Journal of Machine Learning Research 18 (2018) 1-42 Submitted 6/16; Revised 5/17; Published 4/18

A Tight Bound of Hard Thresholding

Jie Shen JS2007@RUTGERS.EDU
Department of Computer Science
Rutgers University
Piscataway, NJ 08854, USA

Ping Li PINGLI98@GMAIL.COM
Baidu Research
Bellevue, WA 98004, USA

Editor: Sujay Sanghavi

Abstract

This paper is concerned with the hard thresholding operator which sets all but the k largest absolute elements of a vector to zero. We establish a *tight* bound to quantitatively characterize the deviation of the thresholded solution from a given signal. Our theoretical result is universal in the sense that it holds for all choices of parameters, and the underlying analysis depends only on fundamental arguments in mathematical optimization. We discuss the implications for two domains:

Compressed Sensing. On account of the crucial estimate, we bridge the connection between the restricted isometry property (RIP) and the sparsity parameter for a vast volume of hard thresholding based algorithms, which renders an improvement on the RIP condition especially when the true sparsity is unknown. This suggests that in essence, many more kinds of sensing matrices or fewer measurements are admissible for the data acquisition procedure.

Machine Learning. In terms of large-scale machine learning, a significant yet challenging problem is learning accurate sparse models in an efficient manner. In stark contrast to prior work that attempted the ℓ_1-relaxation for promoting sparsity, we present a novel stochastic algorithm which performs hard thresholding in each iteration, hence ensuring such parsimonious solutions. Equipped with the developed bound, we prove the *global linear convergence* for a number of prevalent statistical models under mild assumptions, even though the problem turns out to be non-convex.

Keywords: sparsity, hard thresholding, compressed sensing, stochastic optimization

1. Introduction

Over the last two decades, pursuing sparse representations has emerged as a fundamental technique throughout bioinformatics (Olshausen and Field, 1997), statistics (Tibshirani, 1996; Efron et al., 2004), signal processing (Chen et al., 1998; Donoho et al., 2006; Donoho, 2006; Candès and Wakin, 2008) and mathematical science (Chandrasekaran et al., 2012), to name just a few. In order to obtain a sparse solution, a plethora of practical algorithms have been presented, among which two prominent examples are greedy pursuit and convex relaxation (Tropp and Wright, 2010). For instance, as one of the earliest greedy algorithms, orthogonal matching pursuit (OMP) (Pati et al., 1993) repeatedly picks a coordinate as the potential support of a solution. While OMP may fail for some deterministic sensing matrices, Tropp (2004); Tropp and Gilbert (2007) showed that it recovers the true signal with high probability when using random matrices such as Gaussian. Inspired by the success of OMP, the two concurrent work of compressive sampling matching pursuit (CoSaMP) (Needell

©2018 Jie Shen and Ping Li.

License: CC-BY 4.0, see `https://creativecommons.org/licenses/by/4.0/`. Attribution requirements are provided at `http://jmlr.org/papers/v18/16-299.html`.

and Tropp, 2009) and subspace pursuit (SP) (Dai and Milenkovic, 2009) made improvement by selecting multiple coordinates followed by a pruning step in each iteration, and the recovery condition was framed under the restricted isometry property (RIP) (Candès and Tao, 2005). Interestingly, the more careful selection strategy of CoSaMP and SP leads to an optimal sample complexity. The iterative hard thresholding (IHT) algorithm (Daubechies et al., 2004; Blumensath and Davies, 2008, 2009) gradually refines the iterates by gradient descent along with truncation. Foucart (2011) then developed a concise algorithm termed hard thresholding pursuit (HTP), which combined the idea of CoSaMP and IHT, and showed that HTP is superior to both in terms of the RIP condition. Jain et al. (2011) proposed an interesting variant of the HTP algorithm and obtained a sharper RIP result. Recently, Bahmani et al. (2013) and Yuan et al. (2018) respectively extended CoSaMP and HTP to general objective functions, for which a global convergence was established.

Since the sparsity constraint counts the number of non-zero components which renders the problem non-convex, the ℓ_1-norm was suggested as a convex relaxation dating back to basis pursuit (Chen et al., 1998; Donoho and Tsaig, 2008) and Lasso (Tibshirani, 1996). The difference is that Lasso looks for an ℓ_1-norm constrained solution that minimizes the residual while the principle of basis pursuit is to find a signal with minimal ℓ_1-norm that fits the observation data. Candès and Tao (2005) carried out a detailed analysis on the recovery performance of basis pursuit. Another popular estimator in the high-dimensional statistics is the Dantzig selector (Candès and Tao, 2007) which, instead of constraining the residual of the linear model, penalizes the maximum magnitude of the gradient. From a computational perspective, both basis pursuit and Dantzig selector can be solved by linear programming, while Lasso is formulated as a quadratic problem. Interestingly, under the RIP condition or the uniform uncertainty assumption (Candès et al., 2006), a series of work showed that exact recovery by convex programs is possible as soon as the observation noise vanishes (Candès and Tao, 2005; Candès, 2008; Wainwright, 2009; Cai et al., 2010; Foucart, 2012).

In this paper, we are interested in the hard thresholding (HT) operator underlying a large body of the developed algorithms in compressed sensing (e.g., IHT, CoSaMP, SP), machine learning (Yuan and Zhang, 2013), and statistics (Ma, 2013). Our motivation is two-fold. From a high level, compared to the convex programs, these HT-based algorithms are always orders of magnitude computationally more efficient, hence more practical for large-scale problems (Tropp and Wright, 2010). Nevertheless, they usually require a more stringent condition to guarantee the success. This naturally raises an interesting question of whether we can derive milder conditions for HT-based algorithms to achieve the best of the two worlds. For practitioners, to address the huge volume of data, a popular strategy in machine learning is to appeal to stochastic algorithms that sequentially update the solution. However, as many researchers observed (Langford et al., 2009; Duchi and Singer, 2009; Xiao, 2010), it is hard for the ℓ_1-based stochastic algorithms to preserve the sparse structure of the solution as the batch solvers do. This immediately poses the question of whether we are able to apply the principal idea of hard thresholding to stochastic algorithms while still ensuring a fast convergence.

To elaborate the problem more precisely, let us first turn to some basic properties of hard thresholding along with simple yet illustrative cases. For a general vector $\boldsymbol{b} \in \mathbb{R}^d$, the hard thresholded signal $\mathcal{H}_k(\boldsymbol{b})$ is formed by setting all but the largest (in magnitude) k elements of $\boldsymbol{b}$ to zero. Ties are broken lexicographically. Hence, the hard thresholded signal $\mathcal{H}_k(\boldsymbol{b})$ is always k-sparse, i.e., the number of non-zero components does not exceed k. Moreover, the resultant signal $\mathcal{H}_k(\boldsymbol{b})$ is a best

k-sparse approximation to $\boldsymbol{b}$ in terms of any ℓ_p norm ($p \geq 1$). That is, for any k-sparse vector $\boldsymbol{x}$

$$\|\mathcal{H}_k(\boldsymbol{b}) - \boldsymbol{b}\|_p \leq \|\boldsymbol{x} - \boldsymbol{b}\|_p.$$

In view of the above inequality, a broadly used bound in the literature for the deviation of the thresholded signal is as follows:

$$\|\mathcal{H}_k(\boldsymbol{b}) - \boldsymbol{x}\|_2 \leq 2\,\|\boldsymbol{b} - \boldsymbol{x}\|_2\,. \tag{1.1}$$

To gain intuition on the utility of (1.1) and to spell out the importance of offering a tight bound for it, let us consider the compressed sensing problem as an example for which we aim to recover the true sparse signal $\boldsymbol{x}$ from its linear measurements. Here, $\boldsymbol{b}$ is a good but dense approximation to $\boldsymbol{x}$ obtained by, e.g., full gradient descent. Then (1.1) justifies that in order to obtain a structured (i.e., sparse) approximation by hard thresholding, the distance of the iterate to the true signal $\boldsymbol{x}$ is upper bounded by a multiple of 2 to the one before. For comparison, it is worth mentioning that ℓ_1-based convex algorithms usually utilize the soft thresholding operator which enjoys the non-expansiveness property (Defazio et al., 2014), i.e., the iterate becomes closer to the optimum after projection. This salient feature might partially attribute to the wide range of applications of the ℓ_1-regularized formulations. Hence, to derive comparable performance guarantee, tightening the bound (1.1) is crucial in that it controls how much deviation the hard thresholding operator induces. This turns out to be more demanding for stochastic gradient methods, where the proxy $\boldsymbol{b}$ itself is affected by the randomness of sample realization. In other words, since $\boldsymbol{b}$ does not minimize the objective function (it only optimizes the objective in expectation), the deviation (1.1) makes it more challenging to analyze the convergence behavior. As an example, Nguyen et al. (2014) proposed a stochastic solver for general sparsity-constrained programs but suffered a non-vanishing optimization error due to randomness. This indicates that to mitigate the randomness barrier, we have to seek a better bound to control the precision of the thresholded solution and the variance.

1.1 Summary of Contributions

In this work, we make three contributions:

1. We examine the tightness of (1.1) that has been used for a decade in the literature and show that the equality therein will never be attained. We then improve this bound and quantitatively characterize that the deviation is inversely proportional to the value of $\sqrt{k}$. Our bound is tight, in the sense that the equality we build can be attained for specific signals, hence cannot be improved if no additional information is available. Our bound is universal in the sense that it holds for all choices of k-sparse signals $\boldsymbol{x}$ and for general signals $\boldsymbol{b}$.

2. Owing to the tight estimate, we demonstrate how the RIP (or RIP-like) condition assumed by a wide range of hard thresholding based algorithms can be relaxed. In the context of compressed sensing, it means that in essence, many more kinds of sensing matrices or fewer measurements can be utilized for data acquisition. For machine learning, it suggests that existing algorithms are capable of handling more difficult statistical models.

3. Finally, we present an computationally efficient algorithm that applies hard thresholding in large-scale setting and we prove its linear convergence to a global optimum up to the statistical precision of the problem. We also prove that with sufficient samples, our algorithm identifies

the true parameter for prevalent statistical models. Returning to (1.1), our analysis shows that only when the deviation is controlled below the multiple of 1.15 can such an algorithm succeed. This immediately implies that the conventional bound (1.1) is not applicable in the challenging scenario.

1.2 Notation

Before delivering the algorithm and main theoretical results, let us instate several pieces of notation that are involved throughout the paper. We use bold lowercase letters, e.g., $\boldsymbol{v}$, to denote a vector (either column or row) and its ith element is denoted by v_i. The ℓ_2-norm of a vector $\boldsymbol{v}$ is denoted by $\left\|\boldsymbol{v}\right\|_2$. The support set of $\boldsymbol{v}$, i.e., indices of non-zeros, is denoted by $\mathrm{supp}\left(\boldsymbol{v}\right)$ whose cardinality is written as $\left|\mathrm{supp}\left(\boldsymbol{v}\right)\right|$ or $\left\|\boldsymbol{v}\right\|_0$. We write bold capital letters such as $\boldsymbol{M}$ for matrices and its (i, j)-th entry is denoted by m_{ij}. The capital upright letter C and its subscript variants (e.g., $\mathrm{C}_0, \mathrm{C}_1$) are reserved for absolute constants whose values may change from appearance to appearance.

For an integer $d > 0$, suppose that Ω is a subset of $\{1,\ 2,\ \dots,\ d\}$. Then for a general vector $\boldsymbol{v} \in \mathbb{R}^d$, we define $\mathcal{P}_\Omega\left(\cdot\right)$ as the orthogonal projection onto the support set Ω which retains elements contained in Ω and sets others to zero. That is,

$$\left(\mathcal{P}_\Omega\left(\boldsymbol{v}\right)\right)_i = \begin{cases} v_i, & \text{if } i \in \Omega, \\ 0, & \text{otherwise.} \end{cases}$$

In particular, let Γ be the support set indexing the k largest absolute components of $\boldsymbol{v}$. In this way, the hard thresholding operator is given by

$$\mathcal{H}_k\left(\boldsymbol{v}\right) = \mathcal{P}_\Gamma(\boldsymbol{v}).$$

We will also use the orthogonal projection of a vector $\boldsymbol{v}$ onto an ℓ_2-ball with radius ω. That is,

$$\Pi_\omega(\boldsymbol{v}) = \frac{\boldsymbol{v}}{\max\{1, \left\|\boldsymbol{v}\right\|_2 / \omega\}}.$$

1.3 Roadmap

We present the key tight bound for hard thresholding in Section 2, along with a justification why the conventional bound (1.1) is not tight. We then discuss the implications of the developed tight bound to compressed sensing and machine learning in Section 3, which shows that the RIP or RIP-like condition can be improved for a number of popular algorithms. Thanks to our new estimation, Section 4 develops a novel stochastic algorithm which applies hard thresholding to large-scale problems and establishes the global linear convergence. A comprehensive empirical study on the tasks of sparse recovery and binary classification is carried out in Section 5. Finally, We conclude the paper in Section 6 and all the proofs are deferred to the appendix.

2. The Key Bound

We argue that the conventional bound (1.1) is not tight, in the sense that the equality therein can hardly be attained. To see this, recall how the bound was derived for a k-sparse signal $\boldsymbol{x}$ and a general one $\boldsymbol{b}$:

$$\left\|\mathcal{H}_k\left(\boldsymbol{b}\right) - \boldsymbol{x}\right\|_2 = \left\|\mathcal{H}_k\left(\boldsymbol{b}\right) - \boldsymbol{b} + \boldsymbol{b} - \boldsymbol{x}\right\|_2 \overset{\xi}{\leq} \left\|\mathcal{H}_k\left(\boldsymbol{b}\right) - \boldsymbol{b}\right\|_2 + \left\|\boldsymbol{b} - \boldsymbol{x}\right\|_2 \leq 2\left\|\boldsymbol{b} - \boldsymbol{x}\right\|_2,$$

where the last inequality holds because $\mathcal{H}_k(\boldsymbol{b})$ is a best k-sparse approximation to $\boldsymbol{b}$. The major issue occurs in ξ. Though it is the well-known triangle inequality and the equality could be attained if there is no restriction on the signals $\boldsymbol{x}$ and $\boldsymbol{b}$, we remind here that the signal $\boldsymbol{x}$ does have a specific structure – it is k-sparse. Note that in order to fulfill the equality in ξ, we must have $\mathcal{H}_k(\boldsymbol{b}) - \boldsymbol{b} = \gamma(\boldsymbol{b} - \boldsymbol{x})$ for some $\gamma \geq 0$, that is,

$$\mathcal{H}_k(\boldsymbol{b}) = (\gamma + 1)\boldsymbol{b} - \gamma\boldsymbol{x}. \tag{2.1}$$

One may verify that the above equality holds *if and only if*

$$\boldsymbol{x} = \boldsymbol{b} = \mathcal{H}_k(\boldsymbol{b}). \tag{2.2}$$

To see this, let Ω be the support set of $\mathcal{H}_k(\boldsymbol{b})$ and $\overline{\Omega}$ be the complement. Let $\boldsymbol{b}_1 = \mathcal{P}_\Omega(\boldsymbol{b}) = \mathcal{H}_k(\boldsymbol{b})$ and $\boldsymbol{b}_2 = \mathcal{P}_{\overline{\Omega}}(\boldsymbol{b})$. Likewise, we define $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ as the components of $\boldsymbol{x}$ supported on Ω and $\overline{\Omega}$ respectively. Hence, (2.1) indicates $\boldsymbol{x}_1 = \boldsymbol{b}_1$ and $\boldsymbol{x}_2 = (1+\gamma^{-1})\boldsymbol{b}_2$ where we assume $\gamma > 0$ since $\gamma = 0$ immediately implies $\mathcal{H}_k(\boldsymbol{b}) = \boldsymbol{b}$ and hence the equality of (1.1) does not hold. If $\|\boldsymbol{b}_1\|_0 < k$, then we have $\boldsymbol{x}_2 = \boldsymbol{b}_2 = \boldsymbol{0}$ since $\boldsymbol{b}_1$ contains the k largest absolute elements of $\boldsymbol{b}$. Otherwise, the fact that $\|\boldsymbol{x}\|_0 \leq k$ and $\boldsymbol{x}_1 = \boldsymbol{b}_1$ implies $\boldsymbol{x}_2 = \boldsymbol{0}$, and hence $\boldsymbol{b}_2$. Therefore, we obtain (2.2).

When (2.2) happens, however, we in reality have $\|\mathcal{H}_k(\boldsymbol{b}) - \boldsymbol{x}\|_2 = \|\boldsymbol{b} - \boldsymbol{x}\|_2 = 0$. In other words, the factor of 2 in (1.1) can essentially be replaced with an *arbitrary constant*! In this sense, we conclude that the bound (1.1) is not tight. Our new estimate for hard thresholding is as follows:

Theorem 1 (Tight Bound for Hard Thresholding) *Let $\boldsymbol{b} \in \mathbb{R}^d$ be an arbitrary vector and $\boldsymbol{x} \in \mathbb{R}^d$ be any K-sparse signal. For any $k \geq K$, we have the following bound:*

$$\|\mathcal{H}_k(\boldsymbol{b}) - \boldsymbol{x}\|_2 \leq \sqrt{\nu}\,\|\boldsymbol{b} - \boldsymbol{x}\|_2, \quad \nu = 1 + \frac{\rho + \sqrt{(4+\rho)\rho}}{2}, \quad \rho = \frac{\min\{K, d-k\}}{k - K + \min\{K, d-k\}}.$$

In particular, our bound is tight in the sense that there exist specific vectors of $\boldsymbol{b}$ and $\boldsymbol{x}$ such that the equality holds.

Remark 2 (Maximum of ν) *In contrast to the constant bound* (1.1)*, our result asserts that the deviation resulting from hard thresholding is inversely proportional to $\sqrt{k}$ (when $K \leq d-k$) in a universal manner. When k tends to d, ρ is given by $(d-k)/(d-K)$ which is still decreasing with respect to k. Thus, the maximum value of ρ equals one. Even in this case, we find that $\sqrt{\nu_{\max}} = \sqrt{1 + \frac{\sqrt{5}+1}{2}} = \frac{\sqrt{5}+1}{2} \approx 1.618$.*

Remark 3 *Though for some batch algorithms such as IHT and CoSaMP, the constant bound* (1.1) *suffices to establish the convergence due to specific conditions, we show in Section 4 that it cannot ensure the global convergence for stochastic algorithms.*

Remark 4 *When $\boldsymbol{x}$ is not exactly K-sparse, we still can bound the error by $\|\mathcal{H}_k(\boldsymbol{b}) - \boldsymbol{x}\|_2 \leq \|\mathcal{H}_k(\boldsymbol{b}) - \mathcal{H}_k(\boldsymbol{x})\|_2 + \|\mathcal{H}_k(\boldsymbol{x}) - \boldsymbol{x}\|_2$. Thus, without loss of generality, we assumed that the signal $\boldsymbol{x}$ is K-sparse.*

Proof (Sketch) Our bound follows from fully exploring the sparsity pattern of the signals and from fundamental arguments in optimization. Denote

$$\boldsymbol{w} := \mathcal{H}_k(\boldsymbol{b}).$$

Let Ω be the support set of $\boldsymbol{w}$ and let $\overline{\Omega}$ be its complement. We immediately have $\mathcal{P}_{\Omega}\left(\boldsymbol{b}\right) = \boldsymbol{w}$. Let Ω' be the support set of $\boldsymbol{x}$. Define

$$\boldsymbol{b}_1 = \mathcal{P}_{\Omega \backslash \Omega'}\left(\boldsymbol{b}\right), \quad \boldsymbol{b}_2 = \mathcal{P}_{\Omega \cap \Omega'}\left(\boldsymbol{b}\right), \quad \boldsymbol{b}_3 = \mathcal{P}_{\overline{\Omega} \backslash \Omega'}\left(\boldsymbol{b}\right), \quad \boldsymbol{b}_4 = \mathcal{P}_{\overline{\Omega} \cap \Omega'}\left(\boldsymbol{b}\right).$$

Likewise, we define $\boldsymbol{x}_i$ and $\boldsymbol{w}_i$ for $1 \leq i \leq 4$. Due to the construction, we have $\boldsymbol{w}_1 = \boldsymbol{b}_1, \boldsymbol{w}_2 = \boldsymbol{b}_2, \boldsymbol{w}_3 = \boldsymbol{w}_4 = \boldsymbol{x}_1 = \boldsymbol{x}_3 = \boldsymbol{0}$. Our goal is to estimate the maximum value of $\left\|\boldsymbol{w} - \boldsymbol{x}\right\|_2^2 / \left\|\boldsymbol{b} - \boldsymbol{x}\right\|_2^2$. It is easy to show that when attaining the maximum, $\left\|\boldsymbol{b}_3\right\|_2$ must be zero. Denote

$$\gamma := \frac{\left\|\boldsymbol{w} - \boldsymbol{x}\right\|_2^2}{\left\|\boldsymbol{b} - \boldsymbol{x}\right\|_2^2} = \frac{\left\|\boldsymbol{b}_1\right\|_2^2 + \left\|\boldsymbol{b}_2 - \boldsymbol{x}_2\right\|_2^2 + \left\|\boldsymbol{x}_4\right\|_2^2}{\left\|\boldsymbol{b}_1\right\|_2^2 + \left\|\boldsymbol{b}_2 - \boldsymbol{x}_2\right\|_2^2 + \left\|\boldsymbol{b}_4 - \boldsymbol{x}_4\right\|_2^2}. \tag{2.3}$$

