHT is invariant to the Restricted Isometry Property (RIP), which is required by virtually all the existing primal IHT algorithms without sparsity relaxation. Moreover, a stochastic variant of dual IHT is proposed for large-scale stochastic optimization. Numerical results demonstrate the superiority of dual IHT algorithms to the state-of-the-art primal IHT-style algorithms in model estimation accuracy and computational efficiency.

1. Introduction

Sparse learning has emerged as an effective approach to alleviate model overfitting when feature dimension outnumbers training sample. Given 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 in high-dimensional analysis:

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

Here $l(\cdot;\cdot)$ is a convex loss function, $w \in \mathbb{R}^d$ is the model parameter vector and λ controls the regularization strength. For example, the squared loss $l(a,b) = (b-a)^2$ is used in linear regression and the hinge loss $l(a,b) = \max\{0, 1-ab\}$ in support vector machines. Due to the presence of cardinality constraint $\|w\|_0 \leq k$, the problem (1) is simultaneously non-convex and NP-hard in general, and thus is challenging for optimization. A popular 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.

In this paper, we are interested in algorithms that directly minimize the non-convex formulation in (1). Early efforts mainly lie in compressed sensing for signal recovery, which is a special case of (1) with squared loss. Among others, a family of the so called Iterative Hard Thresholding (IHT) methods (Blumensath & Davies, 2009; Foucart, 2011) have gained significant interests and they have been witnessed to offer the fastest and most scalable solutions in many cases. More recently, IHT-style methods have been generalized to handle generic convex loss functions (Beck & Eldar, 2013; Yuan et al., 2014; Jain et al., 2014) 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.

Although IHT-style methods have been extensively studied, the state-of-the-art is only designed for the primal formulation (1). It remains an open problem to investigate the feasibility of solving the original NP-hard/non-convex formulation in a dual space that might potentially further im-

[1] Department of CS, Rutgers University, Piscataway, NJ, 08854, USA. [2] B-DAT Lab, Nanjing University of Information Science & Technology, Nanjing, Jiangsu, 210044, China. Correspondence to: Bo Liu <lb507@cs.rutgers.edu>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

prove computational efficiency. To fill this gap, inspired by the recent success of dual methods in regularized learning problems, we systematically build a sparse duality theory and propose an IHT-style algorithm along with its stochastic variant for dual optimization.

Overview of our contribution. The core contribution of this work is two-fold in theory and algorithm. As the theoretical contribution, we have established a novel sparse Lagrangian duality theory for the NP-hard/non-convex problem (1) which to the best of our knowledge has not been reported elsewhere in literature. We provide in this part a set of *sufficient and necessary* conditions under which one can safely solve the original non-convex problem through maximizing its concave dual objective function. As the algorithmic contribution, we propose the dual IHT (DIHT) algorithm as a super-gradient method to maximize the non-smooth dual objective. In high level description, DIHT iterates between dual gradient ascent and primal hard thresholding pursuit until convergence. A stochastic variant of DIHT is proposed to handle large-scale learning problems. For both algorithms, we provide non-asymptotic convergence analysis on parameter estimation error, sparsity recovery, and primal-dual gap as well. In sharp contrast to the existing analysis for primal IHT-style algorithms, our analysis is not relying on Restricted Isometry Property (RIP) conditions and thus is less restrictive in real-life high-dimensional statistical settings. Numerical results on synthetic datasets and machine learning benchmark datasets demonstrate that dual IHT significantly outperforms the state-of-the-art primal IHT algorithms in accuracy and efficiency. The theoretical and algorithmic contributions of this paper are highlighted in below:

- Sparse Lagrangian duality theory: we established a sparse saddle point theorem (Theorem 1), a sparse mini-max theorem (Theorem 2) and a sparse strong duality theorem (Theorem 3).
- Dual optimization: we proposed an IHT-style algorithm along with its stochastic extension for non-smooth dual maximization. These algorithms have been shown to converge at the rate of $\frac{1}{\epsilon}\ln\frac{1}{\epsilon}$ in dual parameter estimation error and $\frac{1}{\epsilon^2}\ln\frac{1}{\epsilon^2}$ in primal-dual gap (see Theorem 4 and Theorem 5). These guarantees are invariant to RIP conditions which are required by virtually all the primal IHT-style methods without using relaxed sparsity levels.

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

Organization. The rest of this paper is organized as follows: In §2 we briefly review some relevant work. In §3 we develop a Lagrangian duality theory for sparsity-constrained minimization problems. The dual IHT-style algorithms along with convergence analysis are presented in §4. The numerical evaluation results are reported in §5.1. Finally, the concluding remarks are made in §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 established under proper RIP (or restricted strong condition number) bound conditions (Blumensath, 2013; Yuan et al., 2014; 2016). In (Jain et al., 2014), several relaxed variants of IHT-style algorithms were presented for which the estimation consistency can be established without requiring the RIP conditions. In (Bahmani et al., 2013), a gradient support pursuit algorithm is proposed and analyzed. In large-scale settings where a full gradient evaluation on all data becomes a bottleneck, stochastic and variance reduction techniques have been adopted to improve the computational efficiency of IHT (Nguyen et al., 2014; Li et al., 2016; Chen & Gu, 2016).

