EFFICIENT HIGH-DIMENSIONAL DATA REPRESENTATION LEARNING VIA SEMI-STOCHASTIC BLOCK COORDINATE DESCENT METHODS

Anonymous authors
Paper under double-blind review

ABSTRACT

With the increase of data volume and data dimension, sparse representation learning attracts more and more attention. For high-dimensional data, randomized block coordinate descent methods perform well because they do not need to calculate the gradient along the whole dimension. Existing hard thresholding algorithms evaluate gradients followed by a hard thresholding operation to update the model parameter, which leads to slow convergence. To address this issue, we propose a novel hard thresholding algorithm, called Semi-stochastic Block Coordinate Descent Hard Thresholding Pursuit (SBCD-HTP). Moreover, we present its sparse and asynchronous parallel variants. We theoretically analyze the convergence properties of our algorithms, which show that they have a significantly lower hard thresholding complexity than existing algorithms. Our empirical evaluations on real-world datasets and face recognition tasks demonstrate the superior performance of our algorithms for sparsity-constrained optimization problems.

1 INTRODUCTION

In modern high-dimensional data analytics, where the variable dimension can be equal or even larger than the number of samples, sparse representation learning has become a mainstream method to explore the potential real model of the problem and provides statistically reliable results. It has been applied to many diverse domains such as high-dimensional statistics (Bühlmann & Van De Geer, 2011), signal processing (Lai et al., 2013), and computer vision (Wright et al., 2008). Many sparse learning algorithms such as \( \ell_1 \)-norm convex relaxation methods (Negahban et al., 2009; Van de Geer et al., 2008) have been proposed in the past few decades. In order to get a smaller estimation error, \( \ell_0 \)-norm constrained algorithms are more prominent than \( \ell_1 \)-norm convex relaxation algorithms (Zhang et al., 2010). In this paper, we mainly focus on the following sparsity-constrained optimization problem:

\[
\min_w F(w) := \frac{1}{n} \sum_{i=1}^{n} f_i(w), \quad \text{s.t., } \|w\|_0 \leq s,
\]

where \( F(w) \) is a finite-sum convex and smooth function, each function \( f_i(w) \) is associated with the \( i \)-th sample, \( \|w\|_0 \) is the number of nonzero entries in variable \( w \), and \( s \) represents the sparsity level. The goal of sparse representation learning is to recover \( w^* \) based on the given data. Such a formulation encapsulates plentiful important problems, including sparse graphical model learning (Zhou et al., 2019), sparse linear/logistic regression (Blumensath & Davies, 2009; Foucart, 2011; Pati et al., 1993; Bahmani et al., 2013), and low-rank regression (Rohde et al., 2011).

However, Problem (1) is in general NP hard, which is caused by the non-convexity of the sparsity constraint. It makes us to obtain an approximate solution to Problem (1). One of the most widely used methods for solving this problem is the hard thresholding based gradient descent method. In recent years, sparse representation learning is becoming more common for large-scale and high-dimensional data. However, deterministic gradient descent hard thresholding algorithms such as fast gradient descent hard thresholding (FG-HT) (Jain et al., 2014, Yuan et al., 2014) have to compute the gradients of all \( n \) component functions, which leads to a huge computing overhead. To address this issue, many stochastic hard thresholding methods have been proposed. For instance, Nguyen et al. (2017) proposed a stochastic gradient descent hard thresholding (SG-HT) algorithm, whose
To answer the above problem, we propose an efficient Semi-stochastic Block Coordinate Descent Hard Thresholding Pursuit (SBCD-HTP) algorithm and its sparse Asynchronous variant (ASBCD-HTP). The oracle complexities and statistical estimation error of the proposed algorithms and other hard thresholding methods are summarized in Table 1. We highlight several theoretical advantages of the proposed algorithms over the state-of-the-art methods as follows:

• SBCD-HTP and ASBCD-HTP substantially improve the restricted condition on the sparsity level \( s \), i.e., \( s = \Omega(\kappa_2 s^*) \), while most related algorithms such as SVRG-HT require \( s = \Omega(\kappa_2^2 s^*) \).
The statistical estimation error of both our algorithms is better than those of FG-HT, SVRG-HT and ASBCDHT, as our algorithms have a smaller sparsity level \( s = \Omega(\kappa_2 s^*) \) and thus have a smaller cardinality, i.e., \( \tilde{s} = 2s + s^* \). This makes the restricted condition number \( \kappa_2 \) of SBCD-HTP smaller than other algorithms. Moreover, the statistical error of our algorithms is also smaller than that of SG-HT (i.e., \( \frac{1}{2} \sum_{i=1}^{s} \| \nabla f_i(w^*) \| \) due to the large magnitude of \( \| \nabla f_i(w^*) \| \) (Zhou et al., 2018b).

Both SBCD-HTP and ASBCD-HTP enjoy a \( \kappa_2 \)-independent hard thresholding complexity, which is significantly better than the state-of-the-art hard thresholding algorithms. That is, the hard thresholding complexity of our algorithms is \( \kappa_2 \) times lower than that of FG-HT, SG-HT, SVRG-HT and ASBCDHT with \( \tilde{s} = 2\Omega(\kappa_2^2 s^*) + s^* \). Since both our algorithms have a significantly lower hard thresholding complexity, they are more suitable for handling large-scale sparse representation learning problems, especially for high-dimensional data.

For sparsity-constrained problems, the gradient oracle complexities of both SBCD-HTP and ASBCD-HTP (or \( S^2 \) BCD-HTP) are much lower than those of the state-of-the-art hard thresholding algorithms, and that of \( S^2 \) BCD-HTP is similar to ASBCD-HTP under conditions on the delay. Actually, the gradient oracle complexity of \( S^2 \) BCD-HTP is better than SBCD-HTP when we deal with sparse data sets, since \( S^2 \) BCD-HTP only evaluates the common nonzero point of the random sample and the sampled block. Furthermore, their gradient oracle complexities are significantly lower than those of ASBCDHT and SVRG-HT, and much lower than that of FG-HT, since they require a smaller sparsity level \( s = \Omega(\kappa_2 s^*) \) and also have a smaller value of \( \kappa_2 \) with \( \tilde{s} = 2s + s^* \), compared with the restricted condition number \( \kappa_2 \) used in other hard thresholding algorithms (Zhou et al., 2018b).

2 RELATED WORK

In this section, we briefly discuss the relevant research to our work, which beyonds the sparsity-constrained optimization domain. Traditional gradient descent is computational expensive at each iteration, and stochastic gradient descent has a low per-iteration complexity while obtaining a large variance in estimating. Thus, many stochastic variance reduced algorithms (Defazio et al., 2014; Roux et al., 2012; Shalev-Shwartz & Zhang, 2013; Shang et al., 2018; Johnson & Zhang, 2013) and their variants (Schmidt et al., 2017; Konečný et al., 2015; Xiao & Zhang, 2014; Hannah et al., 2018) have been proposed. The stochastic variance reduced methods are very promising for many machine learning problems including sparse learning problems.

In contrast to gradient descent methods, coordinate descent algorithms have received increasing attention due to their successful applications in high dimensional problems (Breheny & Huang, 2011; Friedman et al., 2007). Among them, randomized block coordinate descent (RBCD) updates a block of coordinate with respect to the entire training instances. The per-iteration cost is significantly lower than gradient descent, while it still has a relationship with \( n \) component functions. Some stochastic block coordinate descent (SBCD) algorithms such as (Dang & Lan, 2015; Xu & Yin, 2015; Konečný et al., 2017) were proposed. Such algorithms compute the stochastic gradient with respect to a random sample restricted to one randomized coordinate block. Therefore, these algorithms randomly sample a block of features and data instances at each iteration. However, they can only achieve a sublinear convergence rate (Zhang & Gu, 2016). Recently, some mini-batch randomized block coordinate descent algorithms such as (Zhao et al., 2014; Wang & Banerjee, 2014) were proposed to accelerate the convergence of stochastic block coordinate gradient descent. Our work studies over the above research by considering a sparsity-constrained optimization problem. The proposed algorithms enjoy a lower computational complexity in both gradient evaluation and hard thresholding computation while obtaining a linear convergence performance. Moreover, its sparse and asynchronous variants are also proposed to deal with sparse high-dimensional data.

3 PRELIMINARIES

Throughout this paper, we use \( w^* \) to denote the optimal solution of Problem (1), and its optimal sparsity level is \( s^* \) with \( \| w^* \|_0 \leq s^* \). \( \| w \|_\infty \) is the Euclidean norm for a vector \( w \in \mathbb{R}^d \) and \( \| w \|_\infty \) is the largest absolute entry in \( w \). The hard thresholding operation \( HT(w, s) \) preserves the \( s \) largest entries of \( w \) in magnitude for vector \( w \) and the rest entries are set to zero. We use \( \text{supp}(w) \) to denote the support set of \( w \), i.e., the indices of its nonzero entries. Given an index set \( \mathcal{I} \), we define \( \mathcal{I}^c \) as the complement set of \( \mathcal{I} \), and \( v_{\mathcal{I}} \in \mathbb{R}^d \), where \( [v_{\mathcal{I}}]_j = v_j \) if \( j \in \mathcal{I} \) and \( [v_{\mathcal{I}}]_j = 0 \) if \( j \notin \mathcal{I} \). Given an integer
A differentiable function $F(\cdot)$ is restricted $\rho^-_s$-strongly convex at sparsity level $s$, if there exists a uniform constant $\rho^-_s > 0$ such that for any $w, w' \in \mathbb{R}^d$ with $\|w - w'\|_0 \leq \hat{s}$, we have
\[
F(w) - F(w') - \langle \nabla F(w'), w - w' \rangle \geq \frac{\rho^-_s}{2} \|w - w'\|^2.
\]

Assumption 2 (Restricted Strong Smoothness) For any $i \in [n]$, the differentiable function $f_i(\cdot)$ is restricted $\rho^+_s$-strongly smooth at sparsity level $s$, if there exists a uniform constant $\rho^+_s > 0$ such that for any $w, w' \in \mathbb{R}^d$ with $\|w - w'\|_0 \leq \hat{s}$, we have
\[
f_i(w) - f_i(w') - \langle \nabla f_i(w'), w - w' \rangle \leq \frac{\rho^+_s}{2} \|w - w'\|^2.
\]

Algorithm 1: Semi-stochastic Block Coordinate Descent Hard Thresholding Pursuit (SBCD-HTP)

Input: Number of outer-loops $R$, number of inner-loops $m$, step size $\eta$, sparsity level $s$.

Initialize: $\tilde{w}^0$.

1: for $r = 0, 1, \ldots, R - 1$ do
2: \hspace{1em} $w^0 = \tilde{w} = \tilde{w}^r$, $\nabla F(\tilde{w}) = \frac{1}{n} \sum_{i=1}^n \nabla f_i(\tilde{w})$, $\tilde{\mathcal{S}} = \text{supp}(\tilde{w})$;
3: for $t = 0, 1, \ldots, m - 1$ do
4: \hspace{2em} Randomly sample a mini-batch $\mathcal{B}$ from $[n]$ uniformly;
5: \hspace{2em} Randomly sample a block $j_t$ from $[k]$ uniformly, and $\mathcal{S} = \tilde{\mathcal{S}} \cup \mathcal{G}_{j_t}$;
6: \hspace{2em} $\nabla_S g(w^t) = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \nabla_S f_i(w^t) - \nabla_S f_i(\tilde{w}) + \nabla_S F(\tilde{w})$;
7: \hspace{2em} $w^{t+1} = w^t - \eta \nabla_S g(w^t)$;
8: end for
9: $\tilde{w}^{r+1} = HT(w^m, s)$;
10: end for

Output: $\tilde{w}^R$.

$n \geq 1$, we define $[n] = \{1, \cdots, n\}$. For a set $\mathcal{B}$, we denote its cardinality by $|\mathcal{B}|$. Moreover, we use the common notations of $\Omega(\cdot)$ and $O(\cdot)$ to characterize the asymptotics of two real sequences.

Throughout the analysis, we make two important assumptions on the objective function, which are commonly used in the analysis of hard thresholding algorithms (Li et al. [2016b]; Shen & Li [2017]; Chen & Gu [2016]; Gao & Huang [2018]). In high-dimensional sparse learning, the per-iteration hard thresholding operation can be time-consuming or even more expensive than gradient computation. Thus, we also take the complexity of hard thresholding into our consideration.

Assumption 1 (Restricted Strong Convexity) A differentiable function $F(\cdot)$ is restricted $\rho^-_s$-strongly convex at sparsity level $\hat{s}$, if there exists a uniform constant $\rho^-_s > 0$ such that for any $w, w' \in \mathbb{R}^d$ with $\|w - w'\|_0 \leq \hat{s}$, we have
\[
F(w) - F(w') - \langle \nabla F(w'), w - w' \rangle \geq \frac{\rho^-_s}{2} \|w - w'\|^2.
\]

Assumption 2 (Restricted Strong Smoothness) For any $i \in [n]$, the differentiable function $f_i(\cdot)$ is restricted $\rho^+_s$-strongly smooth at sparsity level $\hat{s}$, if there exists a uniform constant $\rho^+_s > 0$ such that for any $w, w' \in \mathbb{R}^d$ with $\|w - w'\|_0 \leq \hat{s}$, we have
\[
f_i(w) - f_i(w') - \langle \nabla f_i(w'), w - w' \rangle \leq \frac{\rho^+_s}{2} \|w - w'\|^2.
\]

Moreover, the restricted condition number is defined as: $\kappa_s = \rho^+_s / \rho^-_s$. Since the complexity of one hard thresholding operation is linear with $d$, which is similar to a computational cost of a stochastic gradient, we define the hard thresholding complexity as a metric for hard thresholding algorithms.