Note that the variables here only involve $\boldsymbol{x}$ and $\boldsymbol{b}$. Arranging the equation we obtain

$$(\gamma - 1)\left\|\boldsymbol{b}_2 - \boldsymbol{x}_2\right\|_2^2 + \gamma\left\|\boldsymbol{b}_4 - \boldsymbol{x}_4\right\|_2^2 - \left\|\boldsymbol{x}_4\right\|_2^2 + (\gamma - 1)\left\|\boldsymbol{b}_1\right\|_2^2 = 0. \tag{2.4}$$

It is evident that for specific choices of $\boldsymbol{b}$ and $\boldsymbol{x}$, we have $\gamma = 1$. Since we are interested in the maximum of γ, we assume $\gamma > 1$ below. Fixing $\boldsymbol{b}$, we can view the left-hand side of the above equation as a function of $\boldsymbol{x}$. One may verify that the function has a positive definite Hessian matrix and thus it attains the minimum at stationary point given by

$$\boldsymbol{x}_2^* = \boldsymbol{b}_2, \quad \boldsymbol{x}_4^* = \frac{\gamma}{\gamma - 1}\boldsymbol{b}_4. \tag{2.5}$$

On the other hand, (2.4) implies that the minimum function value should not be greater than zero. Plugging the stationary point back gives

$$\left\|\boldsymbol{b}_1\right\|_2^2\gamma^2 - (2\left\|\boldsymbol{b}_1\right\|_2^2 + \left\|\boldsymbol{b}_4\right\|_2^2)\gamma + \left\|\boldsymbol{b}_1\right\|_2^2 \leq 0.$$

Solving the above inequality with respect to γ, we obtain

$$\gamma \leq 1 + \left(2\left\|\boldsymbol{b}_1\right\|_2^2\right)^{-1}\left(\left\|\boldsymbol{b}_4\right\|_2^2 + \sqrt{\left(4\left\|\boldsymbol{b}_1\right\|_2^2 + \left\|\boldsymbol{b}_4\right\|_2^2\right)\left\|\boldsymbol{b}_4\right\|_2^2}\right). \tag{2.6}$$

To derive an upper bound that is uniform over the choice of $\boldsymbol{b}$, we recall that $\boldsymbol{b}_1$ contains the largest absolute elements of $\boldsymbol{b}$ while $\boldsymbol{b}_4$ has smaller values. In particular, the average in $\boldsymbol{b}_1$ is larger than that in $\boldsymbol{b}_4$, which gives

$$\left\|\boldsymbol{b}_4\right\|_2^2 / \left\|\boldsymbol{b}_4\right\|_0 \leq \left\|\boldsymbol{b}_1\right\|_2^2 / \left\|\boldsymbol{b}_1\right\|_0.$$

Note that $\left\|\boldsymbol{b}_1\right\|_0 = k - \left\|\boldsymbol{b}_2\right\|_0 = k - (K - \left\|\boldsymbol{b}_4\right\|_0)$. Hence, combining with the fact that $0 \leq \left\|\boldsymbol{b}_4\right\|_0 \leq \min\{K, d-k\}$ and optimizing over $\left\|\boldsymbol{b}_4\right\|_0$ in the above inequality gives

$$\left\|\boldsymbol{b}_4\right\|_2^2 \leq \frac{\min\{K, d-k\}}{k - K + \min\{K, d-k\}}\left\|\boldsymbol{b}_1\right\|_2^2. \tag{2.7}$$

Finally, we arrive at a uniform upper bound

$$\gamma \leq 1 + \frac{\rho + \sqrt{(4 + \rho)\rho}}{2}, \quad \rho = \frac{\min\{K, d-k\}}{k - K + \min\{K, d-k\}}.$$

See Appendix B for the full proof. ■

Remark 5 (Tightness) *We construct proper vectors $\boldsymbol{b}$ and $\boldsymbol{x}$ to establish the tightness of our bound by a backward induction. Note that γ equals ν if and only if $\|\boldsymbol{b}_4\|_2^2 = \rho \|\boldsymbol{b}_1\|_2^2$. Hence, we pick*

$$\|\boldsymbol{b}_4\|_2^2 = \rho \|\boldsymbol{b}_1\|_2^2, \quad \boldsymbol{x}_2 = \boldsymbol{b}_2, \quad \boldsymbol{x}_4 = \frac{\nu}{\nu - 1}\boldsymbol{b}_4, \tag{2.8}$$

where $\boldsymbol{x}_2$ and $\boldsymbol{x}_4$ are actually chosen as the stationary point as in (2.5)*. We note that the quantity of ν only depends on d, k and K, not on the components of $\boldsymbol{b}$ or $\boldsymbol{x}$. Plugging the above back to* (2.3) *justifies $\gamma = \nu$.*

It remains to show that our choices in (2.8) *do not violate the definition of $\boldsymbol{b}_i$'s, i.e., we need to ensure that the elements in $\boldsymbol{b}_1$ or $\boldsymbol{b}_2$ are equal to or greater than those in $\boldsymbol{b}_3$ or $\boldsymbol{b}_4$. Note that there is no such constraint for the K-sparse vector $\boldsymbol{x}$. Let us consider the case $K < d - k$ and $\|\boldsymbol{b}_4\|_0 = K$, so that $\|\boldsymbol{b}_1\|_0 = k$ and $\rho = K/k$. Thus, the first equality of* (2.8) *holds as soon as all the entries of $\boldsymbol{b}$ have same magnitude. The fact $\|\boldsymbol{b}_4\|_0 = K$ also implies Ω' is a subset of $\overline{\Omega}$ due to the definition of $\boldsymbol{b}_4$ and the sparsity of $\boldsymbol{x}$, hence we have $\boldsymbol{x}_2 = \boldsymbol{0} = \boldsymbol{b}_2$. Finally, picking $\boldsymbol{x}_4$ as we did in* (2.8) *completes the reasoning since it does not violate the sparsity constraint on $\boldsymbol{x}$.*

As we pointed out and just verified, the bound given by Theorem 1 is tight. However, if there is additional information for the signals, a better bound can be established. For instance, let us further assume that the signal $\boldsymbol{b}$ is r-sparse. If $r \leq k$, then $\boldsymbol{b}_4$ is a zero vector and (2.6) reads as $\gamma \leq 1$. Otherwise, we have $\|\boldsymbol{b}_4\|_0 \leq \min\{K, r-k\}$ and (2.7) is improved to

$$\|\boldsymbol{b}_4\|_2^2 \leq \frac{\min\{K, r-k\}}{k - K + \min\{K, r-k\}} \|\boldsymbol{b}_1\|_2^2 .$$

Henceforth, we can show that the parameter ρ is given by

$$\rho = \frac{\min\{K, r-k\}}{k - K + \min\{K, r-k\}}.$$

Note that the fact $r \leq d$ implies that the above is a tighter bound than the one in Theorem 1.

We would also like to mention that in Lemma 1 of Jain et al. (2014), a closely related bound was established:

$$\|\mathcal{H}_k(\boldsymbol{b}) - \boldsymbol{b}\|_2 \leq \sqrt{\frac{d-k}{d-K}} \|\boldsymbol{b} - \boldsymbol{x}\|_2 . \tag{2.9}$$

One may use this nice result to show that

$$\|\mathcal{H}_k(\boldsymbol{b}) - \boldsymbol{x}\|_2 \leq \|\mathcal{H}_k(\boldsymbol{b}) - \boldsymbol{b}\|_2 + \|\boldsymbol{b} - \boldsymbol{x}\|_2 \leq \left(1 + \sqrt{\frac{d-k}{d-K}}\right) \|\boldsymbol{b} - \boldsymbol{x}\|_2 , \tag{2.10}$$

which also improves on (1.1) provided $k > K$. However, one shortcoming of (2.10) is that the factor depends on the dimension. For comparison, we recall that in the regime $K \leq d - k$, our bound is free of the dimension. This turns out to be a salient feature to integrate hard thresholding into stochastic methods, and we will comment on it more in Section 4.

3. Implications to Compressed Sensing

In this section, we investigate the implications of Theorem 1 for compressed sensing and signal processing. Since most of the HT-based algorithms utilize the deviation bound (1.1) to derive the convergence condition, they can be improved by our new bound. We exemplify the power of our theorem on two popular algorithms: IHT (Blumensath and Davies, 2009) and CoSaMP (Needell and Tropp, 2009). We note that our analysis also applies to their extensions such as Bahmani et al. (2013). To be clear, the purpose of this section is not dedicated to improving the best RIP condition for which recovery is possible by any methods (either convex or non-convex). Rather, we focus on two broadly used greedy algorithms and illustrate how our bound improves on previous results.

We proceed with a brief review of the problem setting in compressed sensing. Compressed sensing algorithms aim to recover the true K-sparse signal $\boldsymbol{x}^* \in \mathbb{R}^d$ from a set of its (perhaps noisy) measurements

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}^* + \boldsymbol{\varepsilon}, \tag{3.1}$$

where $\boldsymbol{\varepsilon} \in \mathbb{R}^d$ is some observation noise and $\boldsymbol{A}$ is a known $n \times d$ sensing matrix with $n \ll d$, hence the name compressive sampling. In general, the model is not identifiable since it is an underdetermined system. Yet, the prior knowledge that $\boldsymbol{x}^*$ is sparse radically changes the premise. That is, if the geometry of the sparse signal is preserved under the action of the sampling matrix $\boldsymbol{A}$ for a restricted set of directions, then it is possible to invert the sampling process. Such a novel idea was quantified as the kth restricted isometry property of $\boldsymbol{A}$ by Candès and Tao (2005), which requires that there exists a constant $\delta \geq 0$, such that for all k-sparse signals $\boldsymbol{x}$

$$(1-\delta)\left\|\boldsymbol{x}\right\|_2^2 \leq \left\|\boldsymbol{A}\boldsymbol{x}\right\|_2^2 \leq (1+\delta)\left\|\boldsymbol{x}\right\|_2^2. \tag{3.2}$$

The kth restricted isometry constant (RIC) δ_k is then defined as the smallest one that satisfies the above inequalities. Note that $\delta_{2k} < 1$ is the minimum requirement for distinguishing all k-sparse signals from the measurements. This is because for two arbitrary k-sparse vectors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ and their respective measurements $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$, the RIP condition reads as

$$(1-\delta_{2k})\left\|\boldsymbol{x}_1 - \boldsymbol{x}_2\right\|_2^2 \leq \left\|\boldsymbol{y}_1 - \boldsymbol{y}_2\right\|_2^2 \leq (1+\delta_{2k})\left\|\boldsymbol{x}_1 - \boldsymbol{x}_2\right\|_2^2,$$

for which $\delta_{2k} < 1$ guarantees that $\boldsymbol{x}_1 \neq \boldsymbol{x}_2$ implies $\boldsymbol{y}_1 \neq \boldsymbol{y}_2$. To date, there are three quintessential examples known to exhibit a profound restricted isometry behavior as long as the number of measurements is large enough: Gaussian matrices (optimal RIP, i.e., very small δ_k), partial Fourier matrices (fast computation) and Bernoulli ensembles (low memory footprint). Notably, it was shown in recent work that random matrices with a heavy-tailed distribution also satisfy the RIP with overwhelming probability (Adamczak et al., 2011; Li et al., 2014).

Equipped with the standard RIP condition, many efficient algorithms have been developed. A partial list includes ℓ_1-norm based convex programs, IHT, CoSaMP, SP and regularized OMP (Needell and Vershynin, 2010), along with much interesting work devoted to improving or sharpening the RIP condition (Wang and Shim, 2012; Mo and Shen, 2012; Cai and Zhang, 2013; Mo, 2015). To see why relaxing RIP is of central interest, note that the standard result (Baraniuk et al., 2008) asserts that the RIP condition $\delta_k \leq \delta$ holds with high probability over the draw of $\boldsymbol{A}$ provided

$$n \geq \mathrm{C}_0 \delta^{-2} k \log(d/k). \tag{3.3}$$

Hence, a slight relaxation of the condition $\delta_k \leq \delta$ may dramatically decrease the number of measurements. That being said, since the constant C_0 above is unknown, in general one cannot tell the precise sample size for greedy algorithms. Estimating the constant is actually the theme of phase transition (Donoho and Tanner, 2010; Donoho et al., 2013). While precise phase transition for ℓ_1-based convex programs has been well understood (Wainwright, 2009), an analogous result for greedy algorithms remains an open problem. Notably, in Blanchard and Tanner (2015), phase transition for IHT/CoSaMP was derived using the constant bound (1.1). We believe that our tight bound shall sharpen these results and we leave it as our future work. In the present paper, we focus on the ubiquitous RIP condition. In the language of RIP, we establish improved results.

3.1 Iterative Hard Thresholding

The IHT algorithm recovers the underlying K-sparse signal $\boldsymbol{x}^*$ by iteratively performing a full gradient descent on the least-squares loss followed by a hard thresholding step. That is, IHT starts with an arbitrary point $\boldsymbol{x}^0$ and at the t-th iteration, it updates the new solution as follows:

$$\boldsymbol{x}^t = \mathcal{H}_k\left(\boldsymbol{x}^{t-1} + \boldsymbol{A}^\top(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}^{t-1})\right). \tag{3.4}$$

Note that Blumensath and Davies (2009) used the parameter $k = K$. However, in practice one may only know to an upper bound on the true sparsity K. Thus, we consider the projection sparsity k as a parameter that depends on K. To establish the global convergence with a geometric rate of 0.5, Blumensath and Davies (2009) applied the bound (1.1) and assumed the RIP condition

$$\delta_{2k+K} \leq 0.18. \tag{3.5}$$

As we have shown, (1.1) is actually not tight and hence, their results, especially the RIP condition can be improved by Theorem 1.

Theorem 6 *Consider the model* (3.1) *and the IHT algorithm* (3.4)*. Pick $k \geq K$ and let $\{\boldsymbol{x}^t\}_{t\geq 1}$ be the iterates produced by IHT. Then, under the RIP condition $\delta_{2k+K} \leq 1/\sqrt{8\nu}$, for all $t \geq 1$*

$$\left\|\boldsymbol{x}^t - \boldsymbol{x}^*\right\|_2 \leq 0.5^t \left\|\boldsymbol{x}^0 - \boldsymbol{x}^*\right\|_2 + \mathrm{C}\left\|\boldsymbol{\varepsilon}\right\|_2,$$

where ν is given by Theorem 1.

Let us first study the vanilla case $k = K$. Blumensath and Davies (2009) required $\delta_{3K} \leq 0.18$ whereas our analysis shows $\delta_{3K} \leq 0.22$ suffices. Note that even a little relaxation on RIP is challenging and may require several pages of mathematical induction (Candès, 2008; Cai et al., 2010; Foucart, 2012). In contrast, our improvement comes from a direct application of Theorem 1 which only modifies several lines of the original proof in Blumensath and Davies (2009). See Appendix C for details. In view of (3.3), we find that the necessary number of measurements for IHT is dramatically reduced with a factor of 0.67 by our new theorem in that the minimum requirement of n is inversely proportional to the square of δ_{2k+K}.

Another important consequence of the theorem is a characterization on the RIP condition and the sparsity parameter, which, to the best of our knowledge, has not been studied in the literature. In Blumensath and Davies (2009), when gradually tuning k larger than K, it always requires $\delta_{2k+K} \leq 0.18$. Note that due to the monotonicity of RIC, i.e., $\delta_r \leq \delta_{r'}$ if $r \leq r'$, the condition turns out to be more and more stringent. Compared to their result, since ν is inversely proportional to $\sqrt{k}$, Theorem 6 is powerful especially when k becomes larger. For example, suppose $k = 20K$. In this case, Theorem 6 justifies that IHT admits the linear convergence as soon as $\delta_{41K} \leq 0.32$ whereas Blumensath and Davies (2009) requires $\delta_{41K} \leq 0.18$. Such a property is appealing in practice, in that among various real-world applications, the true sparsity is indeed unknown and we would like to estimate a conservative upper bound on it.

On the other hand, for a given sensing matrix, there does exist a fundamental limit for the maximum choice of k. To be more precise, the condition in Theorem 6 together with the probabilistic argument (3.3) require

$$1/\sqrt{8\nu} \geq \delta_{2k+K}, \quad \mathrm{C}_1 \nu(2k+K) \log\left(d/(2k+K)\right) \leq n.$$

Although it could be very interesting to derive a quantitative characterization for the maximum value of k, we argue that it is perhaps intractable owing to two aspects: First, it is known that one has to enumerate all the combinations of the $2k+K$ columns of $\boldsymbol{A}$ to compute the restricted isometry constant δ_{2k+K} (Bah and Tanner, 2010, 2014). This suggests that it is NP-hard to estimate the largest admissible value of k. Also, there is no analytic solution of the stationary point for the left-hand side of the second inequality.

3.2 Compressive Sampling Matching Pursuit

The CoSaMP algorithm proposed by Needell and Tropp (2009) is one of the most efficient algorithms for sparse recovery. Let $F(\boldsymbol{x}) = \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2$. CoSaMP starts from an arbitrary initial point $\boldsymbol{x}^0$ and proceeds as follows:

$$\begin{aligned}
\Omega^t &= \mathrm{supp}\left(\nabla F(\boldsymbol{x}^{t-1}),\ k\right) \cup \mathrm{supp}\left(\boldsymbol{x}^{t-1}\right), \\
\boldsymbol{b}^t &= \arg\min_{\boldsymbol{x}}\ F(\boldsymbol{x}),\ \text{s.t.}\ \mathrm{supp}\left(\boldsymbol{x}\right) \subset \Omega^t, \\
\boldsymbol{x}^t &= \mathcal{H}_k\left(\boldsymbol{b}^t\right).
\end{aligned}$$

Compared to IHT which performs hard thresholding after gradient update, CoSaMP prunes the gradient at the beginning of each iteration, followed by solving a least-squares program restricted on a small support set. In particular, in the last step, CoSaMP applies hard thresholding to form a k-sparse iterate for future updates. The analysis of CoSaMP consists of bounding the estimation error in each step. Owing to Theorem 1, we advance the theoretical result of CoSaMP by improving the error bound for its last step, and hence the RIP condition.

Theorem 7 *Consider the model* (3.1) *and the CoSaMP algorithm. Pick $k \geq K$ and let $\{\boldsymbol{x}^t\}_{t\geq 1}$ be the iterates produced by CoSaMP. Then, under the RIP condition*

$$\delta_{3k+K} \leq \frac{\left(\sqrt{32\nu+49}-9\right)^{1/2}}{4\sqrt{\nu-1}},$$

it holds that for all $t \geq 1$

$$\left\|\boldsymbol{x}^t - \boldsymbol{x}^*\right\|_2 \leq 0.5^t \left\|\boldsymbol{x}^0 - \boldsymbol{x}^*\right\|_2 + \mathrm{C}\left\|\boldsymbol{\varepsilon}\right\|_2,$$

where ν is given by Theorem 1.

Roughly speaking, the bound is still inversely proportional to $\sqrt{\nu}$. Hence, it is monotonically increasing with respect to k, indicating our theorem is more effective for a large quantity of k. In fact, for the CoSaMP algorithm, our bound above is superior to the best known result even when $k = K$. To see this, we have the RIP condition $\delta_{4K} \leq 0.31$. In comparison, Needell and Tropp (2009) derived a bound $\delta_{4K} \leq 0.1$ and Foucart and Rauhut (2013, Theorem 6.27) improved it to $\delta_{4K} < 0.29$ for a geometric rate of 0.5. We notice that for binary sparse vectors, Jain et al. (2014) presented a different proof technique and obtained the RIP condition $\delta_{4K} \leq 0.35$ for CoSaMP.

4. Hard Thresholding in Large-Scale Optimization

Now we move on to the machine learning setting where our focus is pursuing an optimal sparse solution that minimizes a given objective function based on a set of training samples $Z_1^n := \{Z_i\}_{i=1}^n$. Different from compressed sensing, we usually have sufficient samples which means n can be very large. Therefore, the computational complexity is of primary interest. Formally, we are interested in optimizing the following program:

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} F(\boldsymbol{x}; Z_1^n) = \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}; Z_i), \quad \text{s.t. } \|\boldsymbol{x}\|_0 \leq K,\ \|\boldsymbol{x}\|_2 \leq \omega. \tag{4.1}$$

The global optimum of the above problem is denoted by $\boldsymbol{x}_{\mathrm{opt}}$. We note that the objective function is presumed to be decomposable with respect to the samples. This is quite a mild condition and most of the popular machine learning models fulfill it. Typical examples include (but not limited to) the sparse linear regression and sparse logistic regression:

- Sparse Linear Regression: For all $1 \leq i \leq n$, we have $Z_i = (\boldsymbol{a}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ and the loss function $F(\boldsymbol{x}; Z_1^n) = \frac{1}{2n}\left\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\right\|_2^2$ is the least-squares and can be explained by $f(\boldsymbol{x}; Z_i) = \frac{1}{2}\left\|\boldsymbol{a}_i \cdot \boldsymbol{x} - y_i\right\|_2^2$.
- Sparse Logistic Regression: For all $1 \leq i \leq n$, we have $Z_i = (\boldsymbol{a}_i, y_i) \in \mathbb{R}^d \times \{+1, -1\}$ and the negative log-likelihood is penalized, i.e., $F(\boldsymbol{x}; Z_1^n) = \frac{1}{n}\sum_{i=1}^n \log\left(1 + \exp\left(-y_i \boldsymbol{a}_i \cdot \boldsymbol{x}\right)\right)$ for which $f(\boldsymbol{x}; Z_i) = \log\left(1 + \exp\left(-y_i \boldsymbol{a}_i \cdot \boldsymbol{x}\right)\right)$.