Dual optimization algorithms have been widely used in various learning tasks including SVMs (Hsieh et al., 2008) and multi-task learning (Lapin et al., 2014). In recent years, stochastic dual optimization methods have gained significant attention in large-scale machine learning (Shalev-Shwartz & Zhang, 2013a;b). 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 & Pock, 2011) and generalized Dantzig selector (Lee et al., 2016). More recently, a number of stochastic variants (Zhang & Xiao, 2015; Yu et al., 2015) and parallel variants (Zhu & Storkey, 2016) were developed to make the primal-dual algorithms more scalable and efficient.

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 build the theoretical foundation of developing dual IHT methods.

From now on we abbreviate $l_i(w^\top x_i) = l(w^\top x_i, y_i)$. 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 $\mathcal{F} \subseteq \mathbb{R}$ be the feasible set of α_i. According to the well-known expression of $l_i(u) = \max_{\alpha_i \in \mathcal{F}} \{\alpha_i u - l_i^*(\alpha_i)\}$, the problem (1) can be reformulated into the following mini-max formulation:

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

The following Lagrangian form will be useful in analysis:

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

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

Definition 1 (Sparse Saddle Point)**.** *A pair $(\bar{w}, \bar{\alpha}) \in \mathbb{R}^d \times \mathcal{F}^N$ is said to be a k-sparse saddle point for L if $\|\bar{w}\|_0 \le k$ and the following holds for all $\|w\|_0 \le k, \alpha \in \mathcal{F}^N$:*

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

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

Theorem 1 (Sparse Saddle Point Theorem)**.** *Let $\bar{w} \in \mathbb{R}^d$ be a k-sparse primal vector and $\bar{\alpha} \in \mathcal{F}^N$ be a dual vector. Then the pair $(\bar{w}, \bar{\alpha})$ is a sparse saddle point for L if and only if the following conditions hold:*

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

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

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

Proof. A proof of this result is given in Appendix A.1. □

Remark 1. *Theorem 1 shows that the conditions (a)∼(c) are sufficient and necessary to guarantee the existence of a sparse saddle point for the Lagrangian form L. This result is different from from the traditional saddle point theorem which requires the use of the Slater Constraint Qualification to guarantee the existence of saddle point.*

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

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

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

Theorem 2 (Sparse Mini-Max Theorem)**.** *The mini-max relationship*

$$\max_{\alpha \in \mathcal{F}^N} \min_{\|w\|_0 \le k} L(w, \alpha) = \min_{\|w\|_0 \le k} \max_{\alpha \in \mathcal{F}^N} L(w, \alpha) \quad (4)$$

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

Proof. A proof of this result is given in Appendix A.2. □

The sparse mini-max result in Theorem 2 provides sufficient and necessary conditions under which one can safely exchange a min-max for a max-min, in the presence of sparsity constraint. The following corollary is a direct consequence of applying Theorem 1 to Theorem 2.

Corollary 1. *The mini-max relationship*

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

holds if and only if there exist a k-sparse primal vector $\bar{w} \in \mathbb{R}^d$ and a dual vector $\bar{\alpha} \in \mathcal{F}^N$ such that the conditions (a)∼(c) in Theorem 1 are satisfied.

The mini-max result in Theorem 2 can be used as a basis for establishing sparse duality theory. Indeed, we have already shown the following:

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

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 $D(\alpha) := \min_{\|w\|_0 \le k} L(w, \alpha)$, i.e.,

$$\max_{\alpha \in \mathcal{F}^N} D(\alpha) = \max_{\alpha \in \mathcal{F}^N} \min_{\|w\|_0 \le k} L(w, \alpha). \quad (5)$$

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

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

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

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

Proof. A proof of this result is given in Appendix A.3. □

Based on Theorem 1 and Theorem 2, we are able to further establish a sparse strong duality theorem which gives the sufficient and necessary conditions under which the optimal values of the primal and dual problems coincide.

Theorem 3 (Sparse Strong Duality Theorem)**.** *Let $\bar{w} \in \mathbb{R}^d$ is a k-sparse primal vector and $\bar{\alpha} \in \mathcal{F}^N$ be a dual vector. Then $\bar{\alpha}$ solves the dual problem in (5), i.e., $D(\bar{\alpha}) \geq D(\alpha)$, $\forall \alpha \in \mathcal{F}^N$, and $P(\bar{w}) = D(\bar{\alpha})$ if and only if the pair $(\bar{w}, \bar{\alpha})$ satisfies the conditions (a)∼(c) in Theorem 1.*

Proof. A proof of this result is given in Appendix A.4. □

We define the sparse primal-dual gap $\epsilon_{PD}(w, \alpha) := P(w) - D(\alpha)$. The main message conveyed by Theorem 3 is that the sparse primal-dual gap reaches zero at the primal-dual pair $(\bar{w}, \bar{\alpha})$ if and only if the conditions (a)∼(c) in Theorem 1 hold.