Definition 1 (Hard Thresholding Complexity) In a hard thresholding operation, a vector $w$ is fed into $HT(\cdot, s)$ and then the output $HT(w, s)$ is obtained.

Both gradient oracle and hard thresholding complexities can more comprehensively reflect the overall computational cost of first-order hard thresholding algorithms. Therefore, their per-iteration cost is dominated by both the gradient evaluation and hard thresholding operations.

We also define the optimization error and statistical error to clearly understand our theoretical results.

Definition 2 (Optimization Error and Statistical Error) The optimization error is the difference which will decrease with the increasing of iterations. A statistical error is the (unknown) difference between the returned value and the true value.

This means that the optimization error will decrease to zero when the number of iteration is quite large. However, a statistical error will still work during the optimized process. Therefore, the larger the statistical error, the more difference to the true value.
4 SEMI-STOCHASTIC BLOCK COORDINATE DESCENT HARD THRESHOLDING PURSUIT

In this section, we propose a novel Semi-stochastic Block Coordinate Descent Hard Thresholding Pursuit (SBCD-HTP) algorithm, and also theoretically analyze its convergence properties.

4.1 OUR SBCD-HTP ALGORITHM

The proposed SBCD-HTP algorithm is summarized in Algorithm 1. In each outer-loop, we select a snapshot point \( \overline{w} \) and compute its full gradient \( \nabla F(\overline{w}) \). In each inner-loop, we uniformly and randomly select a mini-batch samples \( B \) and a block of coordinates \( \mathcal{G}_{j_t} \), where \( j_t \in \{1, 2, \ldots, k\} \). \( \{\mathcal{G}_1, \ldots, \mathcal{G}_k\} \) is a partition of all the \( d \) coordinates and is divided uniformly at random. We usually set \( k = 10 \) in our experiments. Note that, different from ASBCDHT, in line 5 we set \( \mathcal{S} \) to be a union set of the support set of snapshot point \( \overline{w} \) and \( \mathcal{G}_{j_t} \). In this way, the coordinate of the support set \( \mathcal{S} \) which includes the possibly optimal coordinates can be updated, and thus this can reduce the error caused by randomness. Moreover, SBCD-HTP enjoys the dual advantages of randomized block coordinate descent and semi-stochastic gradient descent to optimize the objective as shown in line 6. Note that we perform the hard thresholding operation in only the outer-loop, which is different from all existing hard thresholding algorithms. Performing hard thresholding operations with high frequency or prematurely in inner-loop will make the gradient information lost, which is the main reason for slow convergence and low accuracy of existing hard thresholding sparsity-constrained algorithms. Our hard thresholding algorithm not only guarantees the sparsity of model parameters, but also uses the full dimensional information without the interference of hard thresholding operations to update model parameter \( w^{t+1} \). Our theoretical analysis and experimental results show that SBCD-HTP has a faster convergence rate and more accurate results.

We have to clarify the difference between the proposed SBCD-HTP and ASBCDHT in Chen & Gu (2016) for clear improvements. There are mainly two changes: a) we design a bigger support set with the union of the sampled block \( \mathcal{G}_{j_t} \) with the one of the current support of the snapshot point \( \overline{w} \), since the current support of snapshot point contain the most related optimized information, while that of Chen & Gu (2016) only use the sampled block. b) we move the hard thresholding operation outside of the inner-loop, since the truncated function (hard thresholding) may loss many information and bringing higher complexity. Moreover, the asynchronous variant of SBCD-HTP is also proposed in this paper to optimize the sparse problem in the sparse data. This is the first to design this type of algorithm in sparse learning field, and the theoretical proof is also provided.

4.2 CONVERGENCE ANALYSIS

We first analyze the convergence behavior of SBCD-HTP. The main result is summarized in the following theorem, whose proof is provided in Section C.

**Theorem 1** Suppose \( F(w) \) is \( \rho^+_{z} \)-strongly convex and each function \( f_s(w) \) is \( \rho_{z}^{-} \)-strongly smooth with parameter \( s = 2s + \delta \). Let \( \kappa_{\delta} = \frac{\rho_{z}^{-}}{s} \) and the sparsity level \( s \geq \Omega(\kappa_{\delta} \delta) \). In addition, assume that the learning rate \( \eta \leq \frac{1}{4\rho_{z}^{-}} \), the number of inner-loops \( m \geq 1800\kappa_{\delta} \), and the number of blocks \( k = 10 \), then the convergence rate \( \gamma = \frac{k}{2\eta \rho_{z}^{-} (1-\omega)} + \frac{8\rho_{z}^{-} \eta (n-|B|)}{|B| (1-\omega) (n-1)} \leq \frac{1}{2} \). For the sparsity-constrained problem (1), the output of Algorithm 1 satisfies

\[
\mathbb{E}[F(\tilde{w}^*) - F(w^*)] \leq \frac{1}{2} \mathbb{E}[F(\tilde{w}^0) - F(w^*)] + \frac{\eta}{1-\omega} \|\nabla F(w^*)\|^2, \tag{4}
\]

where \( \tilde{w}^* = \supp(HT(\nabla F(w^*), 2s)) \cup \supp(w^*) \) and \( \omega = 8\rho_{z}^{-} \eta (1 + \frac{n-|B|}{|B|(n-1)}) \).

Theorem 1 shows that SBCD-HTP enjoys a linear convergence rate for solving sparsity-constrained problem (1). In order to ensure that the linear converging term in Eq. (4) satisfies \( \frac{1}{2} \mathbb{E}[F(\tilde{w}^0) - F(w^*)] \leq \epsilon \), we can obtain the following corollary.

**Corollary 1** Suppose the conditions in Theorem 1 hold. To achieve \( \frac{1}{2} \mathbb{E}[F(\tilde{w}^0) - F(w^*)] \leq \epsilon \), the gradient oracle complexity of SBCD-HTP is \( \mathcal{O}((n + \frac{\kappa_{\delta} |B|}{k}) \log(\frac{1}{\epsilon})) \), and its hard thresholding complexity is \( \mathcal{O}(\log(\frac{1}{\epsilon})) \).
Remark 1  This corollary is inferred under the high-dimensional condition. Actually, the worst condition of the gradient oracle complexity of SBCD-HTP is $O((n + \frac{n\kappa}{k}) \log(\frac{1}{\epsilon}))$. Due to the small sparsity level $s$ required in our Theorem 1 and the same oracle complexity as ASBCD-HTP. The detailed description and analysis HTP (S$\tilde{2}$Ω($\frac{1}{s}$))

The gradient oracle complexity of SBCD-HTP is $O((n + \frac{n\kappa}{k}) \log(\frac{1}{\epsilon}))$. The hard thresholding complexity of SBCD-HTP is $O(|B|)$ and is much lower than ASBCDHT, as $\kappa$s of SBCD-HTP is usually comparable to or even smaller than those of others (note that $s$ in $s = 2\kappa + s^*$ is required to be smaller than others as shown below). Moreover, SBCD-HTP allows $s = \Omega(k\kappa^2 s^*)$, which is considerably superior to the condition of $s = \Omega(k\kappa s^*)$ required in other hard thresholding algorithms such as (Zhou et al., 2018b). In particular, with the same parameter settings of $[B]$ and $k$, the gradient oracle complexity of our algorithm outperforms ASBCDHT. SBCD-HTP also attains a lower hard thresholding complexity than the state-of-the-art hard thresholding methods. That is, the hard thresholding complexity of SBCD-HTP is $O(\log(\frac{1}{\epsilon}))$, which is $\kappa^2$-independent and is significantly lower than those of other methods. This is a significant improvement on hard thresholding complexity. Theorem 1 immediately implies the following results.

Corollary 2  Suppose the conditions in Theorem 1 hold. The output of Algorithm 1 satisfies

$$\mathbb{E}\|\bar{w}^r - w^*\| \leq \sqrt{\frac{2(\frac{1}{2})^r[F(\bar{w}^0) - F(w^*)]}{\rho^r_w}} + \frac{2\eta}{1 - \omega} \sqrt{s}\|\nabla F(w^*)\|_\infty.$$  (5)

The right-hand side of Eq. (5) consists of two terms. The first term is the optimization error, which approaches zero with the increase of $r$. The second term corresponds to the statistical error, which is proportional to $\sqrt{s}\|\nabla F(w^*)\|_\infty$. One can also observe that the statistical error bound is better than those of FG-HT and ASBCDHT, since $\tilde{s}$ in the statistical error $O(\sqrt{s}\|\nabla F(w^*)\|_\infty)$ has a smaller cardinality (i.e., $2\Omega(k\kappa^2 s^*) + s^*$), while existing algorithms require a larger sparsity level $s' = \Omega(k\kappa^2 s^*)$ and have a larger cardinality. Therefore, compared with other existing algorithms, SBCD-HTP enjoys a smaller statistical error, especially for the problems with a large restricted condition number. It is usually better than the error bound $O(\sqrt{s}\|\nabla F(w^*)\|_\infty + \|\nabla T F(w^*)\|)$ with $\tilde{s} = 2\Omega(k\kappa^2 s^*) + s^*$ in SVRG-HT. Moreover, the error bound of SBCD-HTP is much smaller than that of SG-HT (Nguyen et al., 2017) (i.e., $O(\frac{1}{\sqrt{s}} \sum_{i=1}^n \|\nabla f_i(w^*)\|)$, since the magnitude of the individual gradient norm $\|\nabla f_i(w^*)\|$ can still be relatively large (Zhou et al., 2018b).

5  Asynchronous Variant of SBCD-HTP

In this section, we propose the serial sparse and asynchronous parallel variants of SBCD-HTP for high-dimensional sparse data sets. Our Asynchronous Sparse Semi-stochastic Block Coordinate Descent Hard Thresholding Pursuit (ASBCD-HTP) algorithm is summarized in Algorithm 2. The main difference between ASBCD-HTP and SBCD-HTP is sparse approximate gradients as in (Mannia et al., 2017) Zhou et al., 2018a). In order to perform fully sparse updates, we use a diagonal matrix $D$ to re-weigh the dense fully gradient $\nabla F(\bar{w})$. The entries of $D$ are the inverse probabilities of the corresponding coordinates belonging to a uniformly sampled support $T_i$ of sample $i$. Let $P_i$ be the projection matrix for the support $T_i$, we can get $D_i$ since $D_i = P_i D$. Then we can find that $\mathbb{E}_i[D_i \nabla F(\bar{w})] = \nabla F(\bar{w})$. In this setting, we can compute the full gradient $\nabla F(\bar{w})$ also in a parallel way. In order to reduce thread interference and fully utilize the sparsity of datasets, we restrict the sparse gradient $\nabla g(\overline{[w_i]}_{T_i})$ to the intersection set of $S$ and $T_i$, and obtain $\nabla g(\overline{[w_i]}_{T_i})$. Specifically, each thread of ASBCD-HTP computes sparse approximate gradient independently and then atomically write this gradient to the shared variable $\bar{w}^r$. Note that Sparse variant of SBCD-HTP (S$^2$BCD-HTP) can be viewed as the single-thread version of ASBCD-HTP, which has linear convergence and the same oracle complexity as SBCD-HTP. The detailed description and analysis of S$^2$BCD-HTP is listed in Section D. Next we give the main theoretical result of ASBCD-HTP.
Theorem 2 Suppose $F(w)$ is $\rho^g_2$-strongly convex and each function $f_i(w)$ is $\rho^g_2$-strongly smooth with parameter $\hat{s} = 2s + s^*$. Let $\bar{s} = \frac{\hat{s}}{\rho^g_2}$ and the sparsity level $s \geq \Omega(\bar{s}s^*)$. We assume that the step size $\eta = \frac{1}{60\rho^g_2}$, the number of inner-loops $m \geq 120\bar{s}$, the number of blocks $k = 10$ and $\tau \leq \min\{\frac{3}{5\sqrt{N}}, 2\bar{s}\sqrt{\frac{2\bar{s}}{\sqrt{A}}}}\}$ (i.e., the linear speedup condition). Then for the sparsity-constrained problem $(\ref{eq:sparsity_constrained})$, the gradient oracle and hard thresholding complexities of Algorithm 2 are
\[ O((n + \bar{s})k \log(1/\epsilon)) \quad \text{and} \quad O(\log(1/\epsilon)), \]
where $\tau$ denotes the maximum number of concurrent threads ($\text{Mania et al., 2017}$) and $\Delta = \max_{j=1,...,d} p_j$, which is a indicator to measure the sparsity of datasets ($\text{Leblond et al., 2017}$).

Theorem 2 implies that ASBCD-HTP (or $S^2$BCD-HTP) can still obtain similar gradient oracle complexity and hard thresholding complexity as SBCD-HTP while obtaining a lower speedup ratio. Compared with ASVRG-HT (Li et al., 2016a), ASBCD-HTP has a much lower gradient complexity, i.e., $O((n + \bar{s})k \log(1/\epsilon))$ for ASBCD-HTP vs. $O((n + \bar{s})k \log(1/\epsilon))$ for ASVRG-HT. For hard thresholding complexity, ASBCD-HTP is $\bar{s}$ times lower than ASVRG-HT, which is a significant improvement in the asynchronous setting. Since ASVRG-HT has to compute hard thresholding operations at each inner-loop, the practical performance may be worse, especially for high-dimensional data. We also show this improvement by our empirical evaluations.