To ease notation, we will often write $F(\boldsymbol{x}; Z_1^n)$ as $F(\boldsymbol{x})$ and $f(\boldsymbol{x}; Z_i)$ as $f_i(\boldsymbol{x})$ for $i = 1, 2, \cdots, n$. It is worth mentioning that the objective function $F(\boldsymbol{x})$ is allowed to be non-convex. Hence, in order to ensure the existence of a global optimum, a natural option is to impose an ℓ_p-norm ($p \geq 1$) constraint (Loh and Wainwright, 2012, 2015). Here we choose the ℓ_2-norm constraint owing to its fast projection. Previous work, e.g., Agarwal et al. (2012) prefers the computationally less efficient ℓ_1-norm to promote sparsity and to guarantee the existence of optimum. In our problem, yet, we already have imposed the hard sparsity constraint so the ℓ_2-norm constraint is a better fit.

The major contribution of this section is a computationally efficient algorithm termed hard thresholded stochastic variance reduced gradient method (HT-SVRG) to optimize (4.1), tackling one of the most important problems in large-scale machine learning: producing sparse solutions by stochastic methods. We emphasize that the formulation (4.1) is in stark contrast to the ℓ_1-regularized programs considered by previous stochastic solvers such as Prox-SVRG (Xiao and Zhang, 2014) and SAGA (Defazio et al., 2014). We target here a stochastic algorithm for the *non-convex* problem that is less exploited in the literature. From a theoretical perspective, (4.1) is more difficult to analyze but it always produces sparse solutions, whereas performance guarantees for convex programs are fruitful but one cannot characterize the sparsity of the obtained solution (usually the solution is not sparse). When we appeal to stochastic algorithms to solve the convex programs, the ℓ_1-norm formulation becomes much less effective in terms of sparsification, naturally owing to the randomness. See Langford et al. (2009); Xiao (2010); Duchi and Singer (2009) for more detailed discussion on the issue. We also remark that existing work such as Yuan et al. (2018); Bahmani et al. (2013); Jain et al. (2014) investigated the sparsity-constrained problem (4.1) in a batch scenario, which is not practical for large-scale learning problems. The perhaps most related work to our new algorithm is Nguyen et al. (2014). Nonetheless, the optimization error therein does not vanish for noisy statistical models.

Our main result shows that for prevalent statistical models, our algorithm is able to recover the true parameter with a linear rate. Readers should distinguish the optimal solution $\boldsymbol{x}_{\mathrm{opt}}$ and the true parameter. For instance, consider the model (3.1). Minimizing (4.1) does not amount to recovering $\boldsymbol{x}^*$ if there is observation noise. In fact, the convergence to $\boldsymbol{x}_{\mathrm{opt}}$ is only guaranteed to an accuracy reflected by the *statistical precision* of the problem, i.e., $\left\|\boldsymbol{x}^* - \boldsymbol{x}_{\mathrm{opt}}\right\|_2$, which is the best one can hope for any statistical model (Agarwal et al., 2012). We find that the global convergence is attributed to both the tight bound and the variance reduction technique to be introduced below, and examining the necessity of them is an interesting future work.

Algorithm 1 Hard Thresholded Stochastic Variance Reduced Gradient Method (HT-SVRG)

```
Require: Training samples $\{Z_i\}_{i=1}^n$, maximum stage count $S$, sparsity parameter $k$, update frequency $m$, learning rate $\eta$, radius $\omega$, initial solution $\widetilde{\boldsymbol{x}}^0$.
Ensure: Optimal solution $\widetilde{\boldsymbol{x}}^S$.
for $s = 1$ to $S$ do
    Set $\widetilde{\boldsymbol{x}} = \widetilde{\boldsymbol{x}}^{s-1},\ \widetilde{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^n \nabla f_i(\widetilde{\boldsymbol{x}}),\ \boldsymbol{x}^0 = \widetilde{\boldsymbol{x}}$.
    for $t = 1$ to $m$ do
        Uniformly pick $i_t \in \{1, 2, \cdots, n\}$ and update the solution
            $\boldsymbol{b}^t = \boldsymbol{x}^{t-1} - \eta\left(\nabla f_{i_t}(\boldsymbol{x}^{t-1}) - \nabla f_{i_t}(\widetilde{\boldsymbol{x}}) + \widetilde{\boldsymbol{\mu}}\right),$
            $\boldsymbol{r}^t = \mathcal{H}_k\left(\boldsymbol{b}^t\right),$
            $\boldsymbol{x}^t = \Pi_\omega(\boldsymbol{r}^t).$
    end for
    Uniformly choose $j^s \in \{0, 1, \cdots, m-1\}$ and set $\widetilde{\boldsymbol{x}}^s = \boldsymbol{x}^{j^s}$.
end for
```

4.1 Algorithm

Our algorithm (Algorithm 1) applies the framework of Johnson and Zhang (2013), where the primary idea is to leverage past gradients for the current update for the sake of variance reduction – a technique that has a long history in statistics (Owen and Zhou, 2000). To guarantee that each iterate is k-sparse, it then invokes the hard thresholding operation. Note that the orthogonal projection for $\boldsymbol{r}^t$ will not change the support set, and hence $\boldsymbol{x}^t$ is still k-sparse. Also note that our sparsity constraint in (4.1) reads as $\left\|\boldsymbol{x}\right\|_0 \leq K$. What we will show below is that when the parameter k is properly chosen (which depends on K), we obtain a globally convergent sequence of iterates.

The most challenging part on establishing the global convergence comes from the hard thresholding operation $\mathcal{H}_k\left(\boldsymbol{r}^t\right)$. Note that it is $\boldsymbol{b}^t$ that reduces the objective value in expectation. If $\boldsymbol{b}^t$ is not k-sparse (usually it is dense), $\boldsymbol{x}^t$ is not equal to $\boldsymbol{b}^t$ so it does not decrease the objective function. In addition, compared with the convex proximal operator (Defazio et al., 2014) which enjoys the non-expansiveness of the distance to the optimum, the hard thresholding step can enlarge the distance up to a multiple of 2 if using the bound (1.1). What makes it a more serious issue is that these inaccurate iterates $\boldsymbol{x}^t$ will be used for future updates, and hence the error might be progressively propagated at an exponential rate.

Our key idea is to first bound the curvature of the function from below and above to establish RIP-like condition, which, combined with Theorem 1, downscales the deviation resulting from hard thresholding. Note that ν is always greater than one (see Theorem 1), hence the curvature bound is necessary. Due to variance reduction, we show that the optimization error vanishes when restricted on a small set of directions as soon as we have sufficient samples. Moreover, with hard thresholding we are able to control the error per iteration and to obtain near-optimal sample complexity.

4.2 Deterministic Analysis

We will first establish a general theorem that characterizes the progress of HT-SVRG for approximating an arbitrary K-sparse signal $\widehat{\boldsymbol{x}}$. Then we will discuss how to properly choose the hyper-parameters of the algorithm. Finally we move on to specify $\widehat{\boldsymbol{x}}$ to develop convergence results for a global optimum of (4.1) and for a true parameter (e.g., $\boldsymbol{x}^*$ of the compressed sensing problem).

4.2.1 Assumption

Our analysis depends on two properties of the curvature of the objective function that have been standard in the literature. Readers may refer to Bickel et al. (2009); Negahban et al. (2009); Jain et al. (2014) for a detailed description.

Definition 8 (Restricted Strong Convexity) *A differentiable function $g : \mathbb{R}^d \rightarrow \mathbb{R}$ is said to satisfy the property of restricted strong convexity (RSC) with parameter $\alpha_r > 0$, if for all vectors $\boldsymbol{x}$, $\boldsymbol{x}' \in \mathbb{R}^d$ with $\left\|\boldsymbol{x} - \boldsymbol{x}'\right\|_0 \leq r$, it holds that*

$$g(\boldsymbol{x}') - g(\boldsymbol{x}) - \left\langle \nabla g(\boldsymbol{x}), \boldsymbol{x}' - \boldsymbol{x} \right\rangle \geq \frac{\alpha_r}{2} \left\|\boldsymbol{x}' - \boldsymbol{x}\right\|_2^2 .$$

Definition 9 (Restricted Smoothness) *A differentiable function $g : \mathbb{R}^d \rightarrow \mathbb{R}$ is said to satisfy the property of restricted smoothness (RSS) with parameter $L_r > 0$, if for all vectors $\boldsymbol{x}$, $\boldsymbol{x}' \in \mathbb{R}^d$ with $\left\|\boldsymbol{x} - \boldsymbol{x}'\right\|_0 \leq r$, it holds that*

$$\left\|\nabla g(\boldsymbol{x}') - \nabla g(\boldsymbol{x})\right\|_2 \leq L_r \left\|\boldsymbol{x}' - \boldsymbol{x}\right\|_2 .$$

With these definitions, we assume the following:

$(A1)$ $F(\boldsymbol{x})$ satisfies the RSC condition with parameter α_{k+K}.

$(A2)$ For all $1 \leq i \leq n$, $f_i(\boldsymbol{x})$ satisfies the RSS condition with parameter L_{3k+K}.

Here, we recall that K was first introduced in (4.1) and the parameter k was used in our algorithm. Compared to the convex algorithms such as SAG (Roux et al., 2012), SVRG (Johnson and Zhang,

2013) and SAGA (Defazio et al., 2014) that assume strong convexity and smoothness everywhere, we only assume these in a restricted sense. This is more practical especially in the high dimensional regime where the Hessian matrix could be degenerate (Agarwal et al., 2012). We also stress that the RSS condition is imposed on each $f_i(\boldsymbol{x})$, whereas prior work requires it for $F(\boldsymbol{x})$ which is milder than ours (Negahban et al., 2009).

4.2.2 UPPER BOUND OF PROGRESS

For brevity, let us denote

$$L := L_{3k+K}, \quad \alpha := \alpha_{k+K}, \quad c := L/\alpha,$$

where we call the quantity c as the condition number of the problem. It is also crucial to measure the ℓ_2-norm of the gradient restricted on sparse directions, and we write

$$\left\|\nabla_{3k+K} F(\boldsymbol{x})\right\|_2 := \max_{\Omega} \left\{ \left\|\mathcal{P}_{\Omega}\left(\nabla F(\boldsymbol{x})\right)\right\|_2 : \ |\Omega| \leq 3k + K \right\}.$$

Note that for convex programs, the above evaluated at a global optimum is zero. As will be clear, $\left\|\nabla_{3k+K} F(\boldsymbol{x})\right\|_2$ reflects how close the iterates returned by HT-SVRG can be to the point $\boldsymbol{x}$. For prevalent statistical models, it vanishes when there are sufficient samples. Related to this quantity, our analysis also involves

$$Q(\boldsymbol{x}) := \left(16\nu\eta^2 L\omega m + \frac{2\omega}{\alpha}\right) \left\|\nabla_{3k+K} F(\boldsymbol{x})\right\|_2 + 4\nu\eta^2 m \left\|\nabla_{3k+K} F(\boldsymbol{x})\right\|_2^2,$$

where we recall that ν is the expansiveness factor given by Theorem 1, η and m are used in the algorithm and ω is a universal constant that upper bounds the ℓ_2-norm of the signal we hope to estimate. Virtually, with an appropriate parameter setting, $Q(\boldsymbol{x})$ scales as $\left\|\nabla_{3k+K} F(\boldsymbol{x})\right\|_2$ which will be clarified. For a particular stage s, we denote $\mathcal{I}^s := \{i_1, i_2, \cdots, i_m\}$, i.e., the samples randomly chosen for updating the solution.

Theorem 10 *Consider Algorithm 1 and a K-sparse signal $\widehat{\boldsymbol{x}}$ of interest. Assume $(A1)$ and $(A2)$. Pick the step size $0 < \eta < 1/(4L)$. If $\nu < 4L/(4L - \alpha)$, then it holds that*

$$\mathbb{E}\left[F(\widetilde{\boldsymbol{x}}^s) - F(\widehat{\boldsymbol{x}})\right] \leq \beta^s \left[F(\widetilde{\boldsymbol{x}}^0) - F(\widehat{\boldsymbol{x}})\right] + \tau(\widehat{\boldsymbol{x}}),$$

where the expectation is taken over $\{\mathcal{I}^1, j^1, \mathcal{I}^2, j^2, \cdots, \mathcal{I}^s, j^s\}$ and $0 < \beta < 1$ provided that m is large enough. In particular, for $1/(1 - \eta\alpha) < \nu < 4L/(4L - \alpha)$, we have

$$\beta = \beta_1 := \frac{1}{(2\nu\eta\alpha - 2\nu\eta^2\alpha L - \nu + 1)\, m} + \frac{2\nu\eta^2\alpha L}{2\nu\eta\alpha - 2\nu\eta^2\alpha L - \nu + 1},$$

$$\tau(\widehat{\boldsymbol{x}}) = \tau_1(\widehat{\boldsymbol{x}}) := \frac{\alpha Q(\widehat{\boldsymbol{x}})}{2(2\nu\eta\alpha - 2\nu\eta^2\alpha L - \nu + 1)(1 - \beta_1)m}.$$

For $\nu \leq 1/(1 - \eta\alpha)$, we have

$$\beta = \beta_2 := \frac{1}{\nu\eta\alpha(1 - 2\eta L)m} + \frac{2\eta L}{1 - 2\eta L}, \quad \tau(\widehat{\boldsymbol{x}}) = \tau_2(\widehat{\boldsymbol{x}}) := \frac{Q(\widehat{\boldsymbol{x}})}{2\nu\eta\alpha(1 - 2\eta L)(1 - \beta_2)m}.$$

The proof can be found in Appendix D.1.

Remark 11 *For the theorem to hold,* $\sqrt{\nu} < \sqrt{4L/(4L-\alpha)} \leq \sqrt{4/3} \approx 1.15$ *due to* $L \geq \alpha$*. Hence, the conventional bound* (1.1) *is not applicable. In contrast, Theorem 1 asserts that this condition can be fulfilled by tuning* k *slightly larger than* K*.*

Remark 12 *With the conditions on* η *and* ν*, the coefficient* β *is always less than one provided that* m *is sufficiently large.*

Remark 13 *The theorem does* not *assert convergence to an arbitrary sparse vector* $\widehat{\boldsymbol{x}}$*. This is because* $F(\widetilde{\boldsymbol{x}}^s) - F(\widehat{\boldsymbol{x}})$ *might be less than zero. However, specifying* $\widehat{\boldsymbol{x}}$ *does give convergence results, as to be elaborated later.*

4.2.3 HYPER-PARAMETER SETTING

Before moving on to the convergence guarantee, let us discuss the minimum requirement on the hyper-parameters k, m and η, and determine how to choose them to simplify Theorem 10.

For the sake of success of HT-SVRG, we require $\nu < 4c/(4c-1)$, which implies $\rho < 1/(16c^2 - 4c)$. Recall that ρ is given in Theorem 1. In general, we are interested in the regime $K \leq k \ll d$. Hence, we have $\rho = K/k$ and the minimum requirement for the sparsity parameter is

$$k > (16c^2 - 4c)K. \tag{4.2}$$

To our knowledge, the idea of relaxed sparsity was first introduced in Zhang (2011) for OMP and in Jain et al. (2014) for projected gradient descent. However, the relaxed sparsity here emerges in a different way in that HT-SVRG is a stochastic algorithm, and their proof technique cannot be used.

We also contrast our tight bound to the inequality (2.10) that is obtained by combining the triangle inequality and Lemma 1 of Jain et al. (2014). Following our proof pipeline, (2.10) gives

$$k \geq \left(1 - \left(\sqrt{4c(4c-1)^{-1}} - 1\right)^2\right) d + \left(\sqrt{4c(4c-1)^{-1}} - 1\right)^2 K$$

which grows with the dimension d, whereas using Theorem 1 the sparsity parameter k depends only on the desired sparsity K. In this regard, we conclude that for the stochastic case, our bound is vital.

Another component of the algorithm is the update frequency m. Intuitively, HT-SVRG performs m number of stochastic gradient update followed by a full gradient evaluation, in order to mitigate the variance. In this light, m should not be too small. Otherwise, the algorithm reduces to the full gradient method which is not computationally efficient. On the other spectrum, a large m leads to a slow convergence that is reflected in the convergence coefficient β. To quantitatively analyze how m should be selected, let us consider the case $\nu \leq 1/(1-\eta\alpha)$ for example. The case $1/(1-\eta\alpha) < \nu < 4L/(4L-\alpha)$ follows in a similar way. In order to ensure $\beta_2 < 1$, we must have $m > 1/\left(\nu\eta\alpha(1-4\eta L)\right)$. In particular, picking

$$\eta = \frac{\eta'}{L}, \quad \eta' \in (0, 1/4), \tag{4.3}$$

we find that the update frequency m has to satisfy

$$m > \frac{c}{\nu\eta'(1-\eta')}, \tag{4.4}$$

which is of the same order as in the convex case (Johnson and Zhang, 2013) when $\eta' = \Theta(1)$. Note that the way we choose the learning rate $\eta = \eta'/L$ is also a common practice in convex optimization (Nesterov, 2004).

With (4.2), (4.3) and (4.4) in mind, we provide detailed choices of the hyper-parameters. Due to $0 < \eta < 1/(4L)$, β_1 is monotonically increasing with respect to ν. By Theorem 1, we know that ν is decreasing with respect to k. Thus, a larger quantity of k results in a smaller value of β_1, and hence a faster rate. Interestingly, for β_2 we discover that the smaller the k is, the faster the algorithm concentrates. Hence, we have the following consequence:

Proposition 14 *Fix η and m. Then the optimal choice of ν in Theorem 10 is $\nu = 1/(1-\eta\alpha)$ in the sense that the convergence coefficient β attains the minimum.*

In light of the proposition, in the sections to follow, we will only consider the setting $\nu = 1/(1-\eta\alpha)$. But we emphasize that our analysis and results essentially apply to any $\nu \leq 4L/(4L-\alpha)$.

Now let

$$\eta = \frac{1}{8L}, \quad m = 4(8c-1), \quad k = 8c(8c-1)K. \tag{4.5}$$

This gives

$$\beta = \frac{2}{3}, \quad \tau(\widehat{\boldsymbol{x}}) = \frac{5\omega}{\alpha} \left\| \nabla_{3k+K} F(\widehat{\boldsymbol{x}}) \right\|_2 + \frac{1}{\alpha L} \left\| \nabla_{3k+K} F(\widehat{\boldsymbol{x}}) \right\|_2^2. \tag{4.6}$$

4.2.4 GLOBAL LINEAR CONVERGENCE

We are in the position to state the global linear convergence to an optimum of the sparsity-constrained optimization program (4.1).

Corollary 15 *Assume (A1) and (A2). Consider the HT-SVRG algorithm with hyper-parameters given in (4.5). Then the sequence $\{\widetilde{\boldsymbol{x}}^s\}_{s\geq 1}$ converges linearly to a global optimum $\boldsymbol{x}_{\mathrm{opt}}$ of (4.1)*

$$\begin{aligned}\mathbb{E}\left[F(\widetilde{\boldsymbol{x}}^s) - F(\boldsymbol{x}_{\mathrm{opt}})\right] \leq \left(\frac{2}{3}\right)^s &\left[F(\widetilde{\boldsymbol{x}}^0) - F(\boldsymbol{x}_{\mathrm{opt}})\right] \\ &+ \frac{5\omega}{\alpha} \left\| \nabla_{3k+K} F(\boldsymbol{x}_{\mathrm{opt}}) \right\|_2 + \frac{1}{\alpha L} \left\| \nabla_{3k+K} F(\boldsymbol{x}_{\mathrm{opt}}) \right\|_2^2.\end{aligned}$$

Proof This is a direct consequence of Theorem 10. ■

Whenever $\nabla_{3k+K} F(\boldsymbol{x}_{\mathrm{opt}}) = \boldsymbol{0}$, the corollary reads as

$$\mathbb{E}\left[F(\widetilde{\boldsymbol{x}}^s) - F(\boldsymbol{x}_{\mathrm{opt}})\right] \leq \left(\frac{2}{3}\right)^s \left[F(\widetilde{\boldsymbol{x}}^0) - F(\boldsymbol{x}_{\mathrm{opt}})\right].$$

It implies that if one is solving a convex problem without the sparsity constraint but the optimal solution happens to be sparse, it is safe to perform hard thresholding without loss of optimality. We exemplify such behavior with another algorithm SAGA (Defazio et al., 2014) in Appendix E. In the noiseless compressed sensing setting where $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}^*$, the corollary guarantees that HT-SVRG exactly recovers the underlying true signal $\boldsymbol{x}^*$ when $F(\boldsymbol{x})$ is chosen as the least-squares loss in that $\boldsymbol{x}_{\mathrm{opt}} = \boldsymbol{x}^*$ and $\nabla F(\boldsymbol{x}^*) = \boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{y}) = \boldsymbol{0}$.