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 (5). 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 $\bar{w} = \mathrm{H}_k\left(-\frac{1}{\lambda N}\sum_{i=1}^N \bar{\alpha}_i x_i\right)$ as given in the condition (c). Since the dual objective function $D(\alpha)$ is shown to be concave, its global maximum can be estimated using any convex/concave optimization method. In the next section, we present a simple projected super-gradient method to solve the dual maximization problem.

4. Dual Iterative Hard Thresholding

Generally, $D(\alpha)$ is a non-smooth function since: 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 $w(\alpha)$. Therefore, smooth optimization methods are not directly applicable here and we resort to subgradient-type methods to solve the non-smooth dual maximization problem in (5).

Algorithm 1 Dual Iterative Hard Thresholding (DIHT)

```
Input : Training set $\{x_i, y_i\}_{i=1}^N$. Regularization strength
        parameter $\lambda$. Cardinality constraint $k$. Step-size $\eta$.
Initialization $w^{(0)} = 0$, $\alpha_1^{(0)} = ... = \alpha_N^{(0)} = 0$.
for $t = 1, 2, ..., T$ do
    (S1) Dual projected super-gradient ascent: $\forall\ i \in \{1, 2, ..., N\}$,
        $\alpha_i^{(t)} = \mathrm{P}_{\mathcal{F}}\left(\alpha_i^{(t-1)} + \eta^{(t-1)} g_i^{(t-1)}\right)$,    (6)
    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 $\mathrm{P}_{\mathcal{F}}(\cdot)$ is the Euclidian projection
    operator with respect to feasible set $\mathcal{F}$.
    (S2) Primal hard thresholding:
        $w^{(t)} = \mathrm{H}_k\left(-\frac{1}{\lambda N}\sum_{i=1}^N \alpha_i^{(t)} x_i\right)$.    (7)
end
Output: $w^{(T)}$.
```

4.1. Algorithm

The Dual Iterative Hard Thresholding (DIHT) algorithm, as outlined in Algorithm 1, is essentially a projected super-gradient method for maximizing $D(\alpha)$. The procedure generates a sequence of prima-dual pairs $(w^{(0)}, \alpha^{(0)}), (w^{(1)}, \alpha^{(1)}), \ldots$ from an initial pair $w^{(0)} = 0$ and $\alpha^{(0)} = 0$. At the t-th iteration, the dual update step **S1** conducts the projected super-gradient ascent in (6) to update $\alpha^{(t)}$ from $\alpha^{(t-1)}$ and $w^{(t-1)}$. Then in the primal update step **S2**, the primal variable $w^{(t)}$ is constructed from $\alpha^{(t)}$ using a k-sparse truncation operation in (7).

When a batch estimation of super-gradient $D'(\alpha)$ becomes expensive in large-scale applications, it is natural to consider the stochastic implementation of DIHT, namely SDIHT, as outlined in Algorithm 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 (8). Then in the primal update step **S2.1**, we incrementally update an intermediate accumulation vector $\tilde{w}^{(t)}$ which records $-\frac{1}{\lambda N}\sum_{i=1}^N \alpha_i^{(t)} x_i$ as a weighted sum of samples. In **S2.2**, the primal vector $w^{(t)}$ is updated by applying k-sparse truncation on $\tilde{w}^{(t)}$. The SDIHT is essentially a block-coordinate super-gradient method for the dual problem. Particularly, in the extreme case $m = 1$, SDIHT reduces to the batch DIHT. At the opposite extreme end with $m = N$, i.e., each block contains one sample, SDIHT becomes a stochastic coordinate-wise super-gradient method.