6 Experiments

In this section, we evaluate the performance of SBCD-HTP, $S^2$BCD-HTP and ASBCD-HTP for solving sparse linear regression and sparse logistic regression on real-world datasets. All the dense algorithms were implemented in MATLAB, while the sparse and asynchronous algorithms were implemented in C++ and executed through MATLAB interface for a fair comparison. We also test the performance for face recognition tasks. The detailed information is described in Section 4. Since there is no ground truth on real-world datasets, we run all these baselines sufficiently long until $\|w^r - w^{r+1}\|/\|w^r\| \leq 10^{-6}$, and then we use the minimum $F^c(w^r)$ as the approximate optimal value $F^*$. We use the approximate optimal value for sub-optimality estimation in our experimental results (Zhou et al., 2018b).

6.1 Dense Algorithms

We ran the experiments of all the dense algorithms on a PC with an Intel i7-7700 CPU and 32GB RAM. We compare SBCD-HTP with several state-of-the-art sparsity-constrained algorithms, including FG-HT (Yuan et al., 2014), SG-HT (Nguyen et al., 2017), SVRG-HT (Li et al., 2016b), ASBCDHT (Chen & Gu, 2016) and FNHTP (Chen & Gu, 2017). The detailed descriptions of the
algorithms and datasets are provided in Section F. We set the sparsity level \( s = 200 \) for real-world datasets. For the number of inner-loops, we set \( m = 2n \) for SBCD-HT and ASBCDHT, \( m = n \) for SVRG-HT as suggested in [Li et al. 2016b, Chen & Gu 2016] and 1000 (sparse linear regression)/3000 (sparse logistic regression) for FNHTP. We also set the batch-size of FNHTP to \( s \), and that of ASBCDHT and SBCD-HTP to 5. For the block coordinate descent methods, we set the number of coordinate blocks \( k \) to 10. All the algorithms are tuned to their best performance.

We report the performance of all the algorithms in terms of both effective passes and CPU time. More experimental results are presented in Section F. Figures 1 and 2 show that for both sparse linear regression and sparse logistic regression problems, our algorithm converges significantly faster than the baseline algorithms in terms of both effective passes and CPU time. The main reason is that we use all the information of gradients, while the related algorithms use less information caused by too frequent and premature hard thresholding operation in each inner-loop.

### 6.2 Asynchronous Algorithms

We ran the experiments of all the asynchronous algorithms on a PC with an Intel Xeon(R) Gold 5120 CPU and 64GB RAM. We compare our sparse and asynchronous algorithms with several state-of-the-art algorithms, including ASG-HT, ASVRG-HT and \( \Lambda^2 \) SBCD-HT (asynchronous variant of ABSCDHT). For the algorithms, we set the number of threads to 20.

We only report the performance of all the algorithms in terms of CPU time since the performance in terms of effective passes is similar to the dense case. More experimental results are reported in Section F. Figure 3 shows that ASBCD-HTP and \( S^2 \) BCD-HT significantly outperform than other algorithms. Moreover, ASBCD-HTP has a faster convergence performance than \( S^2 \) BCD-HT because of asynchronous parallel acceleration. There are three main reasons for the advantages of our ASBCD-HTP: a) We get the set \( S \) from the support set of snapshot point \( \bar{w} \) and \( \mathcal{C}_j \). In this way, the coordinate of the support set which appears as the possible optimal solution can be updated, and thus reducing the error caused by randomness. b) Because we put the hard thresholding operation in each outer-loop, the computation cost per thread of our algorithm at each inner-loop is less, while each thread of the baseline algorithms has a hard thresholding operation, which is expensive especially for high-dimensional datasets. c) Our algorithm updates \( \bar{w} \) with all information of gradients, while the baseline algorithms update \( \bar{w} \) with less information caused by too frequent and premature hard thresholding operation in each inner-loop. We also evaluate the improvement of asynchronous parallel by running the same passes with different numbers of threads. We calculate the speed-up...
ratio based on the running time of a single thread. From Figure 4, we can see that our ASBCD-HTP is accelerated by a nearly linear ratio.

7 Conclusion

In this paper, we proposed an efficient semi-stochastic block coordinate descent hard thresholding pursuit (SBCD-HTP) method for sparse representation learning. We proved that SBCD-HTP attains the gradient oracle complexity of $O\left(\left(n + \frac{\kappa}{\epsilon} \log \left( \frac{1}{\epsilon} \right) \right) \log \left( \frac{1}{\epsilon} \right) \right)$, and the hard thresholding complexity of $O\left(\log \left( \frac{1}{\epsilon} \right) \right)$, respectively, which is an improvement over the existing algorithms. Moreover, we also presented the sparse and asynchronous parallel variants of SBCD-HTP. As far as we know, our algorithms are the first to use the full information of gradient to update. This is an improvement on sparse representation learning because it can be easily extended to other sparse learning algorithms.

References


**APPENDIX**

**A  KEY LEMMAS**

**Lemma 1 ([Jain et al., 2014])** For any index set \( I \), any \( w \in \mathbb{R}^I \) and \( HT(\cdot, s) : \mathbb{R}^d \rightarrow \mathbb{R}^d \) be the hard thresholding operator, which keeps the largest \( s \) entries (in magnitude) and sets the other entries equal to zero. Then for any \( w^* \in \mathbb{R}^I \) such that \( \|w^*\|_0 \leq s^* \), we have

\[
\|HT(w, s) - w\|^2 \leq \frac{|I| - s}{|I| - s^*} \|w - w^*\|^2.
\]

**Lemma 2 ([Li et al., 2016])** Let \( w^* \in \mathbb{R}^d \) be the optimal sparse vector such that \( \|w^*\|_0 \leq s^* \) and \( HT(\cdot, s) : \mathbb{R}^d \rightarrow \mathbb{R}^d \) be the hard thresholding operator, which keeps the largest \( s \) entries (in magnitude) and sets the other entries equal to zero. Given \( s > s^* \) for any vector \( w \in \mathbb{R}^d \), we have

\[
\|HT(w, s) - w^*\|^2 \leq (1 + \frac{2\sqrt{s^*}}{\sqrt{s - s^*}}) \|w - w^*\|^2.
\]

**Lemma 3 ([Nesterov, 2004])** For any vector \( w, w' \in \mathbb{R}^d \), suppose \( F(\cdot) \) also satisfies RSC and RSS conditions, then the following inequality holds

\[
\|\nabla F(w) - \nabla F(w')\|^2 \leq 2\rho_s^+ [F(w) - F(w') + \langle \nabla F(w'), w - w' \rangle].
\]

**Lemma 4** Suppose \( F(w) \) is \( \rho^+ \)-strongly convex and each function \( f_i(w) \) is \( \rho^+_i \)-strongly smooth with parameter \( \delta = 2s + s^*, \) \( I = \text{supp}(w^*) \cup \text{supp}(\tilde{w}^r) \cup \text{supp}(\tilde{w}^{r+1}) \), \( |B| \) is the batch size. For any \( w' \in \mathbb{R}^d \) and the sample \( i_t \in B \), denote \( \nu_i = (\nabla f_i(w') - (\nabla f_i(w^*)) - (\nabla F(w') - \nabla F(w^*)) \), we have

\[
\mathbb{E} \left[ \frac{1}{|B|} \sum_{i_t \in B} \nu_i \right]^2 = \frac{n - |B|}{|B| (n - 1)} \mathbb{E} \|\nabla f_i(w') - \nabla f_i(w^*) + \nabla F(w^*) - \nabla F(w')\|^2. \tag{7}
\]

**Proof.** See Section B.1 for the proof of Lemma 2.

**Lemma 5** Suppose \( F(w) \) is \( \rho^+_\tilde{s} \)-strongly convex and each function \( f_i(w) \) is \( \rho^+_i \)-strongly smooth with parameter \( \tilde{s} = 2s + s^* \). Let \( w^* \in \mathbb{R}^d \) be the optimal sparse vector with \( \|w^*\|_0 \leq s^* \). \( \tilde{I} = \text{supp}(w^*) \cup \text{supp}(\tilde{w}^r) \cup \text{supp}(\tilde{w}^{r+1}) \), \( \delta > 1 \) is a uniform constant factor, and \( |B| \) is the batch size. For any \( w' \in \mathbb{R}^d \) and the sample \( i_t \), denote \( \nabla g(w') = \frac{1}{|B|} \sum_{i_t \in B} \nabla f_i(w') + \nabla F(\tilde{w}) \), then we can bound \( \mathbb{E} \|\nabla g(w')\|^2 \) as follows:

\[
\mathbb{E} \|\nabla g(w')\|^2 \leq \frac{1}{k} (16\delta \rho^+_\tilde{s} + 1 + \frac{n - |B|}{|B| (n - 1)}) \mathbb{E} [F(w') - F(w^*)] + 4 \|\nabla F(w^*)\|^2 + 16\delta \rho^+_\tilde{s} \frac{n - |B|}{|B| (n - 1)} \mathbb{E} [F(\tilde{w}) - F(w^*)],
\]

**Proof.** See Section B.2 for the detailed proof of Lemma 5.

**Lemma 6** Suppose \( F(w) \) is \( \rho^+_{\tilde{s}_j} \)-strongly convex and each function \( f_i(w) \) is \( \rho^+_i \)-strongly smooth with parameter \( \tilde{s}_j = 2s + s^* \). Let \( w^* \in \mathbb{R}^d \) be the optimal sparse vector with \( \|w^*\|_0 \leq s^* \). \( \tilde{I} = \text{supp}(w^*) \cup \text{supp}(\tilde{w}^r) \cup \text{supp}(\tilde{w}^{r+1}) \), \( \delta > 1 \) is a uniform constant factor, and \( D_m = \max_i 1, \ldots, d \frac{1}{p_i} \). For any \( w' \in \mathbb{R}^d \) and the sample \( i_t \), denote \( \nabla g(w') = \nabla f_i(w') - \nabla f_i(\tilde{w}) + D_i \nabla F(\tilde{w}) \), we can bound \( \mathbb{E} \|\nabla g(w')\|^2 \) as follows:

\[
\mathbb{E} \|\nabla g(w')\|^2 \leq \frac{1}{k} (12\delta \rho^+_{\tilde{s}_j} [F(w') - F(w^*)] + (3 + 12(D_m^2 - 1)) \|\nabla F(w^*)\|^2
\]

\[
+ (24\delta \rho^+_{\tilde{s}_j} + 48\delta \rho^+_{\tilde{s}_j} (D_m^2 - 1)) [F(\tilde{w}) - F(w^*)]),
\]

where \( p_j \) is the probability that dimension \( j \) belongs to the support set of randomly sampling sample \( i_t \).
Proof. See Section B.3 for the proof of Lemma \[6\]. □

This lemma is different from other variance reduced results for sparsity-constrained problem (Mania et al., 2017; Johnson & Zhang, 2013; Chen & Gu, 2016; Shang et al., 2018). Since this variance reduced approximate gradients introduce an approximate full gradient to make fully sparse update in cardinality constraint problem.

B PROOFS OF KEY LEMMAS

In this section, we prove three key lemmas.

B.1 PROOF OF LEMMA 4

Proof. Let \( \nu_i = \langle \nabla \mathcal{I} f_i(w^t) - \nabla \mathcal{I} f_i(w^*) \rangle - \langle \nabla \mathcal{I} F(w^t) - \nabla \mathcal{I} F(w^*) \rangle \). We have

\[
\mathbb{E} \left\| \frac{1}{|B|} \sum_{i \in B} \nu_i \right\|^2 = \frac{1}{|B|^2} \mathbb{E} \sum_{i \neq i'} \nu_i^\top \nu_{i'} + \frac{1}{|B|} \mathbb{E} \| \nu_i \|^2
\]

(11)

This proof is similar to that of (Zheng & Kwok, 2016). □

B.2 PROOF OF LEMMA 5

Proof. It is obvious that the stochastic variance reduced gradient satisfies

\[
\mathbb{E}[\nabla g(w^t)] = \mathbb{E} \left[ \frac{1}{|B|} \sum_{i \in B} \langle \nabla f_i(w^t) - \nabla f_i(w^*) \rangle + \mu \right] = \nabla F(w^t).
\]

(12)

Thus, \( \nabla g(w^t) \) is an unbiased estimator of the full gradient \( \nabla F(w^t) \). Now, we bound \( \mathbb{E}[\| \nabla g_z(w^t) \|^2] \).