On the other side, the RSC property implies that

$$\left\| \widetilde{\boldsymbol{x}}^s - \widehat{\boldsymbol{x}} \right\|_2 \leq \sqrt{\frac{2\max\{F(\widetilde{\boldsymbol{x}}^s) - F(\widehat{\boldsymbol{x}}), 0\}}{\alpha}} + \frac{2\left\| \nabla_{k+K} F(\widehat{\boldsymbol{x}}) \right\|_2}{\alpha}.$$

The proof is straightforward and can be found in Lemma 14 of Shen and Li (2017a). Now we specify $\widehat{\boldsymbol{x}}$ as the true parameter of some statistical model, for instance, $\boldsymbol{x}^*$ in (3.1). It is hence possible to establish recovery guarantee of $\boldsymbol{x}^*$, which is known as the problem of parameter estimation.

Corollary 16 *Assume* $(A1)$ *and* $(A2)$. *Let* L' *be the RSS parameter of* $F(\boldsymbol{x})$ *at the sparsity level* $3k+K$. *Consider the HT-SVRG algorithm with hyper-parameters given in* (4.5). *Then the sequence* $\{\widetilde{\boldsymbol{x}}^s\}_{s\geq 1}$ *recovers a* K*-sparse signal* $\boldsymbol{x}^*$ *with a geometric rate*

$$\mathbb{E}\left[\, \|\widetilde{\boldsymbol{x}}^s - \boldsymbol{x}^*\|_2 \,\right] \leq \sqrt{\frac{2L'}{\alpha}} \cdot \left(\frac{2}{3}\right)^{\frac{s}{2}} \left\|\widetilde{\boldsymbol{x}}^0 - \boldsymbol{x}^*\right\|_2 + \sqrt{\frac{10\omega}{\alpha^2}\, \|\nabla_{3k+K} F(\boldsymbol{x}^*)\|_2} + \left(\sqrt{\frac{2}{\alpha^3}} + \frac{3}{\alpha}\right) \|\nabla_{3k+K} F(\boldsymbol{x}^*)\|_2 \,.$$

The proof can be found in Appendix D.2.

Remark 17 *The RSS parameter* L' *of* $F(\boldsymbol{x})$ *always ranges in* $[\alpha, L]$, *which is simply by definition.*

4.2.5 COMPUTATIONAL COMPLEXITY

We compare the computational complexity of HT-SVRG to that of projected gradient descent (PGD) studied in Jain et al. (2014), which is a batch counterpart to HT-SVRG. First, we remark that the analysis of PGD is based on the smoothness parameter L' of $F(\boldsymbol{x})$ at sparsity level $2k + K$. We write $c' = L'/\alpha$. To achieve a given accuracy $\epsilon > 0$, PGD requires $\mathcal{O}\left(c' \log(1/\epsilon)\right)$ iterations. Hence the total computational complexity is $\mathcal{O}\left(nc'd \log(1/\epsilon)\right)$. For HT-SVRG, in view of Corollary 15, the convergence coefficient is a constant. Hence, HT-SVRG needs $\mathcal{O}\left(\log(1/\epsilon)\right)$ iterations where we note that the error term $\|\nabla_{3k+K} F(\boldsymbol{x}^*)\|_2$ can be made as small as ϵ with sufficient samples (to be clarified in the sequel). In each stage, HT-SVRG computes a full gradient $\widetilde{\boldsymbol{\mu}}$ followed by m times stochastic updates. Therefore, the total complexity of HT-SVRG is given by $\mathcal{O}\left((n + c)d \log(1/\epsilon)\right)$ by noting the fact $m = \mathcal{O}\left(c\right)$. In the scenario $c < n(c' - 1)$, HT-SVRG significantly improves on PGD in terms of time cost.

4.3 Statistical Results

The last ingredient of our theorem is the term $\tau(\widehat{\boldsymbol{x}})$ which measures how close the iterates could be to a given sparse signal $\widehat{\boldsymbol{x}}$. With appropriate hyper-parameter settings, the quantity relies exclusively on $\|\nabla_{3k+K} F(\widehat{\boldsymbol{x}})\|_2$, as suggested by (4.6). Thereby, this section is dedicated to characterizing $\|\nabla_{3k+K} F(\widehat{\boldsymbol{x}})\|_2$. We will also give examples for which HT-SVRG is computationally more efficient than PGD. For the purpose of a concrete result, we study two problems: sparse linear regression and sparse logistic regression. These are two of the most popular statistical models in the literature and have found a variety of applications in machine learning and statistics (Raskutti et al., 2011). Notably, it is known that similar statistical results can be built for low-rank matrix regression, sparse precision matrix estimation, as suggested in Negahban et al. (2009); Agarwal et al. (2012).

4.3.1 SPARSE LINEAR REGRESSION

For sparse linear regression, the observation model is given by

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}^* + \boldsymbol{\varepsilon}, \quad \|\boldsymbol{x}^*\|_0 \leq K,\ \|\boldsymbol{x}^*\|_2 \leq \omega, \tag{4.7}$$

where $\boldsymbol{A} \in \mathbb{R}^{n \times d}$ is the design matrix, $\boldsymbol{y} \in \mathbb{R}^n$ is the response, $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ is some noise, and $\boldsymbol{x}^*$ is the K-sparse true parameter we hope to estimate from the knowledge of $\boldsymbol{A}$ and $\boldsymbol{y}$. Note that when we have the additional constraint $n \ll d$, the model above is exactly that of compressed sensing (3.1).

In order to (approximately) estimate the parameter, a natural approach is to optimize the following non-convex program:

$$\min_{\boldsymbol{x}} \ F(\boldsymbol{x}) := \frac{1}{2n} \sum_{i=1}^{n} \left\| y_i - \boldsymbol{a}_i \cdot \boldsymbol{x} \right\|_2^2, \quad \text{s.t. } \left\| \boldsymbol{x} \right\|_0 \leq K, \ \left\| \boldsymbol{x} \right\|_2 \leq \omega. \tag{4.8}$$

For our analysis, we assume the following on the design matrix and the noise:

$(A3)$ $\boldsymbol{a}_1, \boldsymbol{a}_2, \dots, \boldsymbol{a}_n$ are independent and identically distributed (i.i.d.) Gaussian random vectors $N(\boldsymbol{0}, \boldsymbol{\Sigma})$. All the diagonal elements of $\boldsymbol{\Sigma}$ satisfy $\Sigma_{jj} \leq 1$. The noise $\boldsymbol{\varepsilon}$ is independent of $\boldsymbol{A}$ and its entries are i.i.d. Gaussian random variables $N(0, \sigma^2)$.

Proposition 18 *Consider the sparse linear regression model* (4.7) *and the program* (4.8). *Assume* $(A3)$. *Then for a sparsity level* r,

- *with probability at least* $1 - \exp(-\mathrm{C}_0 n)$,

$$\alpha_r = \lambda_{\min}(\boldsymbol{\Sigma}) - \mathrm{C}_1 \frac{r \log d}{n}, \quad L'_r = \lambda_{\max}(\boldsymbol{\Sigma}) + \mathrm{C}_2 \frac{r \log d}{n};$$

- *with probability at least* $1 - \mathrm{C}_3 r/d$

$$L_r = \mathrm{C}_4 r \log d;$$

- *and with probability at least* $1 - \mathrm{C}_5/d$

$$\left\| \nabla_r F(\boldsymbol{x}^*) \right\|_2 \leq \mathrm{C}_6 \sigma \sqrt{\frac{r \log d}{n}}, \quad \left\| \nabla_r F(\boldsymbol{x}_{\mathrm{opt}}) \right\|_2 \leq L'_r \left\| \boldsymbol{x}_{\mathrm{opt}} - \boldsymbol{x}^* \right\|_2 + \mathrm{C}_6 \sigma \sqrt{\frac{r \log d}{n}}.$$

Above, $\lambda_{\min}(\boldsymbol{\Sigma})$ *and* $\lambda_{\max}(\boldsymbol{\Sigma})$ *are the minimum and maximum singular values of* $\boldsymbol{\Sigma}$ *respectively.*

We recall that α_r and L_r are involved in our assumptions $(A1)$ and $(A2)$, and L'_r is the RSS parameter of $F(\boldsymbol{x})$. The estimation for α_r, L'_r and $\left\| \nabla_r F(\boldsymbol{x}^*) \right\|_2$ follows from standard results in the literature (Raskutti et al., 2011), while that for L_r follows from Proposition E.1 in Bellec et al. (2016) by noting the fact that bounding L_r amounts to estimating $\max_i \left\| \mathcal{H}_r(\boldsymbol{a}_i) \right\|_2^2$. In order to estimate $\left\| \nabla_r F(\boldsymbol{x}_{\mathrm{opt}}) \right\|_2$, notice that

$$\begin{aligned} \left\| \nabla_r F(\boldsymbol{x}_{\mathrm{opt}}) \right\|_2 &\leq \left\| \nabla_r F(\boldsymbol{x}_{\mathrm{opt}}) - \nabla_r F(\boldsymbol{x}^*) \right\|_2 + \left\| \nabla_r F(\boldsymbol{x}^*) \right\|_2 \\ &\leq \left\| \nabla F(\boldsymbol{x}_{\mathrm{opt}}) - \nabla F(\boldsymbol{x}^*) \right\|_2 + \left\| \nabla_r F(\boldsymbol{x}^*) \right\|_2 \\ &\leq L'_r \left\| \boldsymbol{x}_{\mathrm{opt}} - \boldsymbol{x}^* \right\|_2 + \left\| \nabla_r F(\boldsymbol{x}^*) \right\|_2 , \end{aligned}$$

where we use the definition of RSS in the last inequality.

Now we let $r = 3k + K = \text{const} \cdot c^2 K$ and get $\alpha = \lambda_{\min}(\boldsymbol{\Sigma}) - \mathrm{C}_1 \frac{c^2 K \log d}{n}$, $L = \mathrm{C}_4 c^2 K \log d$. Suppose that $\lambda_{\min}(\boldsymbol{\Sigma}) = 2\mathrm{C}_4 (K \log d)^2$ and $n = q \cdot \frac{\mathrm{C}_1}{\mathrm{C}_4} K \log d$ with $q \geq 1$. Then our assumptions $(A1)$ and $(A2)$ are met with high probability with

$$\alpha = \mathrm{C}_4 (K \log d)^2, \ L = \mathrm{C}_4 (K \log d)^3, \ \text{and } c = K \log d.$$

For Corollary 15, as far as

$$s \geq \mathrm{C}_7 \log\left(\frac{F(\widetilde{\boldsymbol{x}}^0) - F(\boldsymbol{x}_{\mathrm{opt}})}{\epsilon}\right),\ n = \mathrm{C}_7\,(\omega\sigma)^2\,\epsilon^{-2} K \log d,$$

we have

$$\mathbb{E}\left[F(\widetilde{\boldsymbol{x}}^s) - F(\boldsymbol{x}_{\mathrm{opt}})\right] \leq \epsilon + \frac{\lambda_{\max}(\boldsymbol{\Sigma})}{\lambda_{\min}(\boldsymbol{\Sigma})}\left\|\boldsymbol{x}_{\mathrm{opt}} - \boldsymbol{x}^*\right\|_2 + \left(\frac{\lambda_{\max}(\boldsymbol{\Sigma})}{\lambda_{\min}(\boldsymbol{\Sigma})}\left\|\boldsymbol{x}_{\mathrm{opt}} - \boldsymbol{x}^*\right\|_2\right)^2$$

for some accuracy parameter $\epsilon > 0$. This suggests that it is possible for HT-SVRG to approximate a global optimum of (4.1) up to $\left\|\boldsymbol{x}_{\mathrm{opt}} - \boldsymbol{x}^*\right\|_2$, namely the statistical precision of the problem.

Returning to Corollary 16, to guarantee that

$$\mathbb{E}\left[\,\left\|\widetilde{\boldsymbol{x}}^s - \boldsymbol{x}^*\right\|_2\,\right] \leq \epsilon,$$

it suffices to pick

$$s \geq \mathrm{C}_8 \log(\omega\sqrt{c'}/\epsilon), \quad n = \mathrm{C}_8(\omega\sigma)^2\epsilon^{-4} K \log d.$$

Finally, we compare the computational cost to PGD. It is not hard to see that under the same situation $\lambda_{\min}(\boldsymbol{\Sigma}) = 2\mathrm{C}_4(K\log d)^2$ and $n = \frac{\mathrm{C}_1}{\mathrm{C}_4} K \log d$,

$$L' = \mathrm{C}_4(K\log d)^3,\ c' = K \log d,\ \text{provided that } \lambda_{\max}(\boldsymbol{\Sigma}) = \mathrm{C}_4(K\log d)^3 - \frac{\mathrm{C}_2\mathrm{C}_4}{\mathrm{C}_1}(K\log d)^2.$$

Thus $c < n(c' - 1)$, i.e., HT-SVRG is more efficient than PGD. It is also possible to consider other regimes of the covariance matrix and the sample size, though we do not pursue it here.

4.3.2 SPARSE LOGISTIC REGRESSION

For sparse logistic regression, the observation model is given by

$$\Pr(y_i \mid \boldsymbol{a}_i;\ \boldsymbol{x}^*) = \frac{1}{1 + \exp(-y_i \boldsymbol{a}_i \cdot \boldsymbol{x}^*)},\quad \left\|\boldsymbol{x}^*\right\|_0 \leq K,\ \left\|\boldsymbol{x}\right\|_2 \leq \omega,\ \forall\, 1 \leq i \leq n, \tag{4.9}$$

where y_i is either 0 or 1. It then learns the parameter by minimizing the negative log-likelihood:

$$\min_{\boldsymbol{x}}\ F(\boldsymbol{x}) := \frac{1}{n}\sum_{i=1}^{n} \log\left(1 + \exp(-y_i \boldsymbol{a}_i \cdot \boldsymbol{x})\right), \quad \text{s.t. } \left\|\boldsymbol{x}\right\|_0 \leq K,\ \left\|\boldsymbol{x}\right\|_2 \leq \omega. \tag{4.10}$$

There is a large body of work showing that the statistical property is rather analogous to that of linear regression. See, for example, Negahban et al. (2009). In fact, the statistical results apply to generalized linear models as well.

4.4 A Concurrent Work

After we posted the first version Shen and Li (2016) on arXiv, Li et al. (2016) made their work public where a similar algorithm to HT-SVRG was presented. Their theoretical analysis applies to convex objective functions while we allow the function $F(\boldsymbol{x})$ to be non-convex. We also fully characterize the convergence behavior of the algorithm by showing the trade-off between the sparsity parameter k and the convergence coefficient β (Proposition 14).

5. Experiments

In this section, we present a comprehensive empirical study for the proposed HT-SVRG algorithm on two tasks: sparse recovery (compressed sensing) and image classification. The experiments on sparse recovery is dedicated to verifying the theoretical results we presented, and we visualize the classification models learned by HT-SVRG to demonstrate the practical efficacy.

5.1 Sparse Recovery

To understand the practical behavior of our algorithm as well as to justify the theoretical analysis, we perform experiments on synthetic data. The experimental settings are as follows:

- **Data Generation.** The data dimension d is fixed as 256 and we generate an $n \times d$ Gaussian random sensing matrix $\boldsymbol{A}$ whose entries are i.i.d. with zero mean and variance $1/n$. Then 1000 K-sparse signals $\boldsymbol{x}^*$ are independently generated, where the support of each signal is uniformly chosen. That is, we run our algorithm and the baselines for 1000 trials. The measurements $\boldsymbol{y}$ for each signal $\boldsymbol{x}^*$ is obtained by $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}^*$ which is noise free. In this way, we are able to study the convergence rate by plotting the logarithm of the objective value since the optimal objective value is known to be zero.

- **Baselines.** We mainly compare with two closely related algorithms: IHT and PGD. Both of them compute the full gradient of the least-squares loss followed by hard thresholding. Yet, PGD is more general, in the sense that it allows the sparsity parameter k to be larger than the true sparsity K ($k = K$ for IHT) and also considers a flexible step size η ($\eta = 1$ for IHT). Hence, PGD can be viewed as a batch counterpart to our method HT-SVRG.

- **Evaluation Metric.** We say a signal $\boldsymbol{x}^*$ is successfully recovered by a solution $\boldsymbol{x}$ if

  $$\frac{\|\boldsymbol{x} - \boldsymbol{x}^*\|_2}{\|\boldsymbol{x}^*\|_2} < 10^{-3}.$$

  In this way, we can compute the percentage of success over the 1000 trials for each algorithm.

- **Hyper-Parameters.** If not specified, we use $m = 3n$, $k = 9K$, and $S = 10000$ for HT-SVRG. We also use the heuristic step size $\eta = 2/\mathrm{svds}(\boldsymbol{A}\boldsymbol{A}^\top)$ for HT-SVRG and PGD, where $\mathrm{svds}(\boldsymbol{A}\boldsymbol{A}^\top)$ returns the largest singular value of the matrix $\boldsymbol{A}\boldsymbol{A}^\top$. Since for each stage, HT-SVRG computes the full gradient for $(2m/n + 1)$ times, we run the IHT and PGD for $(2m/n + 1)S$ iterations for fair comparison, i.e., all of the algorithms have the same number of full gradient evaluations.

5.1.1 PHASE TRANSITION

Our first simulation aims at offering a big picture on the recovery performance. To this end, we vary the number of measurements n from 1 to 256, roughly with a step size 8. We also study the performance with respect to the true sparsity parameter K, which ranges from 1 to 26, roughly with step size 2. The results are illustrated in Figure 1, where a brighter block means a higher percentage of success and the brightest ones indicate exact sparse recovery. It is apparent that PGD and HT-SVRG require fewer measurements for an accurate recovery than IHT, possibly due to the flexibility in choosing the sparsity parameter and the step size. We also observe that as a stochastic algorithm, HT-SVRG performs comparably to PGD. This suggests that HT-SVRG is an appealing solution to large-scale sparse learning problems in that HT-SVRG is computationally more efficient.

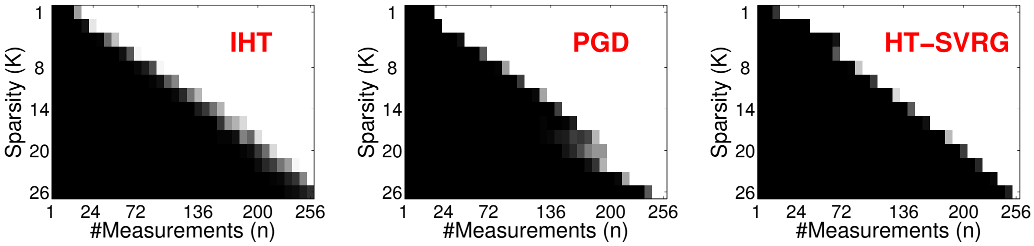

Figure 1: **Percentage of successful recovery under various sparsity and sample size.** The values range from 0 to 100, where a brighter color means a higher percentage of success (the brightest blocks correspond to the value of 100). PGD admits a higher percentage of recovery compared to IHT because it flexibly chooses the step size and sparsity parameter. As a stochastic variant, HT-SVRG performs comparably to the batch counterpart PGD.

In Figure 2, we exemplify some of the results obtained from HT-SVRG by plotting two kinds of curves: the success of percentage against the sample size n and that against the signal sparsity K. In this way, one can examine the detailed values and can determine the minimum sample size for a particular sparsity. For instance, the left panel tells that to ensure that 80% percents of the 16-sparse signals are recovered, we have to collect 175 measurements. We can also learn from the right panel that using 232 measurements, any signal whose sparsity is 22 or less can be reliably recovered.