Algorithm 2 Stochastic Dual Iterative Hard Thresholding (SDIHT)

```
Input : Training set $\{x_i, y_i\}_{i=1}^N$. Regularization strength parameter $\lambda$. Cardinality constraint $k$. Step-size $\eta$. A block disjoint partition $\{B_1, ..., B_m\}$ of the sample index set $[1, ..., N]$.
Initialization $w^{(0)} = \tilde{w}^{(0)} = 0$, $\alpha_1^{(0)} = ... = \alpha_N^{(0)} = 0$.
for $t = 1, 2, ..., T$ do
    (S1) Dual projected super-gradient ascent: Uniformly randomly select an index block $B_i^{(t)} \in \{B_1, ..., B_m\}$. For all $j \in B_i^{(t)}$ update $\alpha_j^{(t)}$ as
        $$\alpha_j^{(t)} = \mathrm{P}_{\mathcal{F}}\left(\alpha_j^{(t-1)} + \eta^{(t-1)} g_j^{(t-1)}\right). \qquad (8)$$
    Set $\alpha_j^{(t)} = \alpha_j^{(t-1)}, \forall j \notin B_i^{(t)}$.
    (S2) Primal hard thresholding:
      – (S2.1) Intermediate update:
        $$\tilde{w}^{(t)} = \tilde{w}^{(t-1)} - \frac{1}{\lambda N} \sum_{j \in B_i^{(t)}} (\alpha_j^{(t)} - \alpha_j^{(t-1)}) x_j. \qquad (9)$$
      – (S2.2) Hard thresholding: $w^{(t)} = \mathrm{H}_k(\tilde{w}^{(t)})$.
end
Output: $w^{(T)}$.
```

The dual update (8) in SDIHT is much more efficient than DIHT as the former only needs to access a small subset of samples at a time. If the hard thresholding operation in primal update becomes a bottleneck, e.g., in high-dimensional settings, we suggest to use SDIHT with relatively smaller number of blocks so that the hard thresholding operation in **S2.2** can be less frequently called.

4.2. Convergence analysis

We now analyze the non-asymptotic convergence behavior of DIHT and SDIHT. In the following analysis, we will denote $\bar{\alpha} = \arg\max_{\alpha \in \mathcal{F}^N} D(\alpha)$ and use the abbreviation $\epsilon_{PD}^{(t)} := \epsilon_{PD}(w^{(t)}, \alpha^{(t)})$. Let $r = \max_{a \in \mathcal{F}} |a|$ be the bound of the dual feasible set $\mathcal{F}$ and $\rho = \max_{i, a \in \mathcal{F}} |l_i^{*'}(a)|$. For example, such quantities exist when l_i and l_i^* are Lipschitz continuous (Shalev-Shwartz & Zhang, 2013b). We also assume without loss of generality that $\|x_i\| \le 1$. Let $X = [x_1, ..., x_N] \in \mathbb{R}^{d \times N}$ be the data matrix. Given an index set F, we denote X_F as the restriction of X with *rows* restricted to F. The following quantities will be used in our analysis:

$$\sigma_{\max}^2(X, s) = \sup_{u \in \mathbb{R}^n, F} \left\{ u^\top X_F^\top X_F u \mid |F| \le s, \|u\| = 1 \right\},$$
$$\sigma_{\min}^2(X, s) = \inf_{u \in \mathbb{R}^n, F} \left\{ u^\top X_F^\top X_F u \mid |F| \le s, \|u\| = 1 \right\}.$$

Particularly, $\sigma_{\max}(X, d) = \sigma_{\max}(X)$ and $\sigma_{\min}(X, d) = \sigma_{\min}(X)$. We say a univariate differentiable function $f(x)$ is γ-smooth if $\forall x, y$, $f(y) \le f(x) + \langle f'(x), y - x \rangle + \frac{\gamma}{2}|x - y|^2$. The following is our main theorem on the dual parameter estimation error, support recovery and primal-dual gap of DIHT.

Theorem 4. *Assume that l_i is $1/\mu$-smooth. Set $\eta^{(t)} = \frac{\lambda N^2}{(\lambda N \mu + \sigma_{\min}(X, k))(t+1)}$. Define constants $c_1 = \frac{N^3 (r + \lambda\rho)^2}{(\lambda N \mu + \sigma_{\min}(X, k))^2}$ and $c_2 = (r + \lambda\rho)^2 \left(1 + \frac{\sigma_{\max}(X, k)}{\mu \lambda N}\right)^2$.*

(a) **Parameter estimation error**: *The sequence $\{\alpha^{(t)}\}_{t \ge 1}$ generated by Algorithm 1 satisfies the following estimation error inequality:*

$$\|\alpha^{(t)} - \bar{\alpha}\|^2 \le c_1 \left(\frac{1}{t} + \frac{\ln t}{t}\right),$$

(b) **Support recovery and primal-dual gap**: *Assume additionally that $\bar{\epsilon} := \bar{w}_{\min} - \frac{1}{\lambda}\|P'(\bar{w})\|_\infty > 0$. Then, supp$(w^{(t)})$ = supp$(\bar{w})$ when*

$$t \ge t_0 = \left\lceil \frac{12 c_1 \sigma_{\max}^2(X)}{\lambda^2 N^2 \bar{\epsilon}^2} \ln \frac{12 c_1 \sigma_{\max}^2(X)}{\lambda^2 N^2 \bar{\epsilon}^2} \right\rceil.$$

Moreover, for any $\epsilon > 0$, the primal-dual gap satisfies $\epsilon_{PD}^{(t)} \le \epsilon$ when $t \ge \max\{t_0, t_1\}$ where $t_1 = \left\lceil \frac{3 c_1 c_2}{\lambda^2 N \epsilon^2} \ln \frac{3 c_1 c_2}{\lambda^2 N \epsilon^2} \right\rceil$.