For any \( i \in \{1, 2, \cdots, n\} \), we define the following function

\[
h_i(w) = f_i(w) - f_i(w^*) - \langle \nabla f_i(w^*), w - w^* \rangle.
\]

(13)

It is easy to find that \( \nabla h_i(w^*) = 0 \), which implies that \( h_i(w^*) = \min_w h_i(w) \). For any vector \( w \), we have

\[
0 = h_i(w^*) \leq \min_{\eta} h_i(w - \eta \nabla \nabla h_i(w))
\]

\[
\leq \min_{\eta} h_i(w) - \eta \langle \nabla h_i(w), \nabla \nabla h_i(w) \rangle + \frac{\eta^2 \rho_i^+}{2} \| \nabla \nabla h_i(w) \|^2
\]

\[
= \min_{\eta} h_i(w) - \eta \| \nabla \nabla h_i(w) \|^2 + \frac{\eta^2 \rho_i^+}{2} \| \nabla \nabla h_i(w) \|^2
\]

\[
= h_i(w) - \frac{1}{2 \rho_i^+} \| \nabla \nabla h_i(w) \|^2,
\]

(14)

where the second inequality follows from the RSS condition, and the second equality holds due to the fact that \( \langle \nabla h_i(w), \nabla \nabla h_i(w) \rangle = \| \nabla \nabla h_i(w) \|^2 \), and the last equality holds due to the fact that \( \eta = \frac{1}{\rho_i^+} \) minimizes the function. Then, we have

\[
\| \nabla \mathcal{I} f_i(w) - \nabla \mathcal{I} f_i(w^*) \|^2 \leq 2 \rho_i^+ [f_i(w) - f_i(w^*) - \langle \nabla \mathcal{I} f_i(w^*), w - w^* \rangle].
\]

(15)
Since the sampling \(i\) is chosen uniformly from \(\{1, 2, \cdots, n\}\), we have
\[
\mathbb{E}\|\nabla_{\mathcal{I}} f_i(w) - \nabla_{\mathcal{I}} f_i(w^*)\|^2 = \mathbb{E}\|\nabla_{\mathcal{I}} F(w) - \nabla_{\mathcal{I}} F(w^*)\|^2 = \frac{1}{n} \sum_{i=1}^{n} \|\nabla_{\mathcal{I}} f_i(w) - \nabla_{\mathcal{I}} f_i(w^*)\|^2
\]
(16)
where the last inequality follows from the restricted strong convexity of \(F(w)\). And we use a simple underlying consensus that there must exist a uniform factor \(\delta > 1\) which makes the following inequality true:
\[
\delta (F(w) - F(w^*)) \geq |\langle \nabla_{\mathcal{I}} F(w^*), w - w^* \rangle| + \frac{\rho_s}{2} \|w - w^*\|^2.
\]
This property is called the extended restricted strong convexity used in the following proof. Therefore, we have
\[
\mathbb{E}\|\nabla_{\mathcal{I}} g_i(w^t)\|^2 = \mathbb{E}\left[\left|\frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \nabla_{\mathcal{I}} f_i(w^t) - \nabla_{\mathcal{I}} f_i(w) + \nabla_{\mathcal{I}} F(w)\right|^2\right]
\]
(17)
where the first inequality follows from \(|a + b + c + d|^2 \leq 4|a|^2 + 4|b|^2 + 4|c|^2 + 4|d|^2\), the second inequality follows from Lemma 4, the third equality follows from Lemma 4. The second inequality follows from Lemma 4. The first inequality follows from the restricted strong convexity of \(F(w)\). And we use a simple underlying consensus that there must exist a uniform factor \(\delta > 1\) which makes the following inequality true:
and the last inequality follows from \[16\]. Due to the fact that \(\mathbb{E}\|\nabla g_{\mathcal{S}G}(w^t)\|^2 \leq 1/k\mathbb{E}\|\nabla g_{\mathcal{S}}(w^t)\|^2\) where \(\mathcal{I} = \text{supp}(w^t) \cup \text{supp}(w^{t+1}) \cup \text{supp}(\bar{w}^t)\) and \(\mathcal{S} = \mathcal{G}_t \cup \text{supp}(\bar{w}^t)\), we have

\[
\mathbb{E}\|\nabla g_{\mathcal{S}G}(w^t)\|^2 \leq \frac{1}{k}(16\delta^2 + \frac{n-B}{B(n-1)})\mathbb{E}[F(w^t) - F(w^*)] + 4\|\nabla F(w^*)\|^2
\]

\[\text{(18)}\]

\(+ 16\delta^2 \mathbb{E}[F(\bar{w}) - F(w^*)] ),
\]

where \(k\) is the number of blocks. Note that we take expectation with respect to randomized block in the above equation. This completes the proof. \(\square\)

### B.3 Proof of Lemma 6

**Proof.** By the definition \(\nabla g(w^t) = \nabla f_i(w^t) - \nabla f_i(\bar{w}) + D_i \nabla F(\bar{w})\), we have \(\mathbb{E}[D_i \nabla F(\bar{w})] = \nabla F(\bar{w})\). Therefore, \(\nabla g(w^t)\) is an unbiased estimator of \(\nabla F(w^t)\).

Based on \([16]\), we have the following result:

\[
\mathbb{E}\|\nabla g_{\mathcal{S}}(w^t)\|^2 = \mathbb{E}\|\nabla F(w^t) - D_i \nabla F(\bar{w})\|^2
\]

\[\leq 3\mathbb{E}\|\nabla F(w^t) - D_i \nabla F(\bar{w})\|^2\]

\[+ 6\mathbb{E}\|\nabla F(w^t) - D_i \nabla F(\bar{w})\|^2\]

\[+ 3\mathbb{E}\|\nabla F(w^t) - D_i \nabla F(\bar{w})\|^2\]

\[\leq 12\delta^2 \mathbb{E}[F(w^t) - F(w^*)] + (3 + 12(D^2 - 1))\|\nabla F(w^*)\|^2\]

\[+ (24\delta^2 + 48\delta^2 (D^2 - 1))[F(w^t) - F(w^*)]
\]

where the first inequality follows from \(\|a + b + c\|^2 \leq 3\|a\|^2 + 3\|b\|^2 + 3\|c\|^2\), the second inequality follows from \(\|a + b\|^2 \leq 2\|a\|^2 + 2\|b\|^2\), the third inequality follows from \(\mathbb{E}\|w - \mathbb{E}w\|^2 \leq \mathbb{E}\|w\|^2\) with \(\mathbb{E}[\nabla F(w^t) - \nabla F(w^*)] = \nabla F(\bar{w}) - \nabla F(w^*)\), and \(\mathbb{E}\|\nabla F(w^t) - \nabla F(w^*)\|^2 \leq 4\delta^2 [F(w^t) - F(w^*)]\) with \(\delta > 1\), which is also satisfied for \(w\) with \(\delta > 1\). Since \(D^2 = \max_{v=1,...,d} \left( \frac{1}{p^2} - 1 \right)\|\nabla F(\bar{w})\|^2\)

\[\|F(\bar{w}) - F(w^t)\|^2\]

\[\leq (D^2 - 1)\|\nabla F(\bar{w})\|^2\]

\[\leq (D^2 - 1)\|\nabla F(\bar{w}) - \nabla F(w^*)\|^2\]

\[\leq 2(D^2 - 1)\|\nabla F(\bar{w}) - \nabla F(w^*)\|^2\]

\[\leq 8\delta^2 (D^2 - 1)[F(\bar{w}) - F(w^*)] + 2(D^2 - 1)\|\nabla F(w^*)\|^2\]

Due to the fact that \(\mathbb{E}\|\nabla g_{\mathcal{S}G}(w^t)\|^2 \leq \frac{1}{k}\mathbb{E}\|\nabla g_{\mathcal{S}G}(w^t)\|^2\), we have

\[
\mathbb{E}\|\nabla g_{\mathcal{S}G}(w^t)\|^2 \leq \frac{1}{k}(12\delta^2 \mathbb{E}[F(w^t) - F(w^*)] + (3 + 12(D^2 - 1))\|\nabla F(w^*)\|^2\]

\[+ (24\delta^2 + 48\delta^2 (D^2 - 1))[F(\bar{w}) - F(w^*)]
\]

where \(k\) is the number of blocks. Note that we take expectation with respect to randomized block in the above equation. This completes the proof. \(\square\)

### C Proofs of Theorem 1 and Corollaries 1 and 2

In this section, we give the detailed proofs of Theorem 1 and Corollaries 1 and 2 which can guarantee the convergence properties of our SBCD-HTP.
C.1 Proof of Theorem 1

Proof. In this subsection, we provide the proof of Theorem 1. Let \( w^{t+1} = w^t - \eta \nabla s g_{\mathcal{I}}(w^t) \) and \( \mathcal{I} = \mathcal{I}^* \cup \mathcal{I}^* \cup \mathcal{I}^{t+1} \), where \( \mathcal{I}^* = \text{supp}(w^*) \), \( \mathcal{I}^* = \text{supp}(\bar{w}^*) \) and \( \mathcal{I}^{t+1} = \text{supp}(\bar{w}^{t+1}) \). Conditioning on \( w^t \), we have the following expectation

\[
\mathbb{E}[\|w^{t+1} - w^*\|^2]
\]

\[
= \mathbb{E}[\|w^t - \eta \nabla s g_{\mathcal{I}}(w^t) - w^*\|^2]
\]

\[
= \mathbb{E}[w^t - w^*\|^2 + \eta^2 \mathbb{E}[\|\nabla s g_{\mathcal{I}}(w^t)\|^2 - 2\eta \langle w^t - w^*, \nabla s g_{\mathcal{I}}(w^t) \rangle]
\]

\[
\leq \mathbb{E}[w^t - w^*\|^2 + \eta^2 \mathbb{E}[\|\nabla s g_{\mathcal{I}}(w^t)\|^2] + \frac{2\eta}{k} \langle w^t - w^*, \nabla s F(w^t) \rangle
\]

\[
\leq \mathbb{E}[w^t - w^*\|^2 + \eta^2 \mathbb{E}[\|\nabla s g_{\mathcal{I}}(w^t)\|^2] + \frac{2\eta}{k} [F(w^*) - F(w^t)]
\]

\[
\leq \mathbb{E}[w^t - w^*\|^2 + \frac{1}{k}(16\rho^2 \eta^2 (1 + \frac{n - |\mathcal{B}|}{|\mathcal{B}|(n - 1)}) \mathbb{E}[F(w^t) - F(w^*)] + \frac{16\rho^2 \eta^2}{k} [F(w^*) - F(w^t)]
\]

\[
= \mathbb{E}[w^t - w^*\|^2 + \frac{1}{k}((16\rho^2 \eta^2 m (1 + \frac{n - |\mathcal{B}|}{|\mathcal{B}|(n - 1)}) - 2mn\delta) \mathbb{E}[F(w^m) - F(w^*)] + \frac{16\rho^2 \eta^2 m}{k} ||\nabla s F(w^*)||^2 + \frac{1}{k} ((16\rho^2 \eta^2 m (1 + \frac{n - |\mathcal{B}|}{|\mathcal{B}|(n - 1)}) - 2mn\delta) \mathbb{E}[F(w^m) - F(w^*)] + \frac{16\rho^2 \eta^2 m}{k} ||\nabla s F(w^*)||^2)
\]

where the first inequality follows from the expectation with respect to randomized block, the second inequality follows from our extended restricted strong convexity in the proof of Lemma 5 and the third inequality holds by Lemma 5. Notice that \( \bar{w} = w^0 = \bar{w}^{t-1} \). Summing \( \mathbb{E}[w^m - w^*\|^2 \leq \mathbb{E}[\bar{w}^{t-1} - w^*\|^2 + \frac{1}{k}((16\rho^2 \eta^2 m (1 + \frac{n - |\mathcal{B}|}{|\mathcal{B}|(n - 1)}) - 2mn\delta) \mathbb{E}[F(w^m) - F(w^*)] + \frac{16\rho^2 \eta^2 m}{k} ||\nabla s F(w^*)||^2
\]

\[
= \mathbb{E}[w^m - w^*\|^2 + \frac{1}{k}((16\rho^2 \eta^2 m (1 + \frac{n - |\mathcal{B}|}{|\mathcal{B}|(n - 1)}) - 2mn\delta) \mathbb{E}[F(w^m) - F(w^*)] + \frac{16\rho^2 \eta^2 m}{k} ||\nabla s F(w^*)||^2 + \frac{1}{k} ((16\rho^2 \eta^2 m (1 + \frac{n - |\mathcal{B}|}{|\mathcal{B}|(n - 1)}) - 2mn\delta) \mathbb{E}[F(w^m) - F(w^*)] + \frac{16\rho^2 \eta^2 m}{k} ||\nabla s F(w^*)||^2)
\]

Notice that the convergence property used here is similar to the proofs of the convergence results for many hard thresholding algorithms (Chen & Gu, 2016, 2017, Li et al., 2016b).