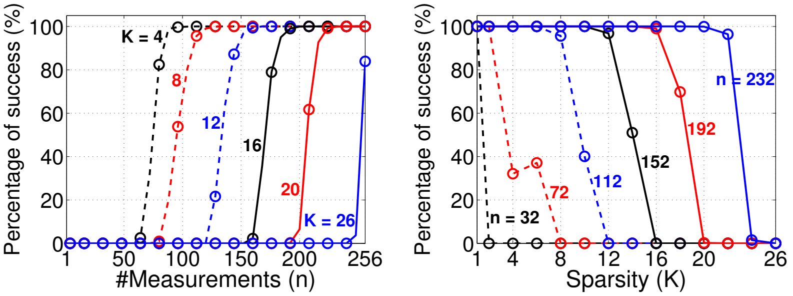

Figure 2: **Percentage of success of HT-SVRG against the number of measurements (left) and the sparsity (right).**

Based on the results in Figure 1 and Figure 2, we have an approximate estimation on the minimum requirement of the sample size which ensures accurate (or exact) recovery. Now we are to investigate how many measurements are needed to guarantee a success percentage of 95% and 99%. To this end, for each signal sparsity K, we look for the number of measurements n_0 from Figure 1 where 90 percents of success are achieved. Then we carefully enlarge n_0 with step size 1 and run the algorithms. The empirical results are recorded in Figure 3, where the circle markers represent the empirical results with different colors indicating different algorithms, e.g., red circle for empirical observation of HT-SVRG. Then we fit these empirical results by linear regression, which are plotted as solid or dashed lines. For example, the green line is a fitted model for IHT. We find that n is almost linear with K. Especially, the curve of HT-SVRG is nearly on top of that of PGD, which again verifies HT-SVRG is an attractive alternative to the batch method.

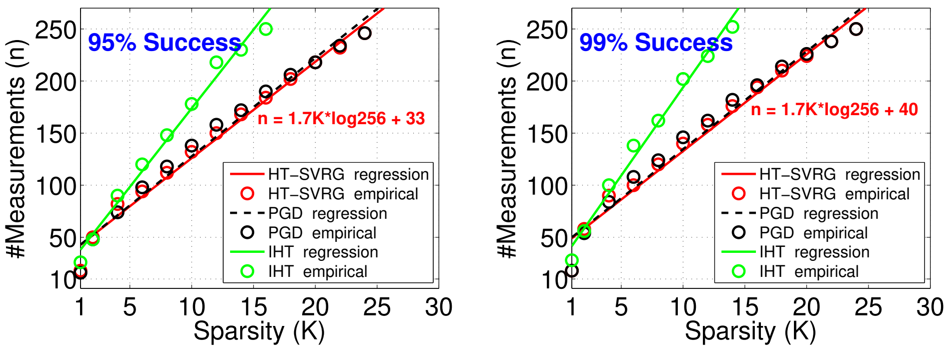

Figure 3: **Minimum number of measurements to achieve** 95% **and** 99% **percentage of success.** Red equation indicates the linear regression of HT-SVRG. The markers and curves for HT-SVRG are almost on top of PGD, which again justifies that HT-SVRG is an appealing stochastic alternative to the batch method PGD.

5.1.2 INFLUENCE OF HYPER-PARAMETERS

Next, we turn to investigate the influence of the hyper-parameters, i.e., the sparsity parameter k, update frequency m and step size η on the convergence behavior of HT-SVRG. We set the true sparsity $K = 4$ and collect 100 measurements for each groundtruth signal, i.e., $n = 100$. Note that the standard setting we employed is $k = 9K = 36$, $m = 3n = 300$ and $\eta = 2/\mathrm{svds}(\boldsymbol{A}\boldsymbol{A}^\top) \approx 0.3$. Each time we vary one of these parameters while fixing the other two, and the results are plotted in Figure 4. We point out that although the convergence result (Theorem 10) is deterministic, the vanishing optimization error (Proposition 18) is guaranteed under a probabilistic argument. Hence, it is possible that for a specific configuration of parameters, 97% of the signals are exactly recovered but HT-SVRG fails on the remaining, as we have observed in, e.g., Figure 2. Clearly, we are not supposed to average all the results to examine the convergence rate. For our purpose, we set a threshold 95%, that is, we average over the success trials if more than 95% percents of the signals are exactly recovered. Otherwise, we say that the set of parameters cannot ensure convergence and we average over these failure signals which will give an illustration of divergence.

The left panel of Figure 4 verifies the condition that k has to be larger than K, while the second panel shows the update frequency m can be reasonably small in the price of a slow convergence rate. Finally, the empirical study demonstrates that our heuristic choice $\eta = 0.3$ works well, and when $\eta > 3$, the objective value exceeds 10^{120} within 3 stages (which cannot be depicted in the figure). For very small step sizes, we plot the convergence curve by gradually enlarging the update frequency m in Figure 5. The empirical results agree with Theorem 10 that for any $0 < \eta < 1/(4L)$, HT-SVRG converges as soon as m is large enough.

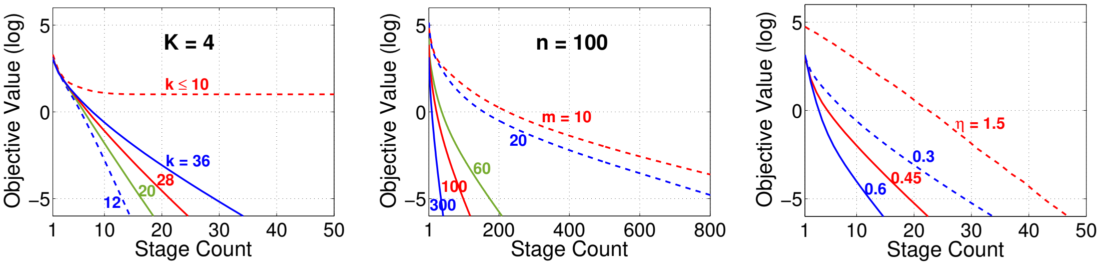

Figure 4: **Convergence of HT-SVRG with different parameters.** We have 100 measurements for the 256-dimensional signal where only 4 elements are non-zero. The standard setting is $k = 36$, $m = 300$ and $\eta = 0.3$. **Left:** If the sparsity parameter k is not large enough, HT-SVRG will not recover the signal. **Middle:** A small m leads to a frequent full gradient evaluation and hence slow convergence. **Right:** We observe divergence when $\eta \geq 3$.

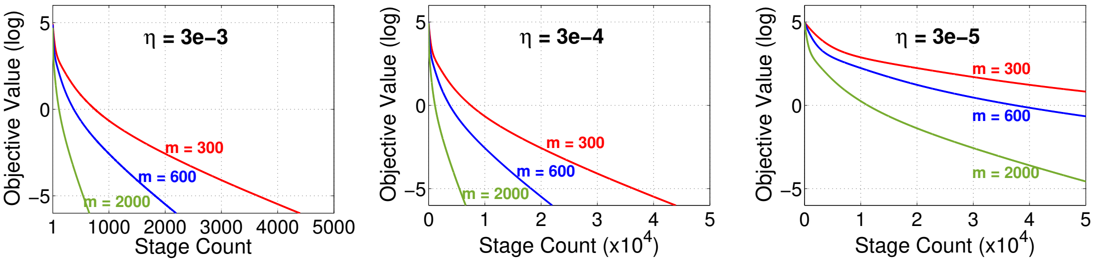

Figure 5: **Convergence behavior under small step size.** We observe that as long as we pick a sufficiently large value for m, HT-SVRG always converges. This is not surprising since our theorem guarantees for any $\eta < 1/(4L)$, HT-SVRG will converge if m is large enough. Also note that the geometric convergence rate is observed after certain iterations, e.g., for $\eta = 3 \times 10^{-5}$, the log(error) decreases linearly after 20 thousands iterations.

5.2 Classification

In addition to the application of sparse recovery, we illustrated that HT-SVRG can deal with binary classification by minimizing the sparse logistic regression problem (4.10). Here, we study the performance on a realistic image dataset MNIST[1], consisting of 60 thousands training samples and 10 thousands samples for testing. There is one digit on each image of size 28-by-28, hence totally 10 classes. Some of the images are shown in Figure 6.

The update frequency m is fixed as $m = 3n$. We compute the heuristic step size η as in the previous section, i.e., $\eta = 2/\mathrm{svds}(\boldsymbol{A}\boldsymbol{A}^\top) \approx 10^{-3}$. Since for the real-world dataset, the true sparsity is actually unknown, we tune the sparsity parameter k and study the performance of the algorithm.

First, we visualize five pair-wise models learned by HT-SVRG in Figure 7, where each row is associated with a binary classification task indicated by the two digits at the leading of the row, and the subsequent red-blue figures are used to illustrate the learned models under different spar-

[1] `http://yann.lecun.com/exdb/mnist/`

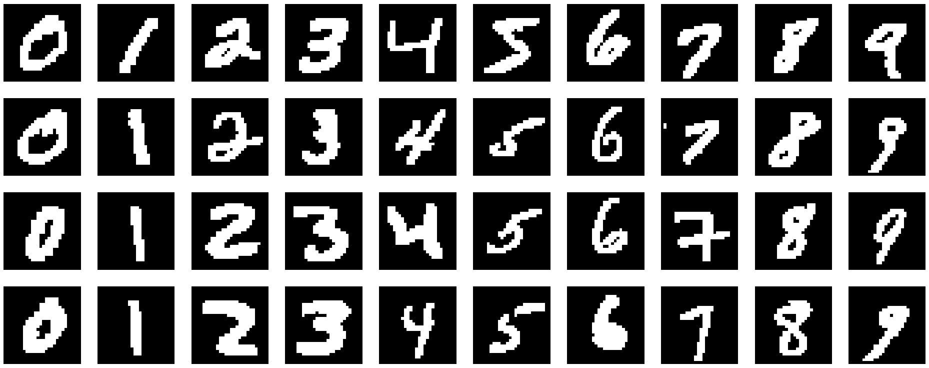

Figure 6: **Sample images in the MNIST database.**

sity parameter. For example, the third colorful figure depicted on the second row corresponds to recognizing a digit is "1" or "7" with the sparsity $k = 30$. In particular, for each pair, we label the small digit as positive and the large one as negative, and the blue and red pixels are the weights with positive and negative values respectively. Apparently, the models we learned are discriminative.

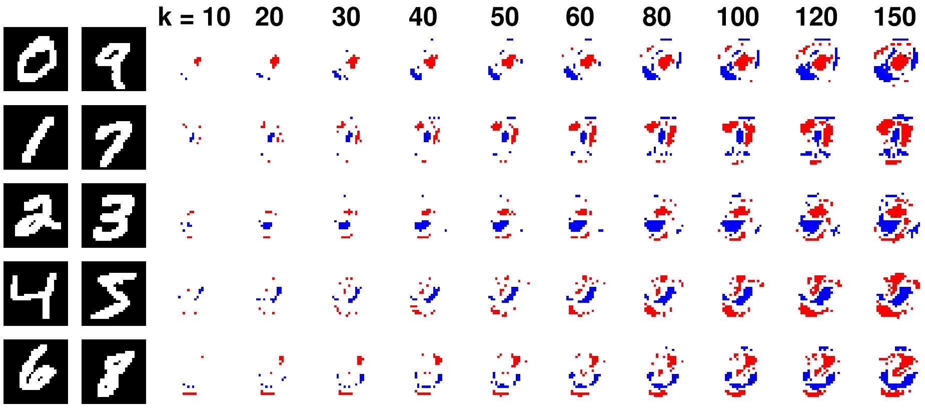

Figure 7: **Visualization of the models.** We visualize 5 models learned by HT-SVRG under different choices of sparsity shown on the top of each column. Note that the feature dimension is 784. From the top row to the bottom row, we illustrate the models of "0 vs 9", "1 vs 7", "2 vs 3", "4 vs 5" and "6 vs 8", where for each pair, we label the small digit as positive and the large one as negative. The red color represents negative weights while the blue pixels correspond with positive weights.

We also quantitatively show the convergence and prediction accuracy curves in Figure 8. Note that here, the y-axis is the objective value $F(\widetilde{\boldsymbol{x}}^s)$ rather than $\log(F(\widetilde{\boldsymbol{x}}^s) - F(\boldsymbol{x}_{\mathrm{opt}}))$, due to the fact that computing the exact optimum of (4.10) is NP-hard. Generally speaking, HT-SVRG converges quite fast and usually attains the minimum of objective value within 20 stages. It is not surprising to see that choosing a large quantity for the sparsity leads to a better (lower) objective value. However, in practice a small assignment for the sparsity, e.g., $k = 70$ facilitates an efficient computation while still suffices to ensure fast convergence and accurate prediction.

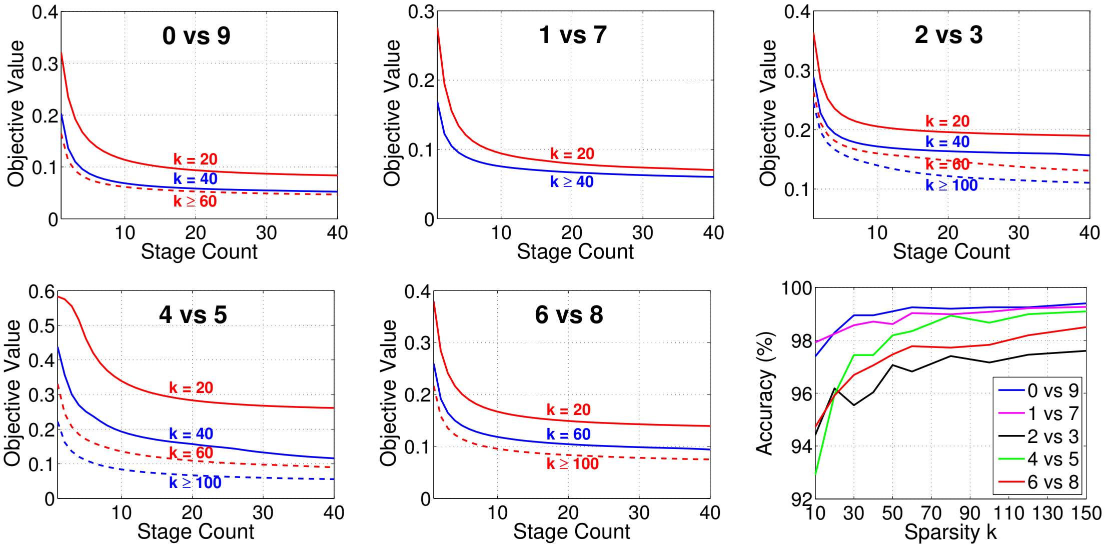

Figure 8: **Quantitative results on convergence and accuracy.** The first 5 figures demonstrate the convergence behavior of HT-SVRG for each binary classification task, where curves with different colors represent the objective value against number of stages under different sparsity k. Generally speaking, HT-SVRG converges within 20 stages which is a very fast rate. The last figure reflects the classification accuracy against the sparsity for all 5 classification tasks, where we find that for a moderate choice, e.g., $k = 70$, it already guarantees an accurate prediction (we recall the dimension is 784).

6. Conclusion and Open Problems

In this paper, we have provided a tight bound on the deviation resulting from the hard thresholding operator, which underlies a vast volume of algorithms developed for sparsity-constrained problems. Our derived bound is universal over all choices of parameters and we have proved that it cannot be improved without further information on the signals. We have discussed the implications of our result to the community of compressed sensing and machine learning, and have demonstrated that the theoretical results of a number of popular algorithms in the literature can be advanced. In addition, we have devised a novel algorithm which tackles the problem of sparse learning in large-scale setting. We have elaborated that our algorithm is guaranteed to produce global optimal solution for prevalent statistical models only when it is equipped with the tight bound, hence justifying that the conventional bound is not applicable in the challenging scenario.

There are several interesting open problems. The first question to ask is whether one can establish sharp RIP condition or sharp phase transition for hard thresholding based algorithms such as IHT and CoSaMP with the tight bound. Moreover, compared to the hard thresholded SGD method (Nguyen et al., 2014), HT-SVRG admits a vanishing optimization error. This poses a question of whether we are able to provably show the necessity of variance reduction for such a sparsity-constrained problem.

Acknowledgments

We would like to thank Jing Wang for insightful discussion since the early stage of the work. We also thank Martin Slawski for helpful discussion on the statistical precision of the problem, and thank Jian Wang for bringing the paper Nguyen et al. (2014) into our attention. We appreciate Huan Xu's high level comments on the work. Finally, we thank the anonymous reviewers for a careful check on our proof and for the encouraging comments. The work was partially funded by NSF-Bigdata-1419210 and NSF-III-1360971.

Appendix A. Technical Lemmas

We present some useful lemmas that will be invoked by subsequent analysis. The following is a characterization of the co-coercivity of the objective function $F(\boldsymbol{x})$. A similar result was obtained in Nguyen et al. (2014) but we present a refined analysis which is essential for our purpose.

Lemma 19 *For a given support set Ω, assume that the continuous function $F(\boldsymbol{x})$ is $L_{|\Omega|}$-RSS and is α_K-RSC for some sparsity level K. Then, for all vectors $\boldsymbol{x}$ and $\boldsymbol{x}'$ with $|\mathrm{supp}\,(\boldsymbol{x}-\boldsymbol{x}')\cup\Omega| \leq K$,*

$$\left\|\nabla_\Omega F(\boldsymbol{x}') - \nabla_\Omega F(\boldsymbol{x})\right\|_2^2 \leq 2L_{|\Omega|}\big(F(\boldsymbol{x}') - F(\boldsymbol{x}) - \left\langle \nabla F(\boldsymbol{x}), \boldsymbol{x}'-\boldsymbol{x}\right\rangle\big).$$

Proof We define an auxiliary function

$$G(\boldsymbol{w}) := F(\boldsymbol{w}) - \langle \nabla F(\boldsymbol{x}), \boldsymbol{w}\rangle .$$

For all vectors $\boldsymbol{w}$ and $\boldsymbol{w}'$, we have

$$\left\|\nabla G(\boldsymbol{w}) - \nabla G(\boldsymbol{w}')\right\|_2 = \left\|\nabla F(\boldsymbol{w}) - \nabla F(\boldsymbol{w}')\right\|_2 \leq L_{|\mathrm{supp}(\boldsymbol{w}-\boldsymbol{w}')|}\left\|\boldsymbol{w}-\boldsymbol{w}'\right\|_2,$$

which is equivalent to

$$G(\boldsymbol{w}) - G(\boldsymbol{w}') - \left\langle \nabla G(\boldsymbol{w}'), \boldsymbol{w}-\boldsymbol{w}'\right\rangle \leq \frac{L_r}{2}\left\|\boldsymbol{w}-\boldsymbol{w}'\right\|_2^2, \tag{A.1}$$

where $r := |\mathrm{supp}\,(\boldsymbol{w}-\boldsymbol{w}')|$. On the other hand, due to the RSC property of $F(\boldsymbol{x})$, we obtain

$$G(\boldsymbol{w}) - G(\boldsymbol{x}) = F(\boldsymbol{w}) - F(\boldsymbol{x}) - \langle \nabla F(\boldsymbol{x}), \boldsymbol{w}-\boldsymbol{x}\rangle \geq \frac{\alpha_{|\mathrm{supp}(\boldsymbol{w}-\boldsymbol{x})|}}{2}\left\|\boldsymbol{w}-\boldsymbol{x}\right\|_2^2 \geq 0,$$

provided that $|\mathrm{supp}\,(\boldsymbol{w}-\boldsymbol{x})| \leq K$. For the given support set Ω, we pick $\boldsymbol{w} = \boldsymbol{x}' - \frac{1}{L_{|\Omega|}}\nabla_\Omega G(\boldsymbol{x}')$. Clearly, for such a choice of $\boldsymbol{w}$, we have $\mathrm{supp}\,(\boldsymbol{w}-\boldsymbol{x}) = \mathrm{supp}\,(\boldsymbol{x}-\boldsymbol{x}')\cup\Omega$. Hence, by assuming that $|\mathrm{supp}\,(\boldsymbol{x}-\boldsymbol{x}')\cup\Omega|$ is not larger than K, we get

$$\begin{aligned}G(\boldsymbol{x}) &\leq G\left(\boldsymbol{x}' - \frac{1}{L_{|\Omega|}}\nabla_\Omega G(\boldsymbol{x}')\right)\\ &\leq G(\boldsymbol{x}') + \left\langle \nabla G(\boldsymbol{x}'), -\frac{1}{L_{|\Omega|}}\nabla_\Omega G(\boldsymbol{x}')\right\rangle + \frac{1}{2L_{|\Omega|}}\left\|\nabla_\Omega G(\boldsymbol{x}')\right\|_2^2\\ &= G(\boldsymbol{x}') - \frac{1}{2L_{|\Omega|}}\left\|\nabla_\Omega G(\boldsymbol{x}')\right\|_2^2,\end{aligned}$$

where the second inequality follows from (A.1). Now expanding $\nabla_\Omega G(\boldsymbol{x}')$ and rearranging the terms gives the desired result. ■