Proof. A proof of this result is given in Appendix A.5. □

Remark 3. *The theorem allows $\mu = 0$ when $\sigma_{\min}(X, k) > 0$. If $\mu > 0$, then $\sigma_{\min}(X, k)$ is allowed to be zero and thus the step-size can be set as $\eta^{(t)} = \frac{N}{\mu(t+1)}$.*

Consider primal sub-optimality $\epsilon_P^{(t)} := P(w^{(t)}) - P(\bar{w})$. Since $\epsilon_P^{(t)} \le \epsilon_{PD}^{(t)}$ always holds, the convergence rates in Theorem 4 are applicable to the primal sub-optimality as well. An interesting observation is that these convergence results on $\epsilon_P^{(t)}$ are not relying on the Restricted Isometry Property (RIP) (or restricted strong condition number) which is required in most existing analysis of IHT-style algorithms (Blumensath & Davies, 2009; Yuan et al., 2014). In (Jain et al., 2014), several relaxed variants of IHT-style algorithms are presented for which the estimation consistency can be established without requiring the RIP conditions. In contrast to the RIP-free sparse recovery analysis in (Jain et al., 2014), our Theorem 4 does not require the sparsity level k to be relaxed.

For SDIHT, we can establish similar non-asymptotic convergence results as summarized in the following theorem.

Theorem 5. *Assume that l_i is $1/\mu$-smooth. Set $\eta^{(t)} = \frac{\lambda m N^2}{(\lambda N \mu + \sigma_{\min}(X, k))(t+1)}$.*

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

$$\mathbb{E}[\|\alpha^{(t)}-\bar{\alpha}\|^2]\le mc_1\left(\frac{1}{t}+\frac{\ln t}{t}\right),$$

(b) **Support recovery and primal-dual gap**: *Assume additionally that $\bar{\epsilon}:=\bar{w}_{\min}-\frac{1}{\lambda}\|P'(\bar{w})\|_\infty>0$. Then, for any $\delta\in(0,1)$, with probability at least $1-\delta$, it holds that supp$(w^{(t)})=$ supp$(\bar{w})$ when*

$$t\ge t_2=\left\lceil\frac{12mc_1\sigma^2_{\max}(X)}{\lambda^2\delta^2N^2\bar{\epsilon}^2}\ln\frac{12mc_1\sigma^2_{\max}(X)}{\lambda^2\delta^2N^2\bar{\epsilon}^2}\right\rceil.$$

Moreover, with probability at least $1-\delta$, the primal-dual gap satisfies $\epsilon^{(t)}_{PD}\le\epsilon$ when $t\ge\max\{4t_2,t_3\}$ where $t_3=\left\lceil\frac{12mc_1c_2}{\lambda^2\delta^2N\epsilon^2}\ln\frac{12mc_1c_2}{\lambda^2\delta^2N\epsilon^2}\right\rceil$.

Proof. A proof of this result is given in Appendix A.6. □

Remark 4. *Theorem 5 shows that, up to scaling factors, the expected or high probability iteration complexity of S-DIHT is almost identical to that of DIHT. The scaling factor m appeared in t_2 and t_3 reflects a trade-off between the decreased per-iteration cost and the increased iteration complexity.*

5. Experiments

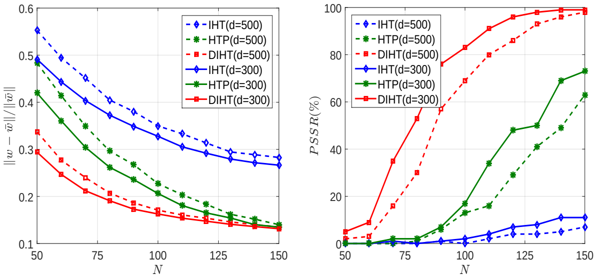
This section dedicates in demonstrating the accuracy and efficiency of the proposed algorithms. We first show the model estimation performance of DIHT when applied to sparse ridge regression models on synthetic datasets. Then we evaluate the efficiency of DIHT/SDIHT on sparse ℓ_2-regularized Huber loss and Hinge loss minimization tasks using real-world datasets.