Moreover, we can obtain

\[
= \mathbb{E}[w^m - w^*\|^2 + \frac{1}{k}((16\rho^2 \eta^2 m (1 + \frac{n - |\mathcal{B}|}{|\mathcal{B}|(n - 1)}) - 2mn\delta) \mathbb{E}[F(w^m) - F(w^*)] + \frac{16\rho^2 \eta^2 m}{k} ||\nabla s F(w^*)||^2 + \frac{1}{k} ((16\rho^2 \eta^2 m (1 + \frac{n - |\mathcal{B}|}{|\mathcal{B}|(n - 1)}) - 2mn\delta) \mathbb{E}[F(w^m) - F(w^*)] + \frac{16\rho^2 \eta^2 m}{k} ||\nabla s F(w^*)||^2)
\]

(24)
where the second inequality holds due to the fact that we have the following results: When \( F(\bar{w}^r) - F(w^m) > 0 \), by using Lemma 3 on \( w^m \) and \( \bar{w}^r \), we have

\[
(\nabla F(w^m), \bar{w}^r - w^m) \leq F(\bar{w}^r) - F(w^m) - \frac{1}{2\rho_s^m} ||\nabla F(\bar{w}^r) - \nabla F(w^m)||^2
\]

where the last inequality holds due to the fact that we assume that there exists a constant factor \( \sigma > 0 \) making this inequality true. Then by using the RSS condition, we have

\[
F(\bar{w}^r) - F(w^m) \leq (\nabla F(w^m), \bar{w}^r - w^m) + \frac{\rho_s^m}{2} ||\bar{w}^r - w^m||^2
\]

and

\[
F(\bar{w}^r) - F(w^m) \leq (\nabla F(w^m), \bar{w}^r - w^m) + \frac{\rho_s^m}{2} ||\bar{w}^r - w^m||^2.
\]

After simplification, we have \( F(\bar{w}^r) - F(w^m) \leq \frac{\rho_s^m}{2\sigma} ||\bar{w}^r - w^m||^2 \). Using Lemma 1, we have

\[
||\bar{w}^r - w^m||^2 = ||H_k(w^m) - w^m||^2 \leq \frac{|I| - s}{|I| - s^*} ||w^m - w^*||^2,
\]

and

\[
F(\bar{w}^r) - F(w^m) \leq \frac{\rho_s^m (d - s)}{2\sigma (d - s^*)} ||w^m - w^*||^2.
\]

When \( F(\bar{w}^r) - F(w^m) < 0 \), we can omit \(-16\delta \rho_s^m \eta^2 m (1 + \frac{n - |B|}{|B|(n - 1)}) - 2 m \eta \delta \|F(\bar{w}^r) - F(w^m)|| \) directly since \( 16\delta \rho_s^m \eta^2 m (1 + \frac{n - |B|}{|B|(n - 1)}) - 2 m \eta \delta \) is less than zero. Because this is the simple situation, we only analyze the complex one in our whole proof. Therefore, we have the following inequality:

\[
(1 + \frac{1}{k} (16\delta \rho_s^m \eta^2 m (1 + \frac{n - |B|}{|B|(n - 1)}) - 2 m \eta \delta \frac{\rho_s^m (d - s)}{2\sigma (d - s^*)}) \|w^m - w^*\|^2
\]

\[
\leq \mathbb{E} \|\bar{w}^r - w^m\|^2 + \frac{1}{k} \left( 16\delta \rho_s^m \eta^2 m (1 + \frac{n - |B|}{|B|(n - 1)}) - 2 m \eta \delta \right) \mathbb{E} [F(\bar{w}^r) - F(w^m)]
\]

\[
+ 16\delta \rho_s^m \eta^2 m \frac{n - |B|}{|B|(n - 1)} \mathbb{E} [F(\bar{w}^r) - F(w^m)] + \frac{4\eta^2 m\alpha}{k} \|\nabla F(w^m)\|^2
\]

Since \( \bar{w}^r = H_k(w^m) \), i.e., \( \bar{w}^r \) is the best \( s \)-sparse approximation of \( w^m \), then we have the following result due to Lemma 2

\[
\|\bar{w}^r - w^*\|^2 \leq \left( 1 + \frac{2\sqrt{s^r}}{\sqrt{s} - s^*} \right) \|w^m - w^*\|^2.
\]

Let \( \alpha = 1 + \frac{2\sqrt{s^r}}{\sqrt{s} - s^*} \), and by combining (29) and (30), we have

\[
(1 + \frac{1}{k} (16\delta \rho_s^m \eta^2 m (1 + \frac{n - |B|}{|B|(n - 1)}) - 2 m \eta \delta \frac{\rho_s^m (d - s)}{2\sigma (d - s^*)}) \|\bar{w}^r - w^*\|^2
\]

\[
\leq \alpha \mathbb{E} \|\bar{w}^r - w^m\|^2 + \alpha \left( 16\delta \rho_s^m \eta^2 m (1 + \frac{n - |B|}{|B|(n - 1)}) - 2 m \eta \delta \right) \mathbb{E} [F(\bar{w}^r) - F(w^m)]
\]

\[
+ 16\delta \rho_s^m \eta^2 m \frac{n - |B|}{|B|(n - 1)} \mathbb{E} [F(\bar{w}^r) - F(w^m)] + \frac{4\eta^2 m\alpha}{k} \|\nabla F(w^m)\|^2
\]

\[
\leq \frac{2\alpha}{\rho_s^m} \mathbb{E} [F(\bar{w}^r) - F(w^m)] + \alpha \left( 16\delta \rho_s^m \eta^2 m (1 + \frac{n - |B|}{|B|(n - 1)}) - 2 m \eta \delta \right) \mathbb{E} [F(\bar{w}^r) - F(w^m)]
\]

\[
+ 16\delta \rho_s^m \eta^2 m \frac{n - |B|}{|B|(n - 1)} \mathbb{E} [F(\bar{w}^r) - F(w^m)] + \frac{4\eta^2 m\alpha}{k} \|\nabla F(w^m)\|^2
\]

where the last inequality follows from the RSC condition and the definition of \( \hat{I} \), i.e., \( \hat{I} = \text{supp}(HT(\nabla F(w^*); 2s)) \cup \text{supp}(w^*) \).
Through proper simplification, we have the following result:

\[
\frac{\alpha}{k}(2m\eta\delta - 16\delta \rho^+ s^2 m (1 + \frac{n - |B|}{|B|(n-1)})\mathbb{E}[F(\bar{w}^r) - F(w^*)] \\
\leq \left(\frac{2\alpha}{\rho^s} + 16\delta \rho^+ s^2 m \frac{n - |B|}{|B|(n-1)}\right)\mathbb{E}[F(\bar{w}^{r-1}) - F(w^*)] + \frac{4\eta^2 m \alpha}{k} \|\nabla_2 F(w^*)\|^2
\]

(32)

where the first term of the LHS of (31) is omitted due to the fact that the coefficient of \(\mathbb{E}\|\bar{w}^r - w^*\|^2\) is larger than zero, when \(8\rho^+ s^2 m \eta^2 (1 + \frac{n - |B|}{|B|(n-1)}) - m \eta + \frac{\sigma_k (\delta - s)}{s \rho^+ (d - s)} > 0\), which can be naturally satisfied under our parameter setting with \(s \geq \kappa \delta s^*\). Following from (32), we have

\[
\mathbb{E}[F(\bar{w}^r) - F(w^*)] \leq \beta \mathbb{E}[F(\bar{w}^{r-1}) - F(w^*)] + \frac{2\eta}{\delta (1 - \omega)} \|\nabla_2 F(w^*)\|^2
\]

(33)

where \(\beta = \frac{k}{\eta m \rho^2 \delta (1 - \omega)} + \frac{8\rho^+ s^2 \eta (n - |B|)}{|B|(1 - \omega)(n-1)}\) and \(\omega = 8\rho^+ s^2 \eta (1 + \frac{n - |B|}{|B|(n-1)})\). By applying (33) recursively, then we have the following desired bound with \(\beta \leq \frac{1}{2} < 1\):

\[
\mathbb{E}[F(\bar{w}^r) - F(w^*)] \leq \left(\frac{1}{2}\right)^r \mathbb{E}[F(\bar{w}^0) - F(w^*)] + \frac{\eta}{1 - \omega} \|\nabla_2 F(w^*)\|^2.
\]

(34)

This completes the proof. \(\square\)

C.2 PROOF OF COROLLARY 1

Proof. We then give the proof for the statistical estimation analysis. Following from the RSC condition of \(F(\cdot)\), we have

\[
F(w^*) \leq F(\bar{w}^r) + \langle \nabla F(w^*), w^* - \bar{w}^r \rangle - \frac{\rho^s}{2} \|\bar{w}^r - w^*\|^2.
\]

(35)

Let \(\phi = (\frac{1}{2})^r \mathbb{E}[F(\bar{w}^0) - F(w^*)] + \frac{\eta}{\delta (1 - \omega)} \|\nabla_2 F(w^*)\|^2\). Combining (34) and (35), we have

\[
\mathbb{E}[F(\bar{w}^r) - \phi] \leq F(w^*) \leq F(\bar{w}^r) + \langle \nabla F(w^*), w^* - \bar{w}^r \rangle - \frac{\rho^s}{2} \|\bar{w}^r - w^*\|^2.
\]

(36)

Using the duality of norms, we have

\[
\mathbb{E}\langle \nabla F(w^*), w^* - \bar{w}^r \rangle \leq \|\nabla F(w^*)\|_{\infty} \mathbb{E}\|\bar{w}^r - w^*\|_1 \leq \sqrt{s} \|\nabla F(w^*)\|_{\infty} \mathbb{E}\|\bar{w}^r - w^*\|.
\]

(37)

Combining (36), (37) and \(\mathbb{E}|w|^2 \leq \mathbb{E}\|w\|^2\), we have

\[
\frac{\rho^s}{2} \mathbb{E}\|\bar{w}^r - w^*\|^2 \leq \frac{\sqrt{s} \|\nabla F(w^*)\|_{\infty}}{\rho^s} \mathbb{E}\|\bar{w}^r - w^*\| + \phi.
\]

(38)

Let \(a = \mathbb{E}\|\bar{w}^r - w^*\|\). From the above inequality, we solve the following quadratic function with respect to \(a\),

\[
\frac{\rho^s}{2} a^2 - \sqrt{s} \|\nabla F(w^*)\|_{\infty} a - \phi \leq 0,
\]

(39)

which yields the bound

\[
\mathbb{E}\|\bar{w}^r - w^*\| \leq \sqrt{\frac{2(\frac{1}{2})^r |F(\bar{w}^0) - F(w^*)|}{\rho^s}} + \sqrt{\frac{2\eta}{1 - \omega} \|\nabla_2 F(w^*)\|} + \frac{2\sqrt{s} \|\nabla F(w^*)\|_{\infty}}{\rho^s}
\]

(40)

\[
\leq \frac{2(\frac{1}{2})^r |F(\bar{w}^0) - F(w^*)|}{\rho^s} + \left(\frac{2}{\rho^s} + \frac{2\eta}{1 - \omega}\right) \sqrt{s} \|\nabla F(w^*)\|_{\infty}.
\]

This completes the proof. \(\square\)
C.3 Bound $\beta < 1$

Now we show that we can guarantee that $\beta \leq 1$ by providing the appropriate choices of constants with $\eta$, $k$ and $m$ used in the theorem. More specifically, let $\eta \leq \frac{C_2}{\rho_s}$, and $|B| = 1$, then we have

$$\frac{8\rho_s^2 \eta (n - |B|)}{|B|(1 - \omega)(n - 1)} \leq \frac{8C_3(n - |B|)}{|B|(n - 1)(1 - 8C_3(1 + \frac{n - |B|}{|B|(n - 1)}))} \leq \frac{1}{4}.$$  \hspace{1cm} (41)

If $\eta \geq \frac{C_2}{\rho_s}$ with $C_2 \leq C_3$, then we have

$$\frac{k}{\eta \delta m \rho_s (1 - \omega)} \leq \frac{1}{2mC_2}.$$  \hspace{1cm} (42)

If we guarantee $\frac{3}{2m\delta \rho_s} \leq \frac{1}{4}$, then we have

$$m \geq \frac{6k\kappa_\delta}{\delta C_2} = C_4\kappa_\delta,$$  \hspace{1cm} (43)

where $\delta > 1$. If we choose $C_2 = \frac{1}{60}$, $C_3 = \frac{1}{50}$, $C_4 = 1800$, $k = 10$ and $\delta = 2$, then we have $\beta \leq \frac{1}{7}$.