Lemma 20 *Consider the HT-SVRG algorithm for a fixed stage s. Let $\widehat{\boldsymbol{x}}$ be the target sparse vector. Let Ω be a support set such that $\mathrm{supp}\left(\boldsymbol{x}^{t-1}\right) \cup \mathrm{supp}\left(\widetilde{\boldsymbol{x}}\right) \cup \mathrm{supp}\left(\widehat{\boldsymbol{x}}\right) \subseteq \Omega$. Put $r = |\Omega|$. Assume $(A2)$. For all $1 \leq t \leq m$, denote $\boldsymbol{v}^t = \nabla f_{i_t}(\boldsymbol{x}^{t-1}) - \nabla f_{i_t}(\widetilde{\boldsymbol{x}}) + \widetilde{\boldsymbol{\mu}}$. Then we have the following:*

$$\begin{aligned}\mathbb{E}_{i_t|\boldsymbol{x}^{t-1}}\left[\left\|\mathcal{P}_\Omega\left(\boldsymbol{v}^t\right)\right\|_2^2\right] \leq 4L_r\left[F(\boldsymbol{x}^{t-1}) - F(\widehat{\boldsymbol{x}})\right] + 4L_r\left[F(\widetilde{\boldsymbol{x}}) - F(\widehat{\boldsymbol{x}})\right]\\ - 4L_r\left\langle\nabla F(\widehat{\boldsymbol{x}}), \boldsymbol{x}^{t-1} + \widetilde{\boldsymbol{x}} - 2\widehat{\boldsymbol{x}}\right\rangle + 4\left\|\mathcal{P}_\Omega\left(\nabla F(\widehat{\boldsymbol{x}})\right)\right\|_2^2.\end{aligned}$$

Proof We have

$$\begin{aligned}\left\|\mathcal{P}_\Omega\left(\boldsymbol{v}^t\right)\right\|_2^2 &= \left\|\mathcal{P}_\Omega\left(\nabla f_{i_t}(\boldsymbol{x}^{t-1}) - \nabla f_{i_t}(\widetilde{\boldsymbol{x}}) + \widetilde{\boldsymbol{\mu}}\right)\right\|_2^2\\ &\leq 2\left\|\mathcal{P}_\Omega\left(\nabla f_{i_t}(\boldsymbol{x}^{t-1}) - \nabla f_{i_t}(\widehat{\boldsymbol{x}})\right)\right\|_2^2 + 2\left\|\mathcal{P}_\Omega\left(\nabla f_{i_t}(\widetilde{\boldsymbol{x}}) - \nabla f_{i_t}(\widehat{\boldsymbol{x}}) - \widetilde{\boldsymbol{\mu}}\right)\right\|_2^2\\ &= 2\left\|\mathcal{P}_\Omega\left(\nabla f_{i_t}(\boldsymbol{x}^{t-1}) - \nabla f_{i_t}(\widehat{\boldsymbol{x}})\right)\right\|_2^2 + 2\left\|\mathcal{P}_\Omega\left(\nabla f_{i_t}(\widetilde{\boldsymbol{x}}) - \nabla f_{i_t}(\widehat{\boldsymbol{x}})\right)\right\|_2^2\\ &\quad + 2\left\|\mathcal{P}_\Omega\left(\widetilde{\boldsymbol{\mu}}\right)\right\|_2^2 - 4\left\langle\mathcal{P}_\Omega\left(\nabla f_{i_t}(\widetilde{\boldsymbol{x}}) - \nabla f_{i_t}(\widehat{\boldsymbol{x}})\right), \mathcal{P}_\Omega\left(\widetilde{\boldsymbol{\mu}}\right)\right\rangle\\ &\overset{\xi_1}{=} 2\left\|\mathcal{P}_\Omega\left(\nabla f_{i_t}(\boldsymbol{x}^{t-1}) - \nabla f_{i_t}(\widehat{\boldsymbol{x}})\right)\right\|_2^2 + 2\left\|\mathcal{P}_\Omega\left(\nabla f_{i_t}(\widetilde{\boldsymbol{x}}) - \nabla f_{i_t}(\widehat{\boldsymbol{x}})\right)\right\|_2^2\\ &\quad + 2\left\|\mathcal{P}_\Omega\left(\widetilde{\boldsymbol{\mu}}\right)\right\|_2^2 - 4\left\langle\nabla f_{i_t}(\widetilde{\boldsymbol{x}}) - \nabla f_{i_t}(\widehat{\boldsymbol{x}}), \mathcal{P}_\Omega\left(\widetilde{\boldsymbol{\mu}}\right)\right\rangle\\ &\overset{\xi_2}{\leq} 4L_r\left[f_{i_t}(\boldsymbol{x}^{t-1}) - f_{i_t}(\widehat{\boldsymbol{x}}) - \left\langle\nabla f_{i_t}(\widehat{\boldsymbol{x}}), \boldsymbol{x}^{t-1} - \widehat{\boldsymbol{x}}\right\rangle\right]\\ &\quad + 4L_r\left[f_{i_t}(\widetilde{\boldsymbol{x}}) - f_{i_t}(\widehat{\boldsymbol{x}}) - \left\langle\nabla f_{i_t}(\widehat{\boldsymbol{x}}), \widetilde{\boldsymbol{x}} - \widehat{\boldsymbol{x}}\right\rangle\right]\\ &\quad + 2\left\|\mathcal{P}_\Omega\left(\widetilde{\boldsymbol{\mu}}\right)\right\|_2^2 - 4\left\langle\nabla f_{i_t}(\widetilde{\boldsymbol{x}}) - \nabla f_{i_t}(\widehat{\boldsymbol{x}}), \mathcal{P}_\Omega\left(\widetilde{\boldsymbol{\mu}}\right)\right\rangle,\end{aligned}$$

where ξ_1 is by algebra, ξ_2 applies Lemma 19 and the fact that $|\Omega| = r$.

Taking the conditional expectation, we obtain the following:

$$\begin{aligned}&\mathbb{E}_{i_t|\boldsymbol{x}^{t-1}}\left[\left\|\mathcal{P}_\Omega\left(\boldsymbol{v}^t\right)\right\|_2^2\right]\\ &\leq 4L_r\left[F(\boldsymbol{x}^{t-1}) - F(\widehat{\boldsymbol{x}})\right] + 4L_r\left[F(\widetilde{\boldsymbol{x}}) - F(\widehat{\boldsymbol{x}})\right]\\ &\quad - 4L_r\left\langle\nabla F(\widehat{\boldsymbol{x}}), \boldsymbol{x}^{t-1} + \widetilde{\boldsymbol{x}} - 2\widehat{\boldsymbol{x}}\right\rangle + 2\left\langle 2\mathcal{P}_\Omega\left(\nabla F(\widehat{\boldsymbol{x}})\right) - \mathcal{P}_\Omega\left(\widetilde{\boldsymbol{\mu}}\right), \mathcal{P}_\Omega\left(\widetilde{\boldsymbol{\mu}}\right)\right\rangle\\ &= 4L_r\left[F(\boldsymbol{x}^{t-1}) - F(\widehat{\boldsymbol{x}})\right] + 4L_r\left[F(\widetilde{\boldsymbol{x}}) - F(\widehat{\boldsymbol{x}})\right]\\ &\quad - 4L_r\left\langle\nabla F(\widehat{\boldsymbol{x}}), \boldsymbol{x}^{t-1} + \widetilde{\boldsymbol{x}} - 2\widehat{\boldsymbol{x}}\right\rangle + \left\|2\mathcal{P}_\Omega\left(\nabla F(\widehat{\boldsymbol{x}})\right)\right\|_2^2\\ &\quad - \left\|2\mathcal{P}_\Omega\left(\nabla F(\widehat{\boldsymbol{x}})\right) - \mathcal{P}_\Omega\left(\widetilde{\boldsymbol{\mu}}\right)\right\|_2^2 - \left\|\mathcal{P}_\Omega\left(\widetilde{\boldsymbol{\mu}}\right)\right\|_2^2\\ &\leq 4L_r\left[F(\boldsymbol{x}^{t-1}) - F(\widehat{\boldsymbol{x}})\right] + 4L_r\left[F(\widetilde{\boldsymbol{x}}) - F(\widehat{\boldsymbol{x}})\right]\\ &\quad - 4L_r\left\langle\nabla F(\widehat{\boldsymbol{x}}), \boldsymbol{x}^{t-1} + \widetilde{\boldsymbol{x}} - 2\widehat{\boldsymbol{x}}\right\rangle + 4\left\|\mathcal{P}_\Omega\left(\nabla F(\widehat{\boldsymbol{x}})\right)\right\|_2^2.\end{aligned}$$

The proof is complete. ■

Corollary 21 *Assume the same conditions as in Lemma 20. If $\nabla F(\widehat{\boldsymbol{x}}) = 0$, we have*

$$\mathbb{E}_{i_t|\boldsymbol{x}^{t-1}}\left[\left\|\mathcal{P}_\Omega\left(\boldsymbol{v}^t\right)\right\|_2^2\right] \leq 4L_r\left[F(\boldsymbol{x}^{t-1}) + F(\widetilde{\boldsymbol{x}}) - 2F(\widehat{\boldsymbol{x}})\right].$$

Appendix B. Proofs for Section 2

B.1 Proof of Theorem 1

Proof The result is true for the trivial case that $\boldsymbol{b}$ is a zero vector. In the following, we assume that $\boldsymbol{b}$ is not a zero vector. Denote

$$\boldsymbol{w} := \mathcal{H}_k\left(\boldsymbol{b}\right).$$

Let Ω be the support set of $\boldsymbol{w}$ and let $\overline{\Omega}$ be its complement. We immediately have $\mathcal{P}_\Omega\left(\boldsymbol{b}\right) = \boldsymbol{w}$.

Let Ω' be the support set of $\boldsymbol{x}$. For the sake of simplicity, let us split the vector $\boldsymbol{b}$ as follows:

$$\begin{aligned}\boldsymbol{b}_1 = \mathcal{P}_{\Omega\backslash\Omega'}\left(\boldsymbol{b}\right), \quad \boldsymbol{b}_2 = \mathcal{P}_{\Omega\cap\Omega'}\left(\boldsymbol{b}\right),\\ \boldsymbol{b}_3 = \mathcal{P}_{\overline{\Omega}\backslash\Omega'}\left(\boldsymbol{b}\right), \quad \boldsymbol{b}_4 = \mathcal{P}_{\overline{\Omega}\cap\Omega'}\left(\boldsymbol{b}\right).\end{aligned}$$

Likewise, we denote

$$\begin{aligned}&\boldsymbol{w}_1 = \mathcal{P}_{\Omega\backslash\Omega'}\left(\boldsymbol{w}\right), \quad \boldsymbol{w}_2 = \mathcal{P}_{\Omega\cap\Omega'}\left(\boldsymbol{w}\right), \quad \boldsymbol{w}_3 = \mathcal{P}_{\overline{\Omega}\backslash\Omega'}(\boldsymbol{w}) = \boldsymbol{0}, \quad \boldsymbol{w}_4 = \mathcal{P}_{\overline{\Omega}\cap\Omega'}\left(\boldsymbol{w}\right) = \boldsymbol{0},\\ &\boldsymbol{x}_1 = \mathcal{P}_{\Omega\backslash\Omega'}\left(\boldsymbol{x}\right) = \boldsymbol{0}, \quad \boldsymbol{x}_2 = \mathcal{P}_{\Omega\cap\Omega'}\left(\boldsymbol{x}\right), \quad \boldsymbol{x}_3 = \mathcal{P}_{\overline{\Omega}\backslash\Omega'}(\boldsymbol{x}) = \boldsymbol{0}, \quad \boldsymbol{x}_4 = \mathcal{P}_{\overline{\Omega}\cap\Omega'}\left(\boldsymbol{x}\right).\end{aligned}$$

Due to the hard thresholding, we have

$$\boldsymbol{w}_1 = \boldsymbol{b}_1, \quad \boldsymbol{w}_2 = \boldsymbol{b}_2.$$

In this way, by simple algebra we have

$$\begin{aligned}\left\|\boldsymbol{w} - \boldsymbol{x}\right\|_2^2 &= \left\|\boldsymbol{b}_1\right\|_2^2 + \left\|\boldsymbol{b}_2 - \boldsymbol{x}_2\right\|_2^2 + \left\|\boldsymbol{x}_4\right\|_2^2,\\ \left\|\boldsymbol{b} - \boldsymbol{x}\right\|_2^2 &= \left\|\boldsymbol{b}_1\right\|_2^2 + \left\|\boldsymbol{b}_2 - \boldsymbol{x}_2\right\|_2^2 + \left\|\boldsymbol{b}_3\right\|_2^2 + \left\|\boldsymbol{b}_4 - \boldsymbol{x}_4\right\|_2^2.\end{aligned}$$

Our goal is to estimate the maximum of $\left\|\boldsymbol{w} - \boldsymbol{x}\right\|_2^2 / \left\|\boldsymbol{b} - \boldsymbol{x}\right\|_2^2$. It is easy to show that when attaining the maximum value, $\left\|\boldsymbol{b}_3\right\|_2$ must be zero since otherwise one may decrease this term to make the objective larger. Hence, maximizing $\left\|\boldsymbol{w} - \boldsymbol{x}\right\|_2^2 / \left\|\boldsymbol{b} - \boldsymbol{x}\right\|_2^2$ amounts to estimating the upper bound of the following over all choices of $\boldsymbol{x}$ and $\boldsymbol{b}$:

$$\gamma := \frac{\left\|\boldsymbol{b}_1\right\|_2^2 + \left\|\boldsymbol{b}_2 - \boldsymbol{x}_2\right\|_2^2 + \left\|\boldsymbol{x}_4\right\|_2^2}{\left\|\boldsymbol{b}_1\right\|_2^2 + \left\|\boldsymbol{b}_2 - \boldsymbol{x}_2\right\|_2^2 + \left\|\boldsymbol{b}_4 - \boldsymbol{x}_4\right\|_2^2}. \tag{B.1}$$

Firstly, we consider the case of $\left\|\boldsymbol{b}_1\right\|_2 = 0$, which means $\Omega = \Omega'$ implying $\gamma = 1$. In the following, we consider $\left\|\boldsymbol{b}_1\right\|_2 \neq 0$. In particular, we consider $\gamma > 1$ since we are interested in the maximum value of γ.

Arranging (B.1) we obtain

$$(\gamma - 1)\left\|\boldsymbol{b}_2 - \boldsymbol{x}_2\right\|_2^2 + \gamma\left\|\boldsymbol{b}_4 - \boldsymbol{x}_4\right\|_2^2 - \left\|\boldsymbol{x}_4\right\|_2^2 + (\gamma - 1)\left\|\boldsymbol{b}_1\right\|_2^2 = 0. \tag{B.2}$$

Let us fix $\boldsymbol{b}$ and define the function

$$G(\boldsymbol{x}_2, \boldsymbol{x}_4) = (\gamma - 1)\left\|\boldsymbol{b}_2 - \boldsymbol{x}_2\right\|_2^2 + \gamma\left\|\boldsymbol{b}_4 - \boldsymbol{x}_4\right\|_2^2 - \left\|\boldsymbol{x}_4\right\|_2^2 + (\gamma - 1)\left\|\boldsymbol{b}_1\right\|_2^2.$$

Thus, (B.2) indicates that $G(\boldsymbol{x}_2, \boldsymbol{x}_4)$ can attain the objective value of zero. Note that $G(\boldsymbol{x}_2, \boldsymbol{x}_4)$ is a quadratic function and its gradient and Hessian matrix can be computed as follows:

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{x}_2} G(\boldsymbol{x}_2, \boldsymbol{x}_4) &= 2(\gamma - 1)(\boldsymbol{x}_2 - \boldsymbol{b}_2),\\ \frac{\partial}{\partial \boldsymbol{x}_4} G(\boldsymbol{x}_2, \boldsymbol{x}_4) &= 2\gamma(\boldsymbol{x}_4 - \boldsymbol{b}_4) - 2\boldsymbol{x}_4,\\ \nabla^2 G(\boldsymbol{x}_2, \boldsymbol{x}_4) &= 2(\gamma - 1)\boldsymbol{I},\end{aligned}$$

where $\boldsymbol{I}$ is the identity matrix. Since the Hessian matrix is positive definite, $G(\boldsymbol{x}_2, \boldsymbol{x}_4)$ attains the global minimum at the stationary point, which is given by

$$\boldsymbol{x}_2^* = \boldsymbol{b}_2, \quad \boldsymbol{x}_4^* = \frac{\gamma}{\gamma - 1}\boldsymbol{b}_4,$$

resulting in the minimum objective value

$$G(\boldsymbol{x}_2^*, \boldsymbol{x}_4^*) = \frac{\gamma}{1-\gamma} \left\|\boldsymbol{b}_4\right\|_2^2 + (\gamma - 1) \left\|\boldsymbol{b}_1\right\|_2^2.$$

In order to guarantee the feasible set of (B.2) is non-empty, we require that

$$G(\boldsymbol{x}_2^*, \boldsymbol{x}_4^*) \leq 0,$$

implying

$$\left\|\boldsymbol{b}_1\right\|_2^2 \gamma^2 - (2\left\|\boldsymbol{b}_1\right\|_2^2 + \left\|\boldsymbol{b}_4\right\|_2^2)\gamma + \left\|\boldsymbol{b}_1\right\|_2^2 \leq 0.$$

Solving the above inequality with respect to γ, we obtain

$$\gamma \leq 1 + \frac{\left\|\boldsymbol{b}_4\right\|_2^2 + \sqrt{\left(4\left\|\boldsymbol{b}_1\right\|_2^2 + \left\|\boldsymbol{b}_4\right\|_2^2\right)\left\|\boldsymbol{b}_4\right\|_2^2}}{2\left\|\boldsymbol{b}_1\right\|_2^2}. \tag{B.3}$$

To derive an upper bound that is uniform over the choice of $\boldsymbol{b}$, we recall that $\boldsymbol{b}_1$ contains the largest absolute elements of $\boldsymbol{b}$ while $\boldsymbol{b}_4$ has smaller values. In particular, the averaged value of $\boldsymbol{b}_4$ is no greater than that of $\boldsymbol{b}_1$ in magnitude, i.e.,

$$\frac{\left\|\boldsymbol{b}_4\right\|_2^2}{\left\|\boldsymbol{b}_4\right\|_0} \leq \frac{\left\|\boldsymbol{b}_1\right\|_2^2}{\left\|\boldsymbol{b}_1\right\|_0}.$$

Note that $\left\|\boldsymbol{b}_1\right\|_0 = k - \left\|\boldsymbol{b}_2\right\|_0 = k - (K - \left\|\boldsymbol{b}_4\right\|_0)$. Hence, combining with the fact that $0 \leq \left\|\boldsymbol{b}_4\right\|_0 \leq \min\{K, d-k\}$ and optimizing over $\left\|\boldsymbol{b}_4\right\|_0$ gives

$$\left\|\boldsymbol{b}_4\right\|_2^2 \leq \frac{\min\{K, d-k\}}{k - K + \min\{K, d-k\}} \left\|\boldsymbol{b}_1\right\|_2^2.$$

Plugging back to (B.3), we finally obtain

$$\gamma \leq 1 + \frac{\rho + \sqrt{(4+\rho)\rho}}{2}, \quad \rho = \frac{\min\{K, d-k\}}{k - K + \min\{K, d-k\}}.$$

The proof is complete. ■

Appendix C. Proofs for Section 3

C.1 Proof of Theorem 6

We follow the proof pipeline of Blumensath and Davies (2009) and only remark the difference of our proof and theirs, i.e., where Theorem 1 applies. In case of possible confusion due to notation, we follow the symbols in Blumensath and Davies. One may refer to that article for a complete proof.

The first difference occurs in Eq. (22) of Blumensath and Davies (2009), where they reached

$$\text{(Old)} \quad \left\| \boldsymbol{x}^s - \boldsymbol{x}^{[n+1]} \right\|_2 \leq 2 \left\| \boldsymbol{x}^s_{B^{n+1}} - \boldsymbol{a}^{[n+1]}_{B^{n+1}} \right\|_2,$$

while Theorem 1 gives

$$\text{(New)} \quad \left\| \boldsymbol{x}^s - \boldsymbol{x}^{[n+1]} \right\|_2 \leq \sqrt{\nu} \left\| \boldsymbol{x}^s_{B^{n+1}} - \boldsymbol{a}^{[n+1]}_{B^{n+1}} \right\|_2.$$

Combining this new inequality and Eq. (23) therein, we obtain

$$\left\| \boldsymbol{x}^s - \boldsymbol{x}^{[n+1]} \right\|_2 \leq \sqrt{\nu} \left\| (\boldsymbol{I} - \boldsymbol{\Phi}^\top_{B^{n+1}} \boldsymbol{\Phi}_{B^{n+1}}) \boldsymbol{r}^{[n]}_{B^{n+1}} \right\|_2 + \sqrt{\nu} \left\| (\boldsymbol{\Phi}^\top_{B^{n+1}} \boldsymbol{\Phi}_{B^{n+1} \backslash B^{n+1}}) \boldsymbol{r}^{[n]}_{B^{n+1} \backslash B^{n+1}} \right\|_2.$$

By noting the fact that $\left| B^n \cup B^{n+1} \right| \leq 2s + s^*$ where s^* denotes the sparsity of the global optimum and following their reasoning of Eq. (24) and (25), we have a new bound for Eq. (26):

$$\text{(New)} \quad \left\| \boldsymbol{r}^{[n+1]} \right\|_2 \leq \sqrt{2\nu} \delta_{2s+s^*} \left\| \boldsymbol{r}^{[n]} \right\|_2 + \sqrt{(1 + \delta_{s+s^*})\nu} \left\| \boldsymbol{e} \right\|_2.$$

Now our result follows by setting the coefficient of $\left\| \boldsymbol{r}^{[n]} \right\|_2$ to 0.5. Note that specifying $\nu = 4$ gives the result of Blumensath and Davies (2009).