5.1. Model parameter estimation accuracy evaluation

A synthetic model is generated with sparse model parameter $\bar{w}=[\underbrace{1,1,\cdots,1}_{\bar{k}},\underbrace{0,0,\cdots,0}_{d-\bar{k}}]$. Each $x_i\in\mathbb{R}^d$ of the N training data examples $\{x_i\}_{i=1}^N$ is designed to have two components. The first component is the top-$\bar{k}$ feature dimensions drawn from multivariate Gaussian distribution $N(\mu_1,\Sigma)$. Each entry in $\mu_1\in\mathbb{R}^{\bar{k}}$ independently follows standard normal distribution. The entries of covariance $\Sigma_{ij}=\left\{\begin{array}{ll}1 & i=j\\ 0.25 & i\ne j\end{array}\right.$. The second component consists the left $d-\bar{k}$ feature dimensions. It follows $N(\mu_2,I)$ where each entry in $\mu_2\in\mathbb{R}^{d-\bar{k}}$ is drawn from standard normal distribution. We simulate two data parameter settings: (1) $d=500,\bar{k}=100$; (2) $d=300,\bar{k}=100$. In each data parameter setting 150 random data copies are produced independently. The task is to solve the following ℓ_2-regularized sparse linear regression problem:

$$\min_{\|w\|\le k}\frac{1}{N}\sum_{i=1}^N l_{sq}(y_i,w^\top x_i)+\frac{\lambda}{2}\|w\|^2,$$

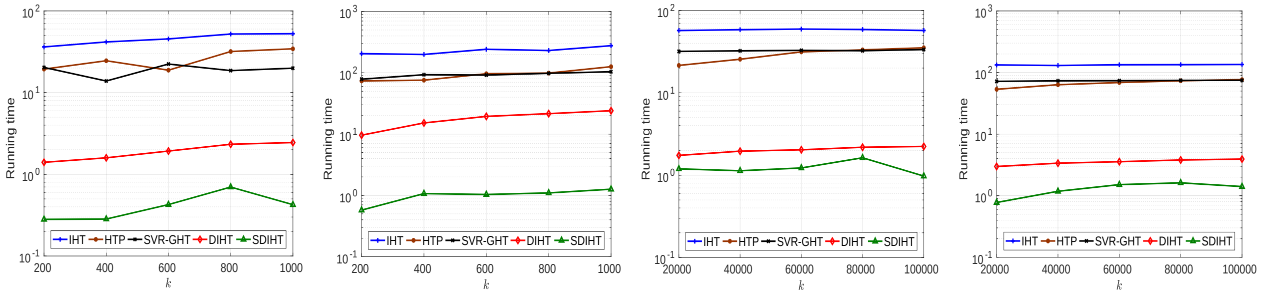
where $l_{sq}(y_i,w^\top x_i)=(y_i-w^\top x_i)^2$. The responses $\{y_i\}_{i=1}^N$ are produced by $y_i=\bar{w}^\top x_i+\varepsilon_i$, where $\varepsilon_i\sim N(0,1)$. The convex conjugate of $l_{sq}(y_i,w^\top x_i)$ is known as $l^*_{sq}(\alpha_i)=\frac{\alpha_i^2}{4}+y_i\alpha_i$ (Shalev-Shwartz & Zhang, 2013b). We consider solving the problem under the sparsity level $k=\bar{k}$. Two measurements are calculated for evaluation. The first is *parameter estimation error* $\|w-\bar{w}\|/\|\bar{w}\|$. Apart from it we calculate the *percentage of successful support recovery* ($PSSR$) as the second performance metric. A successful support recovery is obtained if $supp(\bar{w})=supp(w)$. The evaluation is conducted on the generated batch data copies to calculate the percentage of successful support recovery. We use 50 data copies as validation set to select the parameter λ from $\{10^{-6},...,10^2\}$ and the percentage of successful support recovery is evaluated on the other 100 data copies.



(a) Model estimation error (b) Percentage of support recovery success

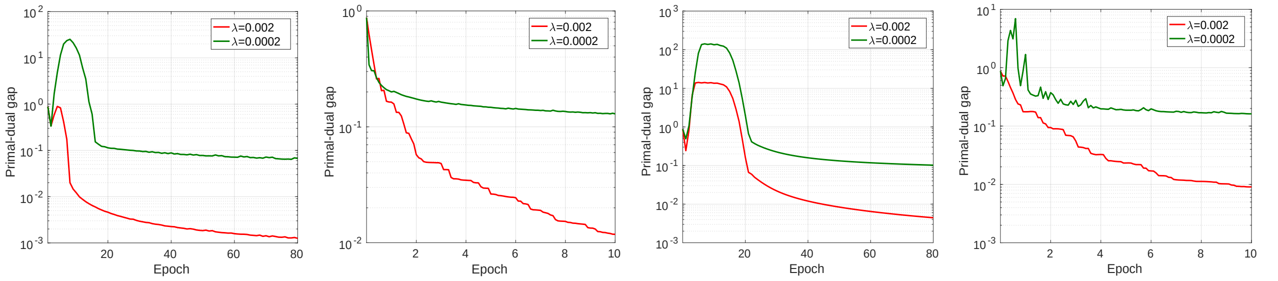
Figure 1. Model parameter estimation performance comparison between DIHT and baseline algorithms on the two synthetic dataset settings. The varying number of training sample is denoted by N.