Note that the convergence rate $\beta$ does not include $\alpha$, therefore, the required value of $k$ comes from the restricted constraint of the coefficient of $||\bar{w}^* - w^*||^2$, which implies that $s = \Omega(\kappa_\delta s^*).$

C.4 Proof of Corollary 2

Proof. \hspace{1cm} $\mathbb{E}[F(\bar{w}^*) - F(w^*)] \leq \epsilon + \frac{\eta}{1 - \omega} \|\nabla_2 F(w^*)\|^2.$  \hspace{1cm} (44)

Let $\varrho_1, \varrho_2, \cdots$ be a non-negative sequence of random variables which is defined as

$$\varrho_r = \max \left\{ F(\bar{w}^*) - F(w^*) - \frac{\eta}{1 - \omega} \|\nabla_2 F(w^*)\|^2, 0 \right\}.$$  \hspace{1cm} (45)

For a fixed $\epsilon > 0$, it follows from (34) and the Markov inequality,

$$\mathbb{P}(\varrho_r \geq \epsilon) \leq \frac{\mathbb{E}\varrho_r}{\epsilon} \leq \left(\frac{1}{2}\right)^r \frac{F(\bar{w}^*) - F(w^*)}{\epsilon}.$$  \hspace{1cm} (46)

For a given $\zeta \in (0, 1)$, let the right side of (46) be not greater than $\zeta$, which requires

$$r \geq \log_2 \frac{F(\bar{w}^*) - F(w^*)}{\epsilon \zeta}.$$  \hspace{1cm} (47)

Therefore, the result in (44) holds with probability at least $1 - \zeta$. Thus, we need $O(\log(\frac{1}{\epsilon}))$ outer iterations to get $\bar{w}^*$ satisfying (44). Since within each outer-iteration, we need to calculate a full gradient and $m$ stochastic variance reduced gradients with mini-batch size $|B|$, the overall oracle complexity is $O((n + \frac{\kappa_\delta |B|}{k}) \log \frac{1}{\epsilon})$. Note that in high-dimensional region, the effect of support of snapshot point can be ignored. On the other hand, we only perform a hard thresholding operation at each outer iteration. Then we have

$$\log_2 \left(\frac{1}{\epsilon} \right) = \frac{\log(\frac{1}{\epsilon})}{\log(2)} = O(\log(\frac{1}{\epsilon})).$$  \hspace{1cm} (48)

Thus, the hard thresholding oracle complexity is $O(\log(\frac{1}{\epsilon})).$ This completes the proof. \hspace{1cm} $\square$

D Proof of Theorem of Sparse Variant

In this section, we theoretically analyze the convergence properties of our Serial Sparse SBCD-HTP ($S^2$BCD-HTP).
Algorithm 3 Serial Sparse SBCD-HTP (S²BCD-HTP)

Input: number of outer-loops $R$, number of inner-loops $m$, step size $\eta$, sparsity level $s$;
Initialize: $\bar{w}^0$;
1: for $r = 0, 1, \ldots, R - 1$ do
2: $w^0 = \bar{w} = \bar{w}^r$;
3: $\nabla F(\bar{w}) = \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(\bar{w})$;
4: $\tilde{G} = \text{supp}(\bar{w})$;
5: for $t = 0, 1, \ldots, m - 1$ do
6: Randomly sample $i_t$ from $\{1, 2, \ldots, n\}$ uniformly;
7: Randomly sample $j_t$ from $\{1, 2, \ldots, k\}$ uniformly;
8: $T_{i_t} := \text{support of sample } i_t$;
9: $S = \tilde{G} \cup G_{j_t}$;
10: $\nabla g_S([w^t]_{T_{i_t}}) = \nabla f_{i_t}([w^t]_{T_{i_t}}) - \nabla f_{i_t}([\bar{w}]_{T_{i_t}}) + D_{i_t} \nabla g(\bar{w})$;
11: $w^{t+1} = w^t - \eta \nabla g_S([w^t]_{T_{i_t}})$;
12: end for
13: $\bar{w}^{r+1} = HT(w^m, s)$;
14: end for
Output: $\bar{w}^R$.

D.1 Proof

Proof. We start with the iterate difference between $w^{t+1}$ and $w^*$. By expanding iterate difference and taking expectation with respect to all randomness, we get

$$
E[\|w^{t+1} - w^*\|^2] = E[\|w^t - \eta \nabla g_S w^t - w^*\|^2]
= \|w^t - w^*\|^2 + \eta^2 \|\nabla g_S w^t\|^2 + 2\eta E(\nabla g_S w^t, w^* - w^t)
\leq \|w^t - w^*\|^2 + \eta^2 \|\nabla g_S w^t\|^2 + \frac{2\eta}{k} \|\nabla F(w^t), w^* - w^t\|,
$$

where the last inequality holds by the unbiasedness of the sparse gradient estimator $\nabla g_S(w^t)$ and the expectation with respect to randomized block. By using the $\rho_s^+$-strongly convex condition, we get the bound for $\langle \nabla F(w^t), w^* - w^t \rangle$ as follows:

$$
\langle \nabla F(w^t), w^* - w^t \rangle \leq F(w^*) - F(w^t) - \frac{\rho_s^+}{2} \|w^* - w^t\|^2.
$$

Due to the support set $S$, we need the extended RSC condition with $\delta > 1$, which has been introduced in the proof of Lemma 5. Based on this underlying consensus, we can obtain the bound for $\langle \nabla F_S(w^t), w^* - w^t \rangle$.

$$
\langle \nabla F_S(w^t), w^* - w^t \rangle \leq \delta [F(w^*) - F(w^t)].
$$

Using Lemma we have the following sparse variance gradient bound,

$$
E[\|\nabla g_S w^t\|^2] \leq \frac{1}{k} (12\delta \rho_s^+ [F(w^t) - F(w^*)] + (3 + 12(D_m^2 - 1)) \|\nabla F(w^*)\|^2
+ (24\delta \rho_s^+ + 48\rho_s^+(D_m^2 - 1)) [F(\bar{w}) - F(w^*)]),
$$

By combining the above inequalities, we have

$$
E[\|w^{t+1} - w^*\|^2] \leq \|w^t - w^*\|^2 + \eta^2 \|\nabla g_S w^t\|^2 - \frac{2\eta}{k} [F(w^t) - F(w^*)]
\leq \|w^t - w^*\|^2 + \frac{12\eta^2 \rho_s^+ - 2\eta}{k} [F(w^t) - F(w^*)]
+ \frac{2\delta \eta^2}{k} (12\rho_s^+ + 24\rho_s^+ (D_m^2 - 1)) [F(\bar{w}) - F(w^*)]
+ \frac{\eta^2}{k} (3 + 12(D_m^2 - 1)) \|\nabla F(w^*)\|^2,
$$
where the second inequality holds by Lemma 5. Summing the above inequality over \( t = 0, \ldots, m - 1 \) and taking expectation with respect to all randomness in this epoch, we have

\[
    \operatorname{E}[\|w^m - w^*\|^2] \leq \operatorname{E}[\|\tilde{w}^{r-1} - w^*\|^2] - \frac{2\eta \delta m(1 - 6\eta \rho^+_s)}{k} [F(w^m) - F(w^*)] \\
    + \frac{2\eta^2 m}{k} (12\rho^+_s + 24\rho^+_s (D_m^2 - 1)) [F(\tilde{w}) - F(w^*)] \\
    + \frac{\eta^2 m}{k} (3 + 12(D_m^2 - 1)) \|\nabla F(w^*)\|^2.
\]

Using (28), we obtain

\[
    \operatorname{E}[\|w^m - w^*\|^2] \leq \operatorname{E}[\|\tilde{w}^{r-1} - w^*\|^2] - \frac{2\eta \delta m(1 - 6\eta \rho^+_s)}{k} [F(\tilde{w}^r) - F(w^*)] \\
    + \frac{2\eta^2 m}{k} (12\rho^+_s + 24\rho^+_s (D_m^2 - 1)) [F(\tilde{w}^{r-1}) - F(w^*)] \\
    + \frac{\eta^2 m}{k} (3 + 12(D_m^2 - 1)) \|\nabla F(w^*)\|^2. \tag{54}
\]

Since \( \tilde{w}^r = HT(w^m, s) \), i.e., \( \tilde{w}^r \) is the best \( s \)-sparse approximation of \( w^m \), and due to Lemma 2, then we have following result,

\[
    \frac{2\eta \delta m}{k} (1 - 6\eta \rho^+_s) [F(\tilde{w}^r) - F(w^*)] + (1 - 2\eta \delta m(1 - 6\eta \rho^+_s) \frac{\rho^+_s (d - s)}{2k\sigma(d - s^*)}) \|\tilde{w}^r - w^*\|^2 \\
    \leq \alpha \operatorname{E}[\|\tilde{w}^{r-1} - w^*\|^2] + \frac{\delta \eta^2 m \alpha}{k} (24\rho^+_s + 48\rho^+_s (D_m^2 - 1)) [F(\tilde{w}^{r-1}) - F(w^*)] \\
    + \frac{\eta^2 m \alpha}{k} (3 + 12(D_m^2 - 1)) \|\nabla F(w^*)\|^2 \tag{55}
\]

where \( \alpha = 1 + \frac{2\sqrt{\rho_s}}{\sqrt{s} - s^*} \), and we use \( \rho^+_s \)-strongly convex of \( F(w) \) to bound \( \|\tilde{w}^{r-1} - w^*\|^2 \) in the last inequality. Through proper simplification, we can obtain

\[
    \frac{2\eta \delta m}{k} (1 - 6\eta \rho^+_s) [F(\tilde{w}^r) - F(w^*)] \\
    \leq \frac{2\alpha}{\rho^+_s} [F(\tilde{w}^{r-1}) - F(w^*)] + \frac{\delta \eta^2 m \alpha}{k} (24\rho^+_s + 48\rho^+_s (D_m^2 - 1)) [F(\tilde{w}^{r-1}) - F(w^*)] \\
    + \frac{\eta^2 m \alpha}{k} (3 + 12(D_m^2 - 1)) \|\nabla F(w^*)\|^2, \tag{56}
\]

where the second term of LHS of (56) is omitted due to the fact that the coefficient of \( \operatorname{E}[\|\tilde{w}^r - w^*\|^2] \) is larger than zero, when \( 6\rho^+_s m \eta^2 - m \eta + \frac{k\sigma(d - s^*)}{\rho^+_s (d - s)} > 0 \), which can be naturally satisfied under our parameter setting with \( s \geq \kappa s^* \). Therefore, we can obtain the following result,

\[
    \operatorname{E}[F(\tilde{w}^r) - F(w^*)] \leq \frac{2\alpha}{\rho^+_s} + (24 + 48(D_m^2 - 1)) \delta m \eta \rho^+_s \eta^2 \frac{\|\tilde{w}^{r-1} - w^*\|^2}{2\eta \delta m (1 - 6\eta \rho^+_s)} \|\nabla F(w^*)\|^2, \tag{58}
\]
where \( \hat{I} = \text{supp}(w^*) \cup \text{supp}(HT(\nabla F(w^*), 2s)) \). We define \( \xi := \frac{[3+12(D^2_m - 1)]\eta}{2m(1-6\rho_s^2)\eta} \|\nabla F(w^*)\|^2 \).

In order to obtain a linear convergence rate, we have to bound

\[
\beta := \frac{2\kappa \eta}{\rho_s^2} + \frac{(24 + 48(D^2_m - 1))\delta m \alpha \rho_s^+ \eta^2}{(\eta - 6\rho_s^2 \eta^2)\delta m \rho_s^2} + \frac{(12 + 24(D^2_m - 1))\rho_s^+ \eta}{1 - 6\rho_s^2 \eta} < 1.
\]

If we choose \( \eta \leq \frac{1}{48D_m^3\rho_s^2} \), then the second term of \( \beta \) will be less than \( \frac{1}{2} \), i.e., \( \frac{(12 + 24(D^2_m - 1))\rho_s^+ \eta}{1 - 6\rho_s^2 \eta} \leq \frac{1}{2} \).

As for the first term, when \( m \geq \frac{50D_m^3\kappa_2}{\rho_s^2} \) with \( \eta \geq \frac{1}{50D_m^3\rho_s^2} \), \( \delta = 4 \) and \( k = 10 \), we have \( \frac{1}{(\eta - 6\rho_s^2 \eta^2)\delta m \rho_s^2} \leq \frac{1}{4} \). Thus, we get the linear convergence for our serial sparse SBCD-HTP (S^2BCD-HTP).

\[
\mathbb{E}[F(\hat{w}^r) - F(w^*)] \leq \frac{3}{4} \cdot [F(\hat{w}^{r-1}) - F(w^*)] + \xi,
\]

which means that the total gradient evaluation oracle complexity is \( \mathcal{O}(n + \frac{\kappa_2}{\epsilon} \log \frac{1}{\epsilon}) \), and the hard thresholding oracle complexity is \( \mathcal{O}(\log(\frac{1}{\epsilon})) \). This completes the proof. \( \square \)

### E PROOF OF THEOREM 2

In this section, we analyze the convergence properties of the proposed Asynchronous Sparse SBCD-HTP (ASBCD-HTP). We first have to specify the iterates labeling order used in our asynchronous analysis.

We use “After Read” labeling order (Leblond et al., 2017) in our proof, which enjoys a simpler analysis but requires the order of randomly sampling step to be an uniform distributed sample. In order to give a clear proof, we adopt the “After Read” labeling order and make the following assumptions:

**Assumption 4** The labeling order increases after Step 8 in Algorithm 2 finished, and thus the future perturbation is not considered in the effect of asynchrony in the current step.

**Assumption 5** We assume that uniform distributed samples and the independence of the sample \( i_{t+1} \) with \( \hat{w}_t \).

Following (Leblond et al., 2017), we can explicit the effect of asynchrony as follows:

\[
\hat{w}_t - w^t = \eta \sum_{u=(t-\tau)_+}^{(t-1)} G_u \nabla g(\hat{w}_u),
\]

where \( w^t \) is the \( t \)-th inner iteration solution. In the proof of Theorem 2, we use \( w^t \) to represent the temporary solution of the \( t \)-th iteration for a clean representation. And \( G_u \) is a diagonal matrix with entries in \{0, +1\}. This explicitly defines the coordinate perturbation from the past updates. Here, \( \tau \) denotes the maximum number of overlaps between concurrent threads. We also denote \( \Delta = \max_{j=1, \ldots, d} p_j \), which provides a measure of sparsity following (Leblond et al., 2017).

### E.1 PROOF

**Proof.** We analyze our asynchronous algorithm based on the “perturbed iterate analysis” framework (Mania et al., 2017) with “After Read” labeling order. By expanding the iterate difference and taking expectation with respect to the sample \( i_{t+1} \), we get
\[ E_{t+1} \| w^{t+1} - w^* \|^2 = E_{t+1} \| w^t - \eta \nabla_S g_{\mathcal{I}}(\hat{w}_t) - w^* \|^2 \]
\[ = \| w^t - w^* \|^2 + \eta^2 \| \nabla_S g_{\mathcal{I}}(\hat{w}_t) \|^2 + 2\eta E_{t+1} \langle w^* - w^t, \nabla g_{\mathcal{I}}(\hat{w}_t) \rangle \]
\[ \leq \| w^t - w^* \|^2 + \eta^2 \| \nabla_S g_{\mathcal{I}}(\hat{w}_t) \|^2 + \frac{2\eta}{k} E_{t+1} \langle w^* - \hat{w}_t, \nabla g_{\mathcal{I}}(\hat{w}_t) \rangle \]
\[ + \frac{2\eta}{k} E_{t+1} \langle \hat{w}_t - w^t, \nabla_S g_{\mathcal{I}}(\hat{w}_t) \rangle, \]

where \( \hat{w}_t \) is the “perturbed” iterate with perturbation \( \psi \) and the last inequality follows from the expectation with respect to randomized block. We need to bound the term, \( 2\eta \langle w^* - \hat{w}_t, \nabla g_{\mathcal{I}}(\hat{w}_t) \rangle \).