C.2 Proof of Theorem 7

We follow the proof technique of Theorem 6.27 in Foucart and Rauhut (2013) which gives the best known RIP condition for the CoSaMP algorithm to date. Since most of the reasoning is similar, we only point out the difference of our proof and theirs, i.e., where Theorem 1 applies. In case of confusion by notation, we follow the symbols used in Foucart and Rauhut (2013). The reader may refer to that book for a complete proof.

The first difference is in Eq. (6.49) of Foucart and Rauhut (2013). Note that to derive this inequality, Foucart and Rauhut invoked the conventional bound (1.1), which gives

$$\text{(Old)} \quad \left\| \boldsymbol{x}_S - \boldsymbol{x}^{n+1} \right\|_2^2 \leq \left\| (\boldsymbol{x}_S - \boldsymbol{u}^{n+1})_{\overline{U^{n+1}}} \right\|_2^2 + 4 \left\| (\boldsymbol{x}_S - \boldsymbol{u}^{n+1})_{U^{n+1}} \right\|_2^2,$$

while utilizing Theorem 1 gives

$$\text{(New)} \quad \left\| \boldsymbol{x}_S - \boldsymbol{x}^{n+1} \right\|_2^2 \leq \left\| (\boldsymbol{x}_S - \boldsymbol{u}^{n+1})_{\overline{U^{n+1}}} \right\|_2^2 + \nu \left\| (\boldsymbol{x}_S - \boldsymbol{u}^{n+1})_{U^{n+1}} \right\|_2^2.$$

Combining this new inequality with Eq. (6.50) and Eq. (6.51) therein, we obtain

$$\begin{aligned}\left\|\boldsymbol{x}_S-\boldsymbol{x}^{n+1}\right\|_2 \leq{}& \sqrt{2}\delta_{3s+s^*}\sqrt{\frac{1+(\nu-1)\delta_{3s+s^*}^2}{1-\delta_{3s+s^*}^2}}\left\|\boldsymbol{x}^n-\boldsymbol{x}_S\right\|_2\\ &+\sqrt{2}\delta_{3s+s^*}\sqrt{\frac{1+(\nu-1)\delta_{3s+s^*}^2}{1-\delta_{3s+s^*}^2}}\left\|(\boldsymbol{A}^*\boldsymbol{e}')_{(S\cup S^n)\Delta T^{n+1}}\right\|_2\\ &+\frac{2}{1-\delta_{3s+s^*}}\left\|(\boldsymbol{A}^*\boldsymbol{e}')_{U^{n+1}}\right\|_2,\end{aligned}$$

where s^* denotes the sparsity of the optimum. Our new bound follows by setting the coefficient of $\left\|\boldsymbol{x}^n-\boldsymbol{x}_S\right\|_2$ to 0.5 and solving the resultant equation. Note that setting $\nu=4$ gives the old bound of Foucart and Rauhut.

Appendix D. Proofs for Section 4

D.1 Proof of Theorem 10

Proof Fix a stage s. Let us denote

$$\boldsymbol{v}^t=\nabla f_{i_t}(\boldsymbol{x}^{t-1})-\nabla f_{i_t}(\widetilde{\boldsymbol{x}})+\widetilde{\boldsymbol{\mu}},$$

so that

$$\boldsymbol{b}^t=\boldsymbol{x}^{t-1}-\eta\boldsymbol{v}^t.$$

By specifying $\Omega=\operatorname{supp}\left(\boldsymbol{x}^{t-1}\right)\cup\operatorname{supp}\left(\boldsymbol{x}^t\right)\cup\operatorname{supp}\left(\widetilde{\boldsymbol{x}}\right)\cup\operatorname{supp}\left(\widehat{\boldsymbol{x}}\right)$, it follows that

$$\boldsymbol{r}^t=\mathcal{H}_k\left(\boldsymbol{b}^t\right)=\mathcal{H}_k\left(\mathcal{P}_\Omega\left(\boldsymbol{b}^t\right)\right).$$

Thus, the Euclidean distance of $\boldsymbol{x}^t$ and $\widehat{\boldsymbol{x}}$ can be bounded as follows:

$$\left\|\boldsymbol{x}^t-\widehat{\boldsymbol{x}}\right\|_2^2\leq\left\|\boldsymbol{r}^t-\widehat{\boldsymbol{x}}\right\|_2^2=\left\|\mathcal{H}_k\left(\mathcal{P}_\Omega\left(\boldsymbol{b}^t\right)\right)-\widehat{\boldsymbol{x}}\right\|_2^2\leq\nu\left\|\mathcal{P}_\Omega\left(\boldsymbol{b}^t\right)-\widehat{\boldsymbol{x}}\right\|_2^2, \tag{D.1}$$

where the first inequality holds because $\boldsymbol{x}^t=\Pi_\omega(\boldsymbol{r}^t)$ and $\left\|\widehat{\boldsymbol{x}}\right\|_2\leq\omega$. We also have

$$\begin{aligned}\left\|\mathcal{P}_\Omega\left(\boldsymbol{b}^t\right)-\widehat{\boldsymbol{x}}\right\|_2^2&=\left\|\boldsymbol{x}^{t-1}-\widehat{\boldsymbol{x}}-\eta\mathcal{P}_\Omega\left(\boldsymbol{v}^t\right)\right\|_2^2\\&=\left\|\boldsymbol{x}^{t-1}-\widehat{\boldsymbol{x}}\right\|_2^2+\eta^2\left\|\mathcal{P}_\Omega\left(\boldsymbol{v}^t\right)\right\|_2^2-2\eta\left\langle\boldsymbol{x}^{t-1}-\widehat{\boldsymbol{x}},\boldsymbol{v}^t\right\rangle,\end{aligned}$$

where the second equality uses the fact that $\left\langle\boldsymbol{x}^{t-1}-\widehat{\boldsymbol{x}},\mathcal{P}_\Omega\left(\boldsymbol{v}^t\right)\right\rangle=\left\langle\boldsymbol{x}^{t-1}-\widehat{\boldsymbol{x}},\boldsymbol{v}^t\right\rangle$. The first term will be preserved for mathematical induction. The third term is easy to manipulate thanks to the unbiasedness of $\boldsymbol{v}^t$. For the second term, we use Lemma 20 to bound it. Put them together, conditioning on $\boldsymbol{x}^{t-1}$ and taking the expectation over i_t for (D.1), we have

$$
\begin{aligned}
&\mathbb{E}_{i_t|\boldsymbol{x}^{t-1}}\Big[\left\|\boldsymbol{x}^t-\widehat{\boldsymbol{x}}\right\|_2^2\Big]\\
&\overset{\xi_1}{\leq} \nu\left\|\boldsymbol{x}^{t-1}-\widehat{\boldsymbol{x}}\right\|_2^2+4\nu\eta^2L\left[F(\boldsymbol{x}^{t-1})-F(\widehat{\boldsymbol{x}})+F(\widetilde{\boldsymbol{x}})-F(\widehat{\boldsymbol{x}})\right]-2\nu\eta\left\langle\boldsymbol{x}^{t-1}-\widehat{\boldsymbol{x}},\nabla F(\boldsymbol{x}^{t-1})\right\rangle\\
&\quad-4\nu\eta^2L\left\langle\nabla F(\widehat{\boldsymbol{x}}),\boldsymbol{x}^{t-1}+\widetilde{\boldsymbol{x}}-2\widehat{\boldsymbol{x}}\right\rangle+4\nu\eta^2\left\|\mathcal{P}_\Omega\left(\nabla F(\widehat{\boldsymbol{x}})\right)\right\|_2^2\\
&\overset{\xi_2}{\leq} \nu(1-\eta\alpha)\left\|\boldsymbol{x}^{t-1}-\widehat{\boldsymbol{x}}\right\|_2^2-2\nu\eta(1-2\eta L)\left[F(\boldsymbol{x}^{t-1})-F(\widehat{\boldsymbol{x}})\right]+4\nu\eta^2L\left[F(\widetilde{\boldsymbol{x}})-F(\widehat{\boldsymbol{x}})\right]\\
&\quad+4\nu\eta^2L\left\|\mathcal{P}_\Omega\left(\nabla F(\widehat{\boldsymbol{x}})\right)\right\|_2\cdot\left\|\boldsymbol{x}^{t-1}+\widetilde{\boldsymbol{x}}-2\widehat{\boldsymbol{x}}\right\|_2+4\nu\eta^2\left\|\mathcal{P}_\Omega\left(\nabla F(\widehat{\boldsymbol{x}})\right)\right\|_2^2\\
&\leq \nu(1-\eta\alpha)\left\|\boldsymbol{x}^{t-1}-\widehat{\boldsymbol{x}}\right\|_2^2-2\nu\eta(1-2\eta L)\left[F(\boldsymbol{x}^{t-1})-F(\widehat{\boldsymbol{x}})\right]\\
&\quad+4\nu\eta^2L\left[F(\widetilde{\boldsymbol{x}})-F(\widehat{\boldsymbol{x}})\right]+4\nu\eta^2Q'(4L\omega+Q')
\end{aligned}
$$

where ξ_1 applies Lemma 20, ξ_2 applies Assumption $(A1)$ and we write $Q':=\left\|\nabla_{3k+K}F(\widehat{\boldsymbol{x}})\right\|_2$ for brevity.

Now summing over the inequalities over $t=1,2,\cdots,m$, conditioning on $\widetilde{\boldsymbol{x}}$ and taking the expectation with respect to $\mathcal{I}^s=\{i_1,i_2,\cdots,i_m\}$, we have

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{I}^s|\widetilde{\boldsymbol{x}}}\Big[\left\|\boldsymbol{x}^m-\widehat{\boldsymbol{x}}\right\|_2^2\Big]\\
&\leq\ \left[\nu(1-\eta\alpha)-1\right]\mathbb{E}_{\mathcal{I}^s|\widetilde{\boldsymbol{x}}}\sum_{t=1}^m\left\|\boldsymbol{x}^{t-1}-\widehat{\boldsymbol{x}}\right\|_2^2+\left\|\boldsymbol{x}^0-\widehat{\boldsymbol{x}}\right\|_2^2+4\nu\eta^2Q'(4L\omega+Q')m\\
&\quad-2\nu\eta(1-2\eta L)\mathbb{E}_{\mathcal{I}^s|\widetilde{\boldsymbol{x}}}\sum_{t=1}^m\left[F(\boldsymbol{x}^{t-1})-F(\widehat{\boldsymbol{x}})\right]+4\nu\eta^2Lm\left[F(\widetilde{\boldsymbol{x}})-F(\widehat{\boldsymbol{x}})\right]\\
&=\ \left[\nu(1-\eta\alpha)-1\right]m\mathbb{E}_{\mathcal{I}^s,j^s|\widetilde{\boldsymbol{x}}}\left\|\widetilde{\boldsymbol{x}}^s-\widehat{\boldsymbol{x}}\right\|_2^2+\left\|\widetilde{\boldsymbol{x}}-\widehat{\boldsymbol{x}}\right\|_2^2+4\nu\eta^2Q'(4L\omega+Q')m\\
&\quad-2\nu\eta(1-2\eta L)m\mathbb{E}_{\mathcal{I}^s,j^s|\widetilde{\boldsymbol{x}}}\left[F(\widetilde{\boldsymbol{x}}^s)-F(\widehat{\boldsymbol{x}})\right]+4\nu\eta^2Lm\left[F(\widetilde{\boldsymbol{x}})-F(\widehat{\boldsymbol{x}})\right]\\
&\leq\ \left[\nu(1-\eta\alpha)-1\right]m\mathbb{E}_{\mathcal{I}^s,j^s|\widetilde{\boldsymbol{x}}}\left\|\widetilde{\boldsymbol{x}}^s-\widehat{\boldsymbol{x}}\right\|_2^2+\left(\frac{2}{\alpha}+4\nu\eta^2Lm\right)\left[F(\widetilde{\boldsymbol{x}})-F(\widehat{\boldsymbol{x}})\right]\\
&\quad-2\nu\eta(1-2\eta L)m\mathbb{E}_{\mathcal{I}^s,j^s|\widetilde{\boldsymbol{x}}}\left[F(\widetilde{\boldsymbol{x}}^s)-F(\widehat{\boldsymbol{x}})\right]+4\nu\eta^2Q'(4L\omega+Q')m+2Q'\omega/\alpha, \qquad \text{(D.2)}
\end{aligned}
$$

where we recall that j^s is the randomly chosen index used to determine $\widetilde{\boldsymbol{x}}^s$ (see Algorithm 1). The last inequality holds due to the RSC condition and $\left\|\boldsymbol{x}^t\right\|_2\leq\omega$. For brevity, we write

$$
Q:=4\nu\eta^2Q'(4L\omega+Q')m+2Q'\omega/\alpha,\quad Q'=\left\|\nabla_{3k+K}F(\widehat{\boldsymbol{x}})\right\|_2.
$$

Based on (D.2), we discuss two cases to examine the convergence of the algorithm.

Case 1. $\nu(1-\eta\alpha)\leq 1$. This immediately results in

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{I}^s|\widetilde{\boldsymbol{x}}}\Big[\left\|\boldsymbol{x}^m-\widehat{\boldsymbol{x}}\right\|_2^2\Big]\\
&\leq\ \left(\frac{2}{\alpha}+4\nu\eta^2Lm\right)\left[F(\widetilde{\boldsymbol{x}})-F(\widehat{\boldsymbol{x}})\right]-2\nu\eta(1-2\eta L)m\,\mathbb{E}_{\mathcal{I}^s,j^s|\widetilde{\boldsymbol{x}}}\left[F(\widetilde{\boldsymbol{x}}^s)-F(\widehat{\boldsymbol{x}})\right]+Q,
\end{aligned}
$$

which implies

$$\nu\eta(1-2\eta L)m\mathbb{E}_{\mathcal{I}^s,j^s|\widetilde{\boldsymbol{x}}}\left[F(\widetilde{\boldsymbol{x}}^s)-F(\widehat{\boldsymbol{x}})\right] \leq \left(\frac{1}{\alpha}+2\nu\eta^2 Lm\right)\left[F(\widetilde{\boldsymbol{x}})-F(\widehat{\boldsymbol{x}})\right]+\frac{Q}{2}.$$

Pick η such that

$$1-2\eta L > 0, \tag{D.3}$$

we obtain

$$\mathbb{E}_{\mathcal{I}^s,j^s|\widetilde{\boldsymbol{x}}}\left[F(\widetilde{\boldsymbol{x}}^s)-F(\widehat{\boldsymbol{x}})\right] \leq \left(\frac{1}{\nu\eta\alpha(1-2\eta L)m}+\frac{2\eta L}{1-2\eta L}\right)\left[F(\widetilde{\boldsymbol{x}})-F(\widehat{\boldsymbol{x}})\right]+\frac{Q}{2\nu\eta\alpha(1-2\eta L)m}.$$

To guarantee the convergence, we must impose

$$\frac{2\eta L}{1-2\eta L} < 1. \tag{D.4}$$

Putting (D.3), (D.4) and $\nu(1-\eta\alpha) \leq 1$ together gives

$$\eta < \frac{1}{4L}, \quad \nu \leq \frac{1}{1-\eta\alpha}. \tag{D.5}$$

The convergence coefficient here is

$$\beta = \frac{1}{\nu\eta\alpha(1-2\eta L)m}+\frac{2\eta L}{1-2\eta L}. \tag{D.6}$$

Thus, we have

$$\mathbb{E}\left[F(\widetilde{\boldsymbol{x}}^s)-F(\widehat{\boldsymbol{x}})\right] \leq \beta^s\left[F(\widetilde{\boldsymbol{x}}^0)-F(\widehat{\boldsymbol{x}})\right]+\frac{Q}{2\nu\eta\alpha(1-2\eta L)(1-\beta)m},$$

where the expectation is taken over $\{\mathcal{I}^1, j^1, \mathcal{I}^2, j^2, \cdots, \mathcal{I}^s, j^s\}$.

Case 2. $\nu(1-\eta\alpha) > 1$. In this case, (D.2) implies

$$\begin{aligned}\mathbb{E}_{\mathcal{I}^s|\widetilde{\boldsymbol{x}}}\Big[\left\|\boldsymbol{x}^m-\widehat{\boldsymbol{x}}\right\|_2^2\Big] \leq\ & \left(\frac{2}{\alpha}+4\nu\eta^2 Lm\right)\left[F(\widetilde{\boldsymbol{x}})-F(\widehat{\boldsymbol{x}})\right]+Q\\ &+\left(\frac{2}{\alpha}\left[\nu(1-\eta\alpha)-1\right]m-2\nu\eta(1-2\eta L)m\right)\mathbb{E}_{\mathcal{I}^s,j^s|\widetilde{\boldsymbol{x}}}\left[F(\widetilde{\boldsymbol{x}}^s)-F(\widehat{\boldsymbol{x}})\right].\end{aligned}$$

Rearranging the terms gives

$$\left(2\nu\eta\alpha-2\nu\eta^2\alpha L-\nu+1\right)m\,\mathbb{E}_{\mathcal{I}^s,j^s|\widetilde{\boldsymbol{x}}}\left[F(\widetilde{\boldsymbol{x}}^s)-F(\widehat{\boldsymbol{x}})\right] \leq \left(1+2\nu\eta^2\alpha Lm\right)\left[F(\widetilde{\boldsymbol{x}})-F(\widehat{\boldsymbol{x}})\right]+\frac{\alpha Q}{2}.$$

To ensure the convergence, the minimum requirements are

$$\begin{aligned}&2\nu\eta\alpha-2\nu\eta^2\alpha L-\nu+1 > 0,\\ &2\nu\eta\alpha-2\nu\eta^2\alpha L-\nu+1 > 2\nu\eta^2\alpha L.\end{aligned}$$

That is,

$$4\nu\alpha L\eta^2 - 2\nu\alpha\eta + \nu - 1 < 0.$$

We need to guarantee the feasible set of the above inequality is non-empty for the positive variable η. Thus, we require

$$4\nu^2\alpha^2 - 4 \times 4\nu\alpha L(\nu - 1) > 0,$$

which is equivalent to

$$\nu < \frac{4L}{4L - \alpha}.$$

Combining it with $\nu(1 - \eta\alpha) > 1$ gives

$$\frac{1}{1 - \eta\alpha} < \nu < \frac{4L}{4L - \alpha}.$$

To ensure the above feasible set is non-empty, we impose

$$\frac{1}{1 - \eta\alpha} < \frac{4L}{4L - \alpha},$$

so that

$$0 < \eta < \frac{1}{4L}, \quad \frac{1}{1 - \eta\alpha} < \nu < \frac{4L}{4L - \alpha}. \tag{D.7}$$

The convergence coefficient for this case is

$$\beta = \frac{1}{(2\nu\eta\alpha - 2\nu\eta^2\alpha L - \nu + 1)\, m} + \frac{2\nu\eta^2\alpha L}{2\nu\eta\alpha - 2\nu\eta^2\alpha L - \nu + 1}. \tag{D.8}$$

Thus,

$$\mathbb{E}\left[F(\widetilde{\boldsymbol{x}}^s) - F(\widehat{\boldsymbol{x}})\right] \le \beta^s \left[F(\widetilde{\boldsymbol{x}}^0) - F(\widehat{\boldsymbol{x}})\right] + \frac{\alpha Q}{2(2\nu\eta\alpha - 2\nu\eta^2\alpha L - \nu + 1)(1 - \beta)m}.$$

By combining (D.5) and (D.7), the minimum requirement for η and ν is

$$0 < \eta < \frac{1}{4L}, \quad \nu < \frac{4L}{4L - \alpha}.$$

The proof is complete. ■

D.2 Proof of Corollary 16

Proof By noting the concavity of the square root function, we have

$$\begin{aligned}\mathbb{E}\left[\sqrt{\max\{F(\widetilde{\boldsymbol{x}}^s) - F(\widehat{\boldsymbol{x}}), 0\}}\right] &\le \sqrt{\mathbb{E}\left[\max\{F(\widetilde{\boldsymbol{x}}^s) - F(\widehat{\boldsymbol{x}}), 0\}\right]} \\ &\le \sqrt{(2/3)^s \max\{F(\widetilde{\boldsymbol{x}}^0) - F(\widehat{\boldsymbol{x}}), 0\} + \tau(\widehat{\boldsymbol{x}})}.\end{aligned}$$

Suppose that $F(\boldsymbol{x})$ satisfies RSS with parameter $L' \in [\alpha, L]$. It follows that

$$F(\widetilde{\boldsymbol{x}}^0) - F(\widehat{\boldsymbol{x}}) \leq \left\langle \nabla F(\widehat{\boldsymbol{x}}), \widetilde{\boldsymbol{x}}^0 - \widehat{\boldsymbol{x}} \right\rangle + \frac{L'}{2} \left\| \widetilde{\boldsymbol{x}}^0 - \widehat{\boldsymbol{x}} \right\|_2^2 \leq \frac{1}{2L'} \left\| \nabla_{k+K} F(\widehat{\boldsymbol{x}}) \right\|_2^2 + L' \left\| \widetilde{\boldsymbol{x}}^0 - \widehat{\boldsymbol{x}} \right\|_2^2.$$