Iterative hard thresholding (IHT) (Blumensath & Davies, 2009) and hard thresholding pursuit (HTP) (Foucart, 2011) are used as the baseline primal algorithms. The parameter estimation error and percentage of successful support recovery curves under varying training size are illustrated in Figure 1. We can observe from this group of curves that DIHT consistently achieves lower parameter estimation error and higher rate of successful support recovery than IHT and HTP. It is noteworthy that most significant performance gap between DIHT and the baselines occurs when the training size N is comparable to or slightly smaller than the sparsity level $\bar{k}$.



(a) RCV1, $\lambda = 0.002$ (b) RCV1, $\lambda = 0.0002$ (c) News20, $\lambda = 0.002$ (d) News20, $\lambda = 0.0002$

Figure 2. Huber loss: Running time (in second) comparison between the considered algorithms.



(a) DIHT on RCV1 (b) SDIHT on RCV1 (c) DIHT on News20 (d) SDIHT on News20

Figure 3. Huber loss: The primal-dual gap evolving curves of DIHT and SDIHT. $k = 600$ for RCV1 and $k = 60000$ for News20.

5.2. Model training efficiency evaluation

5.2.1. Huber loss model learning

We now evaluate the considered algorithms on the following ℓ_2-regularized sparse Huber loss minimization problem:

$$\min_{\|w\|_0 \le k} \frac{1}{N}\sum_{i=1}^{N} l_{Huber}(y_i x_i^\top w) + \frac{\lambda}{2}\|w\|^2, \qquad (10)$$

where

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

It is known that (Shalev-Shwartz & Zhang, 2013b)

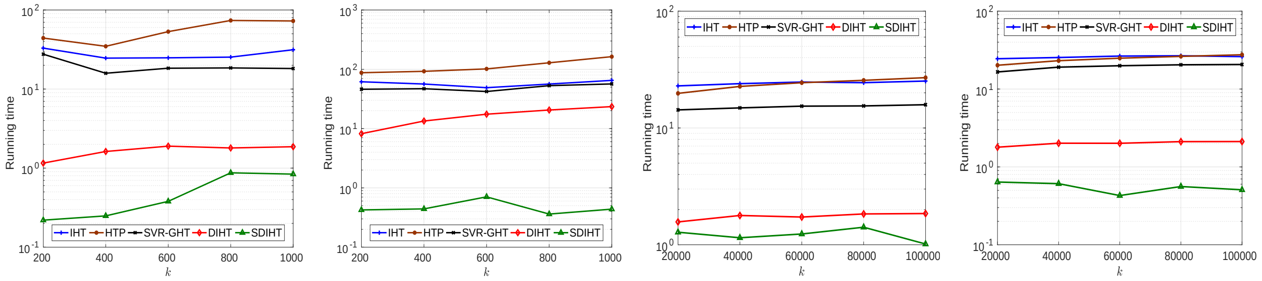
$$l^*_{Huber}(\alpha_i) = \begin{cases} y_i\alpha_i + \frac{\gamma}{2}\alpha_i^2 & \text{if } y_i\alpha_i \in [-1, 0] \\ +\infty & \text{otherwise} \end{cases}.$$

Two binary benchmark datasets from LibSVM data repository[1], RCV1 ($d = 47,236$) and News20 ($d = 1,355,191$), are used for algorithm efficiency evaluation and comparison. We select 0.5 million samples from RCV1 dataset for model training ($N \gg d$). For news20 dataset, all of the $19,996$ samples are used as training data ($d \gg N$).

We evaluate the algorithm efficiency of DIHT and SDIHT by comparing their running time against three primal baseline algorithms: IHT, HTP and gradient hard thresholding with stochastic variance reduction (SVR-GHT) (Li et al., 2016). We first run IHT by setting its convergence criterion to be $\frac{|P(w^{(t)}) - P(w^{(t-1)})|}{P(w^{(t)})} \le 10^{-4}$ or maximum number of iteration is reached. After that we test the time cost spend by other algorithms to make the primal loss reach $P(w^{(t)})$. The parameter update step-size of all the considered algorithms is tuned by grid search. The parameter γ is set to be 0.25. For the two stochastic algorithms SDIHT and SVR-GHT we randomly partition the training data into $|B| = 10$ mini-batches.