Since \( i_{t+1} \) is independent to \( \hat{w}_t \), we can write:
\[ E_{t+1} \langle w^* - \hat{w}_t, \nabla g_{\mathcal{I}}(\hat{w}_t) \rangle = \langle w^* - \hat{w}_t, E_{i_{t+1}} \nabla g_{\mathcal{I}}(\hat{w}_t) \rangle = \langle w^* - \hat{w}_t, \nabla_{\mathcal{I}} F(\hat{w}_t) \rangle. \]

We can use \( \rho^- \)-strong convexity bound of \( F(\hat{w}) \) with constant factor \( \delta > 1 \) as well as a squared triangle inequality to get:
\[ -\langle \hat{w}_t - w^*, \nabla F(\hat{w}_t) \rangle \leq -\delta (F(\hat{w}_t) - F(w^*)) - \frac{\rho^-}{2} \| \hat{w}_t - w^* \|^2, \]
\[ - \| \hat{w}_t - w^* \|^2 \leq \| \hat{w}_t - w^t \|^2 - \frac{1}{2} \| w^t - w^* \|^2, \quad (\| a + b \|^2 \leq 2\| a \|^2 + 2\| b \|^2), \]
\[ \langle w^* - \hat{w}_1, \nabla_{\mathcal{I}} F(\hat{w}_1) \rangle \leq -\delta (F(\hat{w}_t) - F(w^*)) + \frac{\rho^-}{2} \| \hat{w}_t - w^t \|^2 - \frac{\rho^-}{4} \| w^t - w^* \|^2. \]

Then we make an identical deformation, and we have
\[ \frac{2\eta}{k} E_{t+1} \langle w^* - \hat{w}_t, \nabla g_{\mathcal{I}}(\hat{w}_t) \rangle \leq -\frac{\eta \rho^-}{2k} E_{t+1} \| w^t - w^* \|^2 + \frac{\eta \rho^-}{k} E_{t+1} \| \hat{w}_t - w^t \|^2 \]
\[ - \frac{2\eta \delta}{k} (F(\hat{w}_t) - F(w^*)). \]

Putting it all together, we get the initial recursive inequity, written here explicitly:
\[ a^{t+1} \leq (1 - \frac{\eta \rho^-}{2k}) a^t + \eta^2 E_{t+1} \| \nabla_S g_{\mathcal{I}}(\hat{w}_t) \|^2 + \frac{\eta \rho^-}{k} E_{t+1} \| \hat{w}_t - w^t \|^2 \]
\[ + \frac{2\eta}{k} E_{t+1} \langle \hat{w}_t - w^t, \nabla g_{\mathcal{I}}(\hat{w}_t) \rangle - \frac{2\eta \delta}{k} \hat{e}^t, \]

where \( a^t := E_{i_{t+1}} \| w^t - w^* \|^2 \) and \( \hat{e}^t := E_{i_{t+1}} [F(\hat{w}_t) - F(w^*)] \).

From Lemma 1 in [Leblond et al. 2017], we can bound the asynchronous terms \( \| \hat{w}_t - w^t \|^2, \langle \hat{w}_t - w^t \rangle \)

and \( \nabla g_{\mathcal{I}}(\hat{w}_t) \) as follows.
\[ E \langle \hat{w}_t - w^t, \nabla g_{\mathcal{I}}(\hat{w}_t) \rangle = \eta \sum_{u=(t-r)}^{t-1} E \langle G_u \nabla_S g_{\mathcal{I}}(\hat{w}_u), \nabla g_{\mathcal{I}}(\hat{w}_u) \rangle \]
\[ \leq \eta \sum_{u=(t-r)}^{t-1} E \| \nabla_S g_{\mathcal{I}}(\hat{w}_u), \nabla g_{\mathcal{I}}(\hat{w}_u) \| \]
\[ \leq \eta \sum_{u=(t-r)}^{t-1} \| \nabla g_{\mathcal{I}}(\hat{w}_u) \| + E \| \nabla g_{\mathcal{I}}(\hat{w}_u) \|^2 \]
\[ \leq \eta \| \nabla g_{\mathcal{I}}(\hat{w}_u) \|^2 \sum_{u=(t-r)}^{t-1} E \| \nabla g_{\mathcal{I}}(\hat{w}_u) \|^2 + \frac{\eta \sqrt{\Delta}}{2} E \| \nabla g_{\mathcal{I}}(\hat{w}_u) \|^2. \]
\[
\mathbb{E}\|\tilde{w}_t - w^t\|^2 \leq \sum_{u=(t-\tau)_+}^{t-1} \left\langle (G_u \nabla g(w_u), G_v \nabla g(w_v)) \right\rangle \\
\leq \eta^2 \sum_{u=(t-\tau)_+}^{t-1} \|\nabla g(w_u)\|^2 + \eta^2 \sum_{u=(t-\tau)_+, u \neq v}^{t-1} \left\langle (G_u \nabla g(w_u), G_v \nabla g(w_v)) \right\rangle \\
\leq \eta^2 \sum_{u=(t-\tau)_+}^{t-1} \mathbb{E}\|\nabla g(w_u)\|^2 + \eta^2 \sqrt{\Delta} (t-1) \sum_{u=(t-\tau)_+}^{t-1} \mathbb{E}\|\nabla g(w_u)\|^2 \\
= \eta^2 (1 + \sqrt{\Delta}) \sum_{u=(t-\tau)_+}^{t-1} \mathbb{E}\|\nabla g(w_u)\|^2 \\
\leq \eta^2 (1 + \sqrt{\Delta}) \tau \sum_{u=(t-\tau)_+}^{t-1} \mathbb{E}\|\nabla g(w_u)\|^2,
\]
(70)

where \(G_u\) are \(d \times d\) diagonal matrices whose entries in \(\{0, +1\}\). Each update in \(\tilde{w}_t\) is already in \(w^t\) – this is the case of 0. On the contrary, some updates might be late: this is the case of +1. Then by combining (68), (69) and (70), we get

\[
a^{t+1} \leq (1 - \frac{\eta \rho_s^2}{2k})a^t + \eta^2 \|\nabla g(w_t)\|^2 + \eta^3 \rho_s^2 (1 + \sqrt{\Delta} \tau) \sum_{u=(t-\tau)_+}^{t-1} \mathbb{E}_{i_{t+1}} \|\nabla g(w_u)\|^2 - 2\eta \delta \tau e^t \\
+ \eta^2 \sqrt{\Delta} \sum_{u=(t-\tau)_+}^{t-1} \mathbb{E}_{i_{t+1}} \|\nabla g(w_u)\|^2 + \eta^2 \sqrt{\Delta}\tau \mathbb{E}_{i_{t+1}} \|\nabla g(w_t)\|^2 - 2\eta \delta \tau e^t \\
\leq (1 - \frac{\eta \rho_s^2}{2k})a^t + \eta^2 C_1 \sum_{u=(t-\tau)_+}^{t-1} \mathbb{E}_{i_{t+1}} \|\nabla g(w_u)\|^2 + \eta^2 C_2 \sum_{u=(t-\tau)_+}^{t-1} \mathbb{E}_{i_{t+1}} \|\nabla g(w_t)\|^2 - 2\eta \delta \tau e^t,
\]
(71)

where \(C_1 := 1 + \sqrt{\Delta} + \eta \rho_s^2 C_1\). \(\tau\) represents the maximum number of overlaps between concurrent threads. And \(\Delta = \max_{j=1,\ldots,d} p_j\), which is a measure of sparsity with \(1/n \leq \Delta \leq 1\).

Using Lemma 6 to bound the sparse variance term, the above inequality becomes

\[
a^{t+1} \leq (1 - \frac{\eta \rho_s^2}{2k})a^t + \frac{\eta^2 C_1}{k} (12 \delta \rho_s^2 [F(\tilde{w}_t) - F(w^*)] + (24 + 48(D_m^2 - 1)) \delta \rho_s^2 [F(\tilde{w}) - F(w^*)] \\
+ (3 + 12(D_m^2 - 1)) \|\nabla F(w^*)\|^2) + \frac{\eta^2 C_2}{k} \sum_{u=(t-\tau)_+}^{t-1} \mathbb{E}_{i_{t+1}} \|\nabla g(w_u)\|^2 - \frac{2\eta \delta}{k} e^t \\
\leq a^t + \frac{\eta^2 C_2}{k} \sum_{u=(t-\tau)_+}^{t-1} \mathbb{E}_{i_{t+1}} \|\nabla g(w_u)\|^2 - \frac{2\eta \delta (1 - 6 \rho_s^2 \eta C_1)}{k} e^t \\
+ (24 \rho_s^2 + 48 \rho_s^2 (D_m^2 - 1)) \frac{\delta C_1 \eta^2}{k} e^t + (3 + 12(D_m^2 - 1)) \frac{\eta^2 C_1}{k} \|\nabla F(w^*)\|^2,
\]
(72)

where \(\tilde{e} := \mathbb{E}[F(\tilde{w}) - F(w^*)], \hat{e}^t := \mathbb{E}[F(\tilde{w}_t) - F(w^*)]\). Note that we have \(\tilde{w} = w^0 = \tilde{w}^{r-1}\).

Summing (72) over \(t = 0, \ldots, m-1\) and taking expectation with all randomness in this epoch, we get

\[
a^m \leq a^m - \frac{2\eta \delta m}{k} (1 - 6 \rho_s^2 \eta C_1) \hat{e}^m + \frac{\eta^2 C_2}{k} \sum_{u=(t-\tau)_+}^{t-1} \mathbb{E}[\|\nabla g(w_u)\|^2] \\
+ (24 \rho_s^2 + 48 \rho_s^2 (D_m^2 - 1)) \frac{\delta m C_1 \eta^2}{k} \hat{e}^{r-1} + (3 + 12(D_m^2 - 1)) \frac{\eta^2 C_1}{k} \|\nabla F(w^*)\|^2,
\]
(73)
where \( a^m = E \|w^m - w^*\|^2 \), and \( a^0 = E \|w^0 - w^*\|^2 \). Then we focus on upper bounding the third term on the RHS of (73).

\[
\sum_{t=1}^{m} \sum_{u=(t-\tau)_+}^{t-1} E[\|\nabla g_x(\hat{w}_u)\|^2] \leq \tau \sum_{t=1}^{m-1} E[\|\nabla g_x(\hat{w}_t)\|^2] \leq \tau \sum_{t=1}^{m} E[\|\nabla g_x(\hat{w}_t)\|^2]
\]

\[
\leq (12\rho_s^+ \delta m\bar{e}m + \delta m(24\rho_s^+ + 48\rho_s^+ (D^2_m - 1))\bar{e}^{r-1} + m(3 + 12(D^2_m - 1))\|\nabla_x F(w^*)\|^2).
\]

(74)

Substituting the above inequality into (73), we obtain

\[
a^m \leq a^0 - \frac{2\eta\delta m}{k} (1 - 6\rho_s^+ \eta(\tau C_2 + C_1)) \bar{e}^r + \frac{\alpha \rho_s^+ m \eta^2}{2k\sigma(d-s)} \bar{e}^{r-1}
\]

\[
+ \frac{\eta^2 m}{k} (3 + 12(D^2_m - 1))(C_1 + \tau C_2)\|\nabla_x F(w^*)\|^2.
\]

(75)

Then we bring (28) into the above inequality, we have

\[
a^m \leq a^0 - \frac{2\eta\delta m}{k} (1 - 6\rho_s^+ \eta(\tau C_2 + C_1)) \bar{e}^r + \frac{\alpha \rho_s^+ m \eta^2}{2k\sigma(d-s)} \bar{e}^{r-1}
\]

\[
+ \frac{\eta^2 m}{k} (3 + 12(D^2_m - 1))(C_1 + \tau C_2)\|\nabla_x F(w^*)\|^2,
\]

(76)

where the second inequality holds since we use (28) and make a proper simplification. Using Lemma 2 we have

\[
(1 - 2\eta\delta m(1 - 6\rho_s^+ \eta(\tau C_2 + C_1)) \frac{\rho_s^+(d-s)}{2k\sigma(d-s^*)} \bar{e}^r
\]

\[
\leq \frac{2\alpha}{\rho_s^+} \bar{e}^{r-1} - \frac{2\eta\delta m}{k} (1 - 6\rho_s^+ \eta(\tau C_2 + C_1)) \bar{e}^r
\]

\[
+ \frac{\alpha \rho_s^+ m \eta^2}{k} (24 + 48(D^2_m - 1))(\tau C_2 + C_1)\bar{e}^{r-1}
\]

\[
+ \frac{\eta^2 m}{k} (3 + 12(D^2_m - 1))(C_1 + \tau C_2)\|\nabla_x F(w^*)\|^2,
\]