Recall that

$$\tau(\widehat{\boldsymbol{x}}) = \frac{5\omega}{\alpha} \left\| \nabla_{3k+K} F(\widehat{\boldsymbol{x}}) \right\|_2 + \frac{1}{\alpha L} \left\| \nabla_{3k+K} F(\widehat{\boldsymbol{x}}) \right\|_2^2.$$

Hence using $\sqrt{a+b+c+d} \leq \sqrt{a} + \sqrt{b} + \sqrt{c} + \sqrt{d}$ gives

$$\begin{aligned} \mathbb{E}\left[\sqrt{\max\{F(\widetilde{\boldsymbol{x}}^s) - F(\widehat{\boldsymbol{x}}), 0\}} \right] \leq \sqrt{L'} \left(\frac{2}{3}\right)^{\frac{s}{2}} \left\| \widetilde{\boldsymbol{x}}^0 - \widehat{\boldsymbol{x}} \right\|_2 + \sqrt{\frac{5\omega}{\alpha} \left\| \nabla_{3k+K} F(\widehat{\boldsymbol{x}}) \right\|_2} \\ + \left(\frac{1}{\alpha} + \sqrt{\frac{1}{2\alpha}} \right) \left\| \nabla_{3k+K} F(\widehat{\boldsymbol{x}}) \right\|_2. \end{aligned}$$

Finally, the RSC property immediately suggests that (see, e.g., Lemma 20 in Shen and Li (2017b))

$$\begin{aligned} \mathbb{E}\left[\left\| \widetilde{\boldsymbol{x}}^s - \widehat{\boldsymbol{x}} \right\|_2 \right] &\leq \sqrt{\frac{2}{\alpha}} \mathbb{E}\left[\sqrt{\max\{F(\widetilde{\boldsymbol{x}}^s) - F(\widehat{\boldsymbol{x}}), 0\}} \right] + \frac{2 \left\| \nabla_{k+K} F(\widehat{\boldsymbol{x}}) \right\|_2}{\alpha} \\ &\leq \sqrt{\frac{2L'}{\alpha}} \cdot \left(\frac{2}{3}\right)^{\frac{s}{2}} \left\| \widetilde{\boldsymbol{x}}^0 - \widehat{\boldsymbol{x}} \right\|_2 + \sqrt{\frac{10\omega}{\alpha^2} \left\| \nabla_{3k+K} F(\widehat{\boldsymbol{x}}) \right\|_2} \\ &\quad + \left(\sqrt{\frac{2}{\alpha^3}} + \frac{3}{\alpha} \right) \left\| \nabla_{3k+K} F(\widehat{\boldsymbol{x}}) \right\|_2. \end{aligned}$$

The proof is complete. ■

Appendix E. HT-SAGA

We demonstrate that the hard thresholding step can be integrated into SAGA (Defazio et al., 2014) as shown in Algorithm 2. Note that the only difference of Algorithm 2 and the one proposed in Defazio et al. (2014) is that we perform hard thresholding rather than proximal operator. Hence, our algorithm guarantees k-sparse solution.

Theorem 22 *Assume the same conditions as in Defazio et al. (2014). Further assume the optimum of* (4.1) *without the sparsity constraint happens to be k-sparse. Then, the sequence of the solutions produced by Algorithm 2 converges to the optimum with geometric rate for some properly chosen sparsity parameter k.*

Proof Define the Lyapunov function Z as follows:

$$Z^t := Z(\boldsymbol{x}^t, \{\boldsymbol{\phi}_i^t\}) = \frac{1}{n} \sum_{i=1}^{n} f_i(\boldsymbol{\phi}_i^t) - F(\widehat{\boldsymbol{x}}) - \frac{1}{n} \sum_{i=1}^{n} \left\langle \nabla f_i(\widehat{\boldsymbol{x}}), \boldsymbol{\phi}_i^t - \widehat{\boldsymbol{x}} \right\rangle + c \left\| \boldsymbol{x}^t - \widehat{\boldsymbol{x}} \right\|_2^2.$$

Algorithm 2 SAGA with Hard Thresholding (HT-SAGA)

```
Require: The current iterate $\boldsymbol{x}^t$ and of each $\nabla f_i(\boldsymbol{\phi}_i^t)$ at the end of iteration $t$, the step size $\eta$.
Ensure: The new iterate.
Pick $j \in \{1, 2, \cdots, n\}$ uniformly at random.
Take $\boldsymbol{\phi}_j^{t+1} = \boldsymbol{x}^t$ and store $\nabla f_j(\boldsymbol{\phi}_j^{t+1})$ in the table. All other entries in the table remain unchanged.
Update the new iterate $\boldsymbol{x}^{t+1}$ as follows:
    $\boldsymbol{b}^{t+1} = \boldsymbol{x}^t - \eta \left[ \nabla f_j(\boldsymbol{\phi}_j^{t+1}) - \nabla f_j(\boldsymbol{\phi}_j^{t}) + \frac{1}{n} \sum_{i=1}^n \nabla f_i(\boldsymbol{\phi}_i^t) \right],$
    $\boldsymbol{x}^{t+1} = \mathcal{H}_k \left( \boldsymbol{b}^{t+1} \right).$
```

We examine Z^{t+1}. We have

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{n} \sum_i f_i(\boldsymbol{\phi}_i^{t+1}) \right] &= \frac{1}{n} F(\boldsymbol{x}^t) + \left(1 - \frac{1}{n}\right) \frac{1}{n} \sum_i f_i(\boldsymbol{\phi}_i^t), \\
\mathbb{E}\left[-\frac{1}{n} \sum_i \left\langle \nabla f_i(\widehat{\boldsymbol{x}}), \boldsymbol{\phi}_i^{t+1} - \widehat{\boldsymbol{x}} \right\rangle \right] &= -\frac{1}{n} \left\langle \nabla F(\widehat{\boldsymbol{x}}), \boldsymbol{x}^t - \widehat{\boldsymbol{x}} \right\rangle \\
&\quad - \left(1 - \frac{1}{n}\right) \frac{1}{n} \sum_i \left\langle \nabla f_i(\widehat{\boldsymbol{x}}), \boldsymbol{\phi}_i^t - \widehat{\boldsymbol{x}} \right\rangle.
\end{aligned}$$

Also,

$$c \left\| \boldsymbol{x}^{t+1} - \widehat{\boldsymbol{x}} \right\|_2^2 \leq c\nu \left\| \boldsymbol{b}^{t+1} - \widehat{\boldsymbol{x}} \right\|_2^2 = c\nu \left\| \boldsymbol{b}^{t+1} - \widehat{\boldsymbol{x}} + \eta \nabla F(\widehat{\boldsymbol{x}}) \right\|_2^2.$$

For the first term, we have

$$\begin{aligned}
& c\nu \, \mathbb{E} \left\| \boldsymbol{b}^{t+1} - \widehat{\boldsymbol{x}} + \eta \nabla F(\widehat{\boldsymbol{x}}) \right\|_2^2 \\
\leq\, & c\nu(1 - \eta\alpha) \left\| \boldsymbol{x}^t - \widehat{\boldsymbol{x}} \right\|_2^2 + c\nu \left((1+\mu)\eta^2 - \frac{\eta}{L} \right) \mathbb{E} \left\| \nabla f_j(\boldsymbol{x}^t) - \nabla f_j(\widehat{\boldsymbol{x}}) \right\|_2^2 \\
& - \frac{2c\nu\eta(L-\alpha)}{L} \left[F(\boldsymbol{x}^t) - F(\widehat{\boldsymbol{x}}) - \left\langle \nabla F(\widehat{\boldsymbol{x}}), \boldsymbol{x}^t - \widehat{\boldsymbol{x}} \right\rangle \right] - c\nu\eta^2\mu \left\| \nabla F(\boldsymbol{x}^t) - \nabla F(\widehat{\boldsymbol{x}}) \right\|_2^2 \\
& + 2c\nu(1+\mu^{-1})\eta^2 L \left[\frac{1}{n} \sum_i f_i(\boldsymbol{\phi}_i^t) - F(\widehat{\boldsymbol{x}}) - \frac{1}{n} \sum_i \left\langle \nabla f_i(\widehat{\boldsymbol{x}}), \boldsymbol{\phi}_i^t - \widehat{\boldsymbol{x}} \right\rangle \right].
\end{aligned}$$

Therefore,

$$\begin{aligned}
& \mathbb{E}[Z^{t+1}] - Z^t \\
\leq\, & -\frac{1}{\kappa} Z^t + \left(\frac{1}{n} - \frac{2c\nu\eta(L-\alpha)}{L} - 2c\nu\eta^2\alpha\mu \right) \left[F(\boldsymbol{x}^t) - F(\widehat{\boldsymbol{x}}) - \left\langle \nabla F(\widehat{\boldsymbol{x}}), \boldsymbol{x}^t - \widehat{\boldsymbol{x}} \right\rangle \right] \\
& + \left(\frac{1}{\kappa} + 2c\nu(1+\mu^{-1})\eta^2 L - \frac{1}{n} \right) \left[\frac{1}{n} \sum_i f_i(\boldsymbol{\phi}_i^t) - F(\widehat{\boldsymbol{x}}) - \frac{1}{n} \sum_i \left\langle \nabla f_i(\widehat{\boldsymbol{x}}), \boldsymbol{\phi}_i^t - \widehat{\boldsymbol{x}} \right\rangle \right] \\
& + \left(\frac{c}{\kappa} - c\nu\eta\alpha \right) \left\| \boldsymbol{x}^t - \widehat{\boldsymbol{x}} \right\|_2^2 + \left((1+\mu)\eta - \frac{1}{L} \right) c\nu\eta \, \mathbb{E} \left\| \nabla f_j(\boldsymbol{x}^t) - \nabla f_j(\widehat{\boldsymbol{x}}) \right\|_2^2.
\end{aligned}$$

In order to guarantee the convergence, we choose proper values for η, c, κ, μ and ν such that the terms in round brackets are non-positive. That is, we require

$$\begin{aligned} \frac{c}{\kappa} - c\nu\eta\alpha &\leq 0, \\ (1+\mu)\eta - \frac{1}{L} &\leq 0, \\ \frac{1}{n} - \frac{2c\nu\eta(L-\alpha)}{L} - 2c\nu\eta^2\alpha\mu &\leq 0, \\ \frac{1}{\kappa} + 2c\nu(1+\mu^{-1})\eta^2 L - \frac{1}{n} &\leq 0. \end{aligned}$$

Pick

$$\begin{aligned} \eta &= \frac{1}{2(\alpha n + L)}, \\ \mu &= \frac{2\alpha n + L}{L}, \\ \kappa &= \frac{1}{\nu\eta\alpha}, \end{aligned}$$

we fulfill the first two inequalities. Pick

$$c = \frac{1}{2\eta(1-\eta\alpha)n}.$$

Then by the last two equalities, we require

$$1 - \eta\alpha \leq \nu \leq \frac{(1-\eta\alpha)L}{\eta\alpha(1-\eta\alpha)Ln + 1}.$$

On the other hand, by Theorem 1, we have

$$\nu > 1.$$

Thus, we require

$$1 < \nu \leq \frac{(1-\eta\alpha)L}{\eta\alpha(1-\eta\alpha)Ln + 1},$$

By algebra, the above inequalities has non-empty feasible set provided that

$$(6\alpha^2 - 8\alpha^2 L)n^2 + (14\alpha L - \alpha - 16\alpha L^2)n + 8L^2(1-L) < 0.$$

Due to $\alpha \leq L$, we know

$$n \geq \frac{14L + \sqrt{224L^3 + 1}}{2\alpha(8L-6)}$$

suffices where we assume $L > 3/4$. Picking

$$\nu = \frac{(1-\eta\alpha)L}{\eta\alpha(1-\eta\alpha)Ln + 1}$$

completes the proof. ■

References

Radosaw Adamczak, Alexander E. Litvak, Alain Pajor, and Nicole Tomczak-Jaegermann. Restricted isometry property of matrices with independent columns and neighborly polytopes by random sampling. *Constructive Approximation*, 34(1):61–88, 2011.

Alekh Agarwal, Sahand Negahban, and Martin J. Wainwright. Fast global convergence of gradient methods for high-dimensional statistical recovery. *The Annals of Statistics*, 40(5):2452–2482, 2012.

Bubacarr Bah and Jared Tanner. Improved bounds on restricted isometry constants for gaussian matrices. *SIAM Journal on Matrix Analysis Applications*, 31(5):2882–2898, 2010.

Bubacarr Bah and Jared Tanner. Bounds of restricted isometry constants in extreme asymptotics: Formulae for Gaussian matrices. *Linear Algebra and its Applications*, 441:88–109, 2014.

Sohail Bahmani, Bhiksha Raj, and Petros T. Boufounos. Greedy sparsity-constrained optimization. *Journal of Machine Learning Research*, 14(1):807–841, 2013.

Richard Baraniuk, Mark Davenport, Ronald DeVore, and Michael B. Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 28(3):253–263, 2008.

Pierre C. Bellec, Guillaume Lecué, and Alexandre B. Tsybakov. Slope meets lasso: improved oracle bounds and optimality. *CoRR*, abs/1605.08651, 2016.

Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, pages 1705–1732, 2009.

Jeffrey D. Blanchard and Jared Tanner. Performance comparisons of greedy algorithms in compressed sensing. *Numerical Linear Algebra with Applications*, 22(2):254–282, 2015.

Thomas Blumensath and Mike E. Davies. Iterative thresholding for sparse approximations. *Journal of Fourier Analysis and Applications*, 14(5-6):629–654, 2008.

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

Tony T. Cai and Anru Zhang. Sharp RIP bound for sparse signal and low-rank matrix recovery. *Applied and Computational Harmonic Analysis*, 35(1):74–93, 2013.

Tony T. Cai, Lie Wang, and Guangwu Xu. New bounds for restricted isometry constants. *IEEE Transactions on Information Theory*, 56(9):4388–4394, 2010.

Emmanuel J. Candès. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, 346(9):589–592, 2008.

Emmanuel J. Candès and Terence Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

Emmanuel J. Candès and Terence Tao. The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.

Emmanuel J. Candès and Michael B. Wakin. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, 25(2):21–30, 2008.

Emmanuel J. Candès, Justin K. Romberg, and Terence Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006.

Venkat Chandrasekaran, Benjamin Recht, Pablo A. Parrilo, and Alan S. Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012.

Scott Shaobing Chen, David L. Donoho, and Michael A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1998.

Wei Dai and Olgica Milenkovic. Subspace pursuit for compressive sensing signal reconstruction. *IEEE Transactions on Information Theory*, 55(5):2230–2249, 2009.

Ingrid Daubechies, Michel Defrise, and Christine De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics*, 57(11):1413–1457, 2004.

Aaron Defazio, Francis R. Bach, and Simon Lacoste-Julien. SAGA: a fast incremental gradient method with support for non-strongly convex composite objectives. In *Proceedings of the 28th Annual Conference on Neural Information Processing Systems*, pages 1646–1654, 2014.

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

David L. Donoho and Jared Tanner. Precise undersampling theorems. *Proceedings of the IEEE*, 98 (6):913–924, 2010.

David L. Donoho and Yaakov Tsaig. Fast solution of ℓ_1-norm minimization problems when the solution may be sparse. *IEEE Transactions on Information Theory*, 54(11):4789–4812, 2008.

David L. Donoho, Michael Elad, and Vladimir N. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on Information Theory*, 52(1): 6–18, 2006.

David L. Donoho, Iain Johnstone, and Andrea Montanari. Accurate prediction of phase transitions in compressed sensing via a connection to minimax denoising. *IEEE Transactions on Information Theory*, 59(6):3396–3433, 2013.

John C. Duchi and Yoram Singer. Efficient online and batch learning using forward backward splitting. *Journal of Machine Learning Research*, 10:2899–2934, 2009.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

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

Simon Foucart. Sparse recovery algorithms: Sufficient conditions in terms of restricted isometry constants. In *Approximation Theory XIII: San Antonio 2010*, pages 65–77. Springer, New York, NY, 2012.

Simon Foucart and Holger Rauhut. *A Mathematical Introduction to Compressive Sensing*. Applied and Numerical Harmonic Analysis. Birkhäuser, 2013.

Prateek Jain, Ambuj Tewari, and Inderjit S. Dhillon. Orthogonal matching pursuit with replacement. In *Proceedings of the 25th Annual Conference on Neural Information Processing Systems*, pages 1215–1223, 2011.

Prateek Jain, Ambuj Tewari, and Purushottam Kar. On iterative hard thresholding methods for high-dimensional M-estimation. In *Proceedings of the 28th Annual Conference on Neural Information Processing Systems*, pages 685–693, 2014.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Proceedings of the 27th Annual Conference on Neural Information Processing Systems*, pages 315–323, 2013.

John Langford, Lihong Li, and Tong Zhang. Sparse online learning via truncated gradient. *Journal of Machine Learning Research*, 10:777–801, 2009.

Ping Li, Cun-Hui Zhang, and Tong Zhang. Compressed counting meets compressed sensing. In *Proceedings of The 27th Conference on Learning Theory*, pages 1058–1077, 2014.

Xingguo Li, Tuo Zhao, Raman Arora, Han Liu, and Jarvis Haupt. Stochastic variance reduced optimization for nonconvex sparse learning. *CoRR*, abs/1605.02711, 2016.

Po-Ling Loh and Martin J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. *The Annals of Statistics*, 40(3):1637–1664, 2012.

Po-Ling Loh and Martin J. Wainwright. Regularized m-estimators with nonconvexity: statistical and algorithmic theory for local optima. *Journal of Machine Learning Research*, 16:559–616, 2015.

Zongming Ma. Sparse principal component analysis and iterative thresholding. *The Annals of Statistics*, 41(2):772–801, 2013.

Qun Mo. A sharp restricted isometry constant bound of orthogonal matching pursuit. *CoRR*, abs/1501.01708, 2015.

Qun Mo and Yi Shen. A remark on the restricted isometry property in orthogonal matching pursuit. *IEEE Transactions on Information Theory*, 58(6):3654–3656, 2012.

Deanna Needell and Joel A. Tropp. CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis*, 26(3):301–321, 2009.

Deanna Needell and Roman Vershynin. Signal recovery from incomplete and inaccurate measurements via regularized orthogonal matching pursuit. *IEEE Journal of Selected Topics in Signal Processing*, 4(2):310–316, 2010.

Sahand Negahban, Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. In *Proceedings of the 23rd Annual Conference on Neural Information Processing Systems*, pages 1348–1356, 2009.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87 of *Applied Optimization*. Springer US, 2004.

Nam H. Nguyen, Deanna Needell, and Tina Woolf. Linear convergence of stochastic iterative greedy algorithms with sparse constraints. *CoRR*, abs/1407.0088, 2014.

Bruno A. Olshausen and David J. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision research*, 37(23):3311–3325, 1997.

Art Owen and Yi Zhou. Safe and effective importance sampling. *Journal of the American Statistical Association*, 95(449):135–143, 2000.

Yagyensh C. Pati, Ramin Rezaiifar, and Perinkulam S. Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Conference Record of The 27th Asilomar Conference on Signals, Systems and Computers*, pages 40–44, 1993.

Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions on Information Theory*, 57(10): 6976–6994, 2011.

Nicolas Le Roux, Mark W. Schmidt, and Francis R. Bach. A stochastic gradient method with an exponential convergence rate for finite training sets. In *Proceedings of the 26th Annual Conference on Neural Information Processing Systems*, pages 2672–2680, 2012.

Jie Shen and Ping Li. A tight bound of hard thresholding. *CoRR*, abs/1605.01656, 2016.

Jie Shen and Ping Li. On the iteration complexity of support recovery via hard thresholding pursuit. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3115–3124, 2017a.

Jie Shen and Ping Li. Partial hard thresholding: Towards A principled analysis of support recovery. In *Proceedings of the 31st Annual Conference on Neural Information Processing Systems*, pages 3127–3137, 2017b.

Robert Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Joel A. Tropp. Greed is good: algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242, 2004.

Joel A. Tropp and Anna C. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on Information Theory*, 53(12):4655–4666, 2007.

Joel A. Tropp and Stephen J. Wright. Computational methods for sparse solution of linear inverse problems. *Proceedings of the IEEE*, 98(6):948–958, 2010.

Martin J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory*, 55(5): 2183–2202, 2009.

Jian Wang and Byonghyo Shim. On the recovery limit of sparse signals using orthogonal matching pursuit. *IEEE Transactions on Signal Processing*, 60(9):4973–4976, 2012.

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

Lin Xiao and Tong Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.

Xiao-Tong Yuan and Tong Zhang. Truncated power method for sparse eigenvalue problems. *Journal of Machine Learning Research*, 14(1):899–925, 2013.

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

Tong Zhang. Sparse recovery with orthogonal matching pursuit under RIP. *IEEE Transactions on Information Theory*, 57(9):6215–6221, 2011.