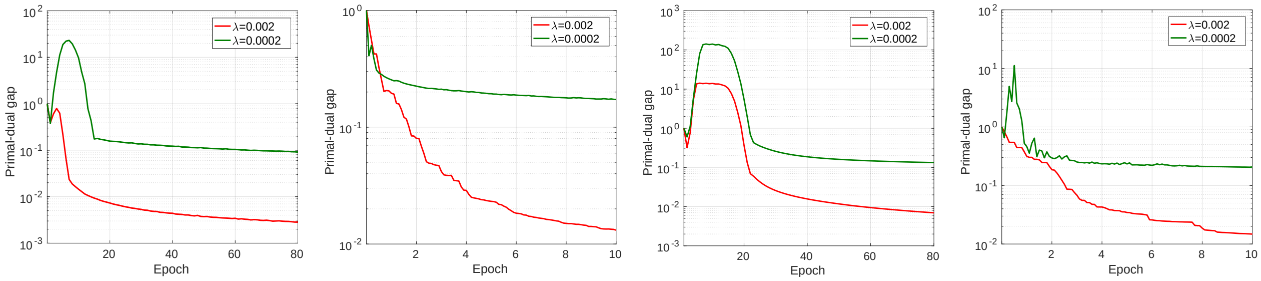
Figure 2 shows the running time curves on both datasets under varying sparsity level k and regularization strength $\lambda = 0.002, 0.0002$. It is obvious that under all tested (k, λ) configurations on both datasets, DIHT and SDIHT need much less time than the primal baseline algorithms, IHT, HTP and SVR-GHT to reach the same primal sub-optimality. Figure 3 shows the primal-dual gap convergence curves with respect to the number of epochs. This group of results support the theoretical prediction in Theorem 4 and 5 that ϵ_{PD} converges non-asymptotically.

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



(a) RCV1, $\lambda = 0.002$ (b) RCV1, $\lambda = 0.0002$ (c) News20, $\lambda = 0.002$ (d) News20, $\lambda = 0.0002$

Figure 4. Hinge loss: Running time (in second) comparison between the considered algorithms.



(a) DIHT on RCV1 (b) SDIHT on RCV1 (c) DIHT on News20 (d) SDIHT on News20

Figure 5. Hinge loss: The primal-dual gap evolving curves of DIHT and SDIHT. $k = 600$ for RCV1 and $k = 60000$ for News20.

5.2.2. Hinge loss model learning

We further test the performance of our algorithms when applied to the following ℓ_2-regularized sparse hinge loss minimization problem:

$$\min_{\|w\|_0 \le k} \frac{1}{N}\sum_{i=1}^{N} l_{Hinge}(y_i x_i^\top w) + \frac{\lambda}{2}\|w\|^2,$$

where $l_{Hinge}(y_i x_i^\top w) = \max(0, 1 - y_i w^\top x_i)$. It is standard to know (Hsieh et al., 2008)

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

We follow the same experiment protocol as in § 5.2.1 to compare the considered algorithms on the benchmark datasets. The time cost comparison is illustrated in Figure 4 and the prima-dual gap sub-optimality is illustrated in Figure 5. This group of results indicate that DIHT and SDIHT still exhibit remarkable efficiency advantage over the considered primal IHT algorithms even when the loss function is non-smooth.

6. Conclusion

In this paper, we systematically investigate duality theory and algorithms for solving the sparsity-constrained minimization problem which is NP-hard and non-convex in its primal form. As a theoretical contribution, we develop a sparse Lagrangian duality theory which guarantees strong duality in sparse settings, under mild sufficient and necessary conditions. This theory opens the gate to solve the original NP-hard/non-convex problem equivalently in a dual space. 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 stochastic variant of DIHT. For both algorithms we have proved sub-linear primal-dual gap convergence rate when the primal loss is smooth, without assuming RIP-style conditions. 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.

Acknowledgements

Xiao-Tong Yuan is supported in part by Natural Science Foundation of China (NSFC) under Grant 61402232, Grant 61522308, and in part by Natural Science Foundation of Jiangsu Province of China (NSFJPC) under Grant BK20141003. Qingshan Liu is supported in part by NSFC under Grant 61532009.

References

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

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

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

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

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

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

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

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

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

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

Jain, Prateek, Rao, Nikhil, and Dhillon, Inderjit. Structured sparse regression via greedy hard-thresholding. *arXiv preprint arXiv:1602.06042*, 2016.

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

Lee, Sangkyun, Brzyski, Damian, and Bogdan, Malgorzata. Fast saddle-point algorithm for generalized dantzig selector and fdr control with the ordered ℓ_1-norm. In *International Conference on Artificial Intelligence and Statistics*, 2016.

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

Nguyen, Nam, Needell, Deanna, and Woolf, Tina. Linear convergence of stochastic iterative greedy algorithms with sparse constraints. *arXiv preprint arXiv:1407.0088*, 2014.

Shalev-Shwartz, Shai and Zhang, Tong. Accelerated minibatch stochastic dual coordinate ascent. In *Advances in Neural Information Processing Systems*, 2013a.

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

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

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

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

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

Zhang, Yuchen and Xiao, Lin. Stochastic primal-dual coordinate method for regularized empirical risk minimization. In *International Conference on Machine Learning*, 2015.

Zhu, Zhanxing and Storkey, Amos J. Stochastic parallel block coordinate descent for large-scale saddle point problems. In *AAAI Conference on Artificial Intelligence*, 2016.