(77)

where \( \alpha = 1 + \frac{2\sqrt{\pi}}{\sqrt{\pi}} \), \( \bar{e}^r = \|\bar{w}^r - w^*\|^2 \), and \( \hat{I} = \text{supp}(w^*) \cup \text{supp}(HT(\nabla F(w^*)), 2s) \). The first term of LHS of (77) is omitted due to the coefficient of \( \bar{e}^r \) is larger than zero, when \( 6\rho_s^+ m(\tau C_2 + C_1)\eta^2 - m_\eta + \frac{\kappa \sigma(d-s^*)}{\delta \rho_s^+(d-s)} > 0 \), which can be naturally satisfied under our parameter settings with \( s \geq \kappa s^* \). Therefore, we obtain the following recursive inequality,

\[
\bar{e}^r \leq \frac{2\alpha}{\rho_s^+} \bar{e}^{r-1} - \frac{2\eta\delta m(1 - 6\rho_s^+ \eta(\tau C_2 + C_1))}{k} \bar{e}^r
\]

\[
+ \frac{\alpha \rho_s^+ m \eta^2}{k} (24 + 48(D^2_m - 1))(\tau C_2 + C_1)\|\nabla_x F(w^*)\|^2.
\]

(78)
We assume that the following settings has satisfied with the above constraint and then by choosing \( m = 120 \kappa^\hat{s}, \eta = \frac{1}{60\rho^2}, k = 10 \), we get
\[
e \leq \frac{20 + \frac{\delta}{75(\rho^2)}(1 + 2(D_m^2 - 1))(\tau C_2 + C_1)}{4\delta(1 - \frac{\delta}{10}(\tau C_2 + C_1))} \cdot \tilde{e}^{r-1} + \xi \cdot \|\nabla_2 F(w^*)\|^2, \tag{79}
\]
where \( \xi = \eta^{(3+12(D_m^2 - 1))(\tau C_2 + C_1)} \). In order to ensure linear speed up, \( \tau \) needs to satisfy the following constraint,
\[
\theta \triangleq \frac{20 + \frac{\delta}{75(\rho^2)}(1 + 2(D_m^2 - 1))(\tau C_2 + C_1)}{4\delta(1 - \frac{\delta}{10}(\tau C_2 + C_1))} \leq 1. \tag{80}
\]
We assume that \( D_m^2 \) and \( \delta = 10 \), then by simply setting \( \tau \leq \min\{\frac{3}{5\sqrt{\Delta}}, 2\kappa^\hat{s}, \sqrt{\frac{2\kappa^\hat{s}}{\sqrt{\Delta}}}\} \), the above constraint is satisfied with \( \theta \leq 0.7538 \), which implies that the oracle gradient evaluation and hard thresholding oracle complexities are \( O((n + \frac{n}{\tau}) \log \frac{1}{\epsilon}) \) and \( O(\log(\frac{1}{\epsilon})) \), respectively. We can simply infer the statistical estimation result as before. Here, we will omit this part. This completes the proof. \( \square \)

E.2 Discussion about Sparse Variance Bound

In the dense update case (i.e., \( D_m = 1 \)), Lemma 6 is very similar to Lemma 5 when \( |B| = 1 \). In the sparse update case, Lemma 6 highly correlates with the sparsity of datasets (\( \propto D_m^2 \)), which could be much loose in some extreme cases. If \( D_m \) is very large (imaging a dataset with some dimensions contain only one entry among the \( n \) samples, so \( D_m = n \)), the step size of sparse (asynchronous) SBCD-HTP will be much smaller than that of SBCD-HTP. Our algorithms can still converge to the optimal solution up to the statistical error. Actually, it is still an open problem whether we can have a tighter variance bound in the sparse update setting that is uncorrelated with \( D_m \).

F More Experimental Results

Table 2: Summary of large-scale and high-dimensional datasets

<table>
<thead>
<tr>
<th>Datasets</th>
<th># Data</th>
<th># Feature</th>
<th>Density</th>
</tr>
</thead>
<tbody>
<tr>
<td>rcv1-train</td>
<td>20,242</td>
<td>47,236</td>
<td>0.16%</td>
</tr>
<tr>
<td>rcv1-test</td>
<td>677,399</td>
<td>47,236</td>
<td>0.15%</td>
</tr>
<tr>
<td>real-sim</td>
<td>72,309</td>
<td>20,958</td>
<td>0.024%</td>
</tr>
<tr>
<td>news20</td>
<td>19,996</td>
<td>30,000</td>
<td>0.49%</td>
</tr>
<tr>
<td>E2006-TFIDF</td>
<td>16,087</td>
<td>80,000</td>
<td>1.54%</td>
</tr>
</tbody>
</table>

F.1 Real-World Datasets and Equipments

We summarize the detailed information of all the real-world datasets in Table 2, including rcv1-train, rcv1-test, real-sim, news20 and E2006-TFIDF. All these datasets are provided in the LibSVM website.

https://www.csie.ntu.edu.tw/~cjlin/libsvm/
Instead of using all features, we randomly select 80,000 features for training. As for the news20 dataset, we randomly choose 30,000 features for training.

F.2 Baseline Methods

We compare our serial dense SBCD-HTP method with the state-of-the-art methods:

- **Fast Gradient with Hard Thresholding** (FG-HT) ([Yuan et al., 2014]): This method is based on a standard gradient descent step and combined with hard thresholding to solve sparsity-constrained problems.

- **Stochastic Gradient with Hard Thresholding** (SG-HT) ([Nguyen et al., 2017]): This method combines stochastic gradient descent with hard thresholding for solving large-scale sparsity-constrained problems.

- **Stochastic Variance Reduced Gradient with Hard Thresholding** (SVRG-HT) ([Li et al., 2016b]): This method incorporates the variance reduction technique in [Johnson & Zhang, 2013] into the hard thresholding algorithm to improve the efficiency of stochastic optimization.

- **Accelerated Stochastic Block Coordinate Gradient Descent with Hard Thresholding** (ASBCD-HT) ([Chen & Gu, 2016]): Block coordinate descent (BCD) is used for solving sparsity-constrained problems.

- **Fast Newton Hard Thresholding Pursuit** (FNHTP) ([Chen & Gu, 2017]): This method tries to iteratively approximate the inverse Hessian matrix and combines a Newton algorithm with hard thresholding.

For the proposed Serial Sparse SBCD-HTP ($S^2$BCD-HTP) and Asynchronous sparse parallel SBCD-HTP (ASBCD-HTP) methods, we compare them with the following asynchronous parallel sparsity-constrained methods:

- **Asynchronous Accelerated Stochastic Block Coordinate Descent Hard Thresholding** ($A^2$SBCD-HT): This algorithm is the asynchronous variant of ASBCD-HT by using our parallel framework.

- **Asynchronous Stochastic Gradient with Hard Thresholding** (ASG-HT): This method is an extension of SG-HT from serial dense setting to asynchronous sparse setting.

- **Asynchronous Stochastic Variance Reduced Gradient Hard Thresholding** (ASVRG-HT) ([Li et al., 2016a]): This method incorporates multicore structure to compute stochastic variance reduced gradient in a parallel way.

F.3 Face Recognition Datasets

Although there are many datasets available for face recognition, we choose a common dataset (i.e. the Extended Yale B database ([Georghiades et al., 2001]). The Extended Yale B database contains 2,414 frontal-face images of 38 people under different controlled lighting conditions ([Georghiades et al., 2001]). For each individual, we randomly choose 26 images for training and 15 images for testing.

F.4 Experimental Setup

For each sparsity-constrained algorithm, the sparse level makes great difference to the solution of sparse representation, especially at different noise levels. Therefore, in order to approach the best performance of all these algorithms, we change the sparsity parameter within a certain range for each algorithm. Thus, we can make sure that all these algorithms achieve the best recognition rates in the parameter setting. In all settings of the experiments, the images are down-sampled to $32 \times 32$ pixels.
Based on the sparse representation-based classification algorithm (SRC) (Wright et al., 2008), a series of processing operations are made to the above dataset. We first rescale the training matrix into $[0,1]$ for the convenience of adding noise, and then add Gaussian noise with zero mean and standard deviation $\vartheta$. Finally, we normalize the columns of the training matrix to have unit $\ell_2$-norm.

The parameter settings of comparison algorithms for face recognition task are the same as those described in Section 6.1.

F.5 RESULTS

We first measure the effect of mini-batch size $|B|$ in the convergence of SBCD-HTP on rcv1-train, as shown in Figure 6. It is obvious that the large mini-batch size can accelerate the convergence rate and make SBCD-HTP to obtain a better solution. This result is exactly in line with our theoretical analysis. However, when the mini-batch size is relatively large, it will deteriorate the performance of running time. One can simply choose the mini-batch size to 3 or 5 as suggested in Figure 6.
F.6 FACE RECOGNITION TASKS

Then, we compare the performance of some hard thresholding algorithms, including SG-HT, SVRG-HT and ASBCDHT, with our SBCD-HTP for solving face recognition tasks on the Extended Yale B dataset. As shown in Figure [5] SBCD-HTP can achieve the best performance among these stochastic hard thresholding methods in a short time. It is clear that SBCD-HTP can not only optimize the objective function efficiently but also obtain a better representative solution on real applications, which is consistent with our theoretical results as described in Section 4.2. This result also shows that our SBCD-HTP has a better generalization accuracy than other hard thresholding algorithms.

Figure 7: Comparison of all the algorithms for solving sparse logistic regression problems. In each plot, the vertical axis shows the objective value minus the minimum, and the horizontal axis is the number of effective passes over data or running time (seconds).

F.7 LINEAR/LOGISTIC REGISTRATION

We show more practical evaluation results for sparse linear regression and sparse logistic regression problems. All hardware environment and parameter settings are the same as in Section 6. For SBCD-HTP, from Figures [7] and [8] we can come to the same conclusion as we did before, that is in both sparse linear regression and sparse logistic regression, our algorithm converges faster than the baseline algorithms in terms of effective passes and running time. However, we find an interesting
result, which is that for the E2006-TFIDF dataset, all the algorithms converge with high precision solutions, so that we have to use some methods to magnify their differences. Of course, our SBCD-HTP has the best performance, but the gap is not that big. We think that this is due to ease with which E2006-TFIDF dataset can be optimized, since this phenomenon does not occur on other datasets.

For $S^2$BCD-HTP and ASBCD-HTP, from Figure 8, our $S^2$BCD-HTP and ASBCD-HTP far exceed the baseline algorithms in terms of running time. Therefore, we can confirmedly obtain the conclusion that our Sparse and Asynchronous algorithms outperform the state-of-the-art hard thresholding algorithms for sparse datasets. This is a significant improvement for hard thresholding algorithms to solve sparse learning problems in the high-dimensional asynchronous setting.

![Graph](image)

Figure 8: Comparison of all the algorithms for solving sparse linear regression problems.

### F.7.1 Upper Bound

In order to conduct an experiment on the upper bound of Theorem 1. We first generate some $n \times d$ synthetic matrices $X$, each row of which is drawn independently from a $d$-dimensional Gaussian distribution with mean 0 and covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$. The response vector is generated from the model $y = Xw^* + \epsilon$, where $w^* \in \mathbb{R}^d$ is the $s^*$-sparse coefficient vector, and we need to generate the noise $\epsilon$ drawn from a multivariate normal distribution $N(0, \sigma^2 I)$ with $\sigma^2 = 0.01$. The nonzero entries in $w^*$ are sampled independently from a uniform distribution over the interval $[-1, 1]$. For the experiments, we construct the following synthetic data set: $n = 5000$, $d = 10000$, $s^* = 500$, and the diagonal entries of the covariance matrix $\Sigma$ are set to 1, and the other entries are set to 0.1.
Figure 9: Comparison of ASG-HT, ASVRG-HT, $A^2$SBCD-HT, $S^2$BCD-HTP and ASBCD-HTP for solving sparse linear regression problems. The four asynchronous parallel algorithms (i.e., ASG-HT, ASVRG-HT, $A^2$SBCD-HT and ASBCD-HTP) run on 20 threads.

We set the step-size $\eta = \frac{1}{48\rho s}$ and batch size $|B| = 1$ during our experiments. The upper bound is then computed according to Theorem 1. Under this condition, we obtain the result, as is shown in Figure 10. As we can see that the upper bound converges faster at first and then it becomes slow since the statistical error becomes the main factor. The convergence performance of SBCD-HTP is always bounded below the black line, which demonstrates the correctness of our theorem.

F.8 LINEAR SVM

In this subsection, we consider the sparsity-constrained $L_2$-SVM problem (i.e.,
$$\min_w \frac{1}{2n} \sum_{i=1}^n \left( \max\{0, 1 - y^{(i)} w^\top x^{(i)}\} \right)^2 + \frac{\lambda}{2} \|w\|^2, \text{ subject to } \|w\|_0 \leq s,$$
where $y^{(i)}$ is the known label and $x^{(i)}$ is the given sample. We fix the regularization parameter $\lambda = 10^{-4}$. The other parameters are also same to the above experiments. The objective value gap and running time of the baselines are provided in Figure 11. We can see that we also make a similar observation as from the previous results. SBCD-HT performs the best among these considered baselines.
Figure 10: The upper bound of Theorem 1 vs the convergence performance.

Figure 11: Comparison of FG-HT, SG-HT, SVRH-HT, ASBCDHT and SBCD-HTP for solving sparse SVM problems.