# Efficient Stochastic Gradient Hard Thresholding

**Pan Zhou**[*] **Xiaotong Yuan**[†] **Jiashi Feng**[*]
[*] Learning & Vision Lab, National University of Singapore, Singapore
[†] B-DAT Lab, Nanjing University of Information Science & Technology, Nanjing, China
pzhou@u.nus.edu xtyuan@nuist.edu.cn elefjia@nus.edu.sg

## Abstract

Stochastic gradient hard thresholding methods have recently been shown to work favorably in solving large-scale empirical risk minimization problems under sparsity or rank constraint. Despite the improved iteration complexity over full gradient methods, the gradient evaluation and hard thresholding complexity of the existing stochastic algorithms usually scales linearly with data size, which could still be expensive when data is huge and the hard thresholding step could be as expensive as singular value decomposition in rank-constrained problems. To address these deficiencies, we propose an efficient hybrid stochastic gradient hard thresholding (HSG-HT) method that can be provably shown to have sample-size-independent gradient evaluation and hard thresholding complexity bounds. Specifically, we prove that the stochastic gradient evaluation complexity of HSG-HT scales linearly with inverse of sub-optimality and its hard thresholding complexity scales logarithmically. By applying the heavy ball acceleration technique, we further propose an accelerated variant of HSG-HT which can be shown to have improved factor dependence on restricted condition number. Numerical results confirm our theoretical affirmation and demonstrate the computational efficiency of the proposed methods.

## 1 Introduction

We consider the following sparsity- or rank-constrained *finite-sum* minimization problems which are widely applied in high-dimensional statistical estimation:

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}) := \frac{1}{n}\sum\nolimits_{i=1}^{n} f_i(\boldsymbol{x}), \quad \text{s.t. } \|\boldsymbol{x}\|_0 \le k \quad \text{or} \quad \mathsf{rank}\,(\boldsymbol{x}) \le k, \tag{1}$$

where each individual loss $f_i(\boldsymbol{x})$ is associated with the $i$-th sample, $\|\boldsymbol{x}\|_0$ denotes the number of nonzero entries in $\boldsymbol{x}$ as a vector variable, $\mathsf{rank}\,(\boldsymbol{x})$ denotes the rank of $\boldsymbol{x}$ as a matrix variable, and $k$ represents the sparsity/low-rankness level. Such a formulation encapsulates several important problems, including $\ell_0$-constrained linear/logistic regression [1, 2, 3], sparse graphical model learning [4], and low-rank multivariate and multi-task regression [5, 6], to name a few.

We are particularly interested in gradient hard thresholding methods [7, 8, 9, 10] which are popular and effective for solving problem (1). The common theme of this class of methods is to iterate between gradient descent and hard thresholding to maintain sparsity/low-rankness of solution while minimizing the loss function. In our problem setting, a plain gradient hard thresholding iteration is given by $\boldsymbol{x}^{t+1} = \Phi_k(\boldsymbol{x}^t - \eta\nabla f(\boldsymbol{x}^t))$, where $\Phi_k(\cdot)$, as defined in Section 2, denotes the hard thresholding operation that preserves the top $k$ entries (in magnitude) of a vector or produces an optimal rank-$k$ approximation to a matrix via singular value decomposition (SVD). When considering gradient hard thresholding methods, two main sources of computational complexity are at play: the gradient evaluation complexity and the hard thresholding complexity. As the per-iteration hard thresholding can be as expensive as SVD in rank-constrained problems, our goal is to develop methods that iterate and converge quickly while using a minimal number of hard thresholding operations.

**Full gradient hard thresholding.** The plain form of full gradient hard thresholding (FG-HT) algorithm has been extensively studied in compressed sensing and sparse learning [7, 9, 10, 11]. At

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

Table 1: Comparison of different hard thresholding algorithms for sparsity- and rank-constrained problem (1). Both computational complexity and statistical error are evaluated w.r.t. the estimation error $\|\tilde{\boldsymbol{x}} - \boldsymbol{x}^*\|$ between the $k$-sparse/rank estimator $\tilde{\boldsymbol{x}}$ and the $k^*$-sparse/rank optimum $\boldsymbol{x}^*$. Both $\kappa_s$ and $\kappa_{\widehat{s}}$ denote the restricted condition numbers with $s = 2k + k^*$ and $\widehat{s} = 3k + k^*$. $\widetilde{\mathcal{I}} = \mathsf{supp}(\Phi_{2k}(\nabla f(\boldsymbol{x}^*)))\cup\mathsf{supp}(\boldsymbol{x}^*)$ and $\widehat{\mathcal{I}} = \mathsf{supp}(\Phi_{3k}(\nabla f(\boldsymbol{x}^*)))\cup\mathsf{supp}(\boldsymbol{x}^*)$ are two support sets.

| | Restriction on $\kappa_s$ | Required value of $k$ | Computational Complexity #IFO | Computational Complexity #Hard Thresholding | Statistical Error on Sparsity-constrained Problem[1] |
|---|---|---|---|---|---|
| FG-HT [9, 10] | — | $\Omega(\kappa_s^2 k^*)$ | $\mathcal{O}\big(n\kappa_s \log\big(\frac{1}{\epsilon}\big)\big)$ | $\mathcal{O}\big(\kappa_s \log\big(\frac{1}{\epsilon}\big)\big)$ | $\mathcal{O}\big(\sqrt{k+k^*}\|\nabla f(\boldsymbol{x}^*)\|_\infty\big)$ |
| SG-HT [17] | $\leq \frac{4}{3}$ | $\Omega(\kappa_s^2 k^*)$ | $\mathcal{O}\big(\kappa_s \log\big(\frac{1}{\epsilon}\big)\big)$ | $\mathcal{O}\big(\kappa_s \log\big(\frac{1}{\epsilon}\big)\big)$ | $\mathcal{O}\big(\frac{1}{n}\sum_{i=1}^n \|\nabla f_i(\boldsymbol{x}^*)\|_2\big)$ |
| SVRG-HT [18] | — | $\Omega(\kappa_s^2 k^*)$ | $\mathcal{O}\big((n+\kappa_s)\log\big(\frac{1}{\epsilon}\big)\big)$ | $\mathcal{O}\big((n+\kappa_s)\log\big(\frac{1}{\epsilon}\big)\big)$ | $\mathcal{O}\big(\sqrt{s}\|\nabla f(\boldsymbol{x}^*)\|_\infty + \|\nabla_{\widetilde{\mathcal{I}}} f(\boldsymbol{x}^*)\|_2\big)$ |
| HSG-HT | — | $\Omega(\kappa_s^2 k^*)$ | $\mathcal{O}\big(\frac{\kappa_s}{\epsilon}\big)$ | $\mathcal{O}\big(\kappa_s \log\big(\frac{1}{\epsilon}\big)\big)$ | $\mathcal{O}\big(\|\nabla_{\widetilde{\mathcal{I}}} f(\boldsymbol{x}^*)\|_2\big)$ |
| AHSG-HT | — | $\Omega(\kappa_{\widehat{s}} k^*)$ | $\mathcal{O}\big(\frac{\sqrt{\kappa_{\widehat{s}}}}{\epsilon}\big)$ | $\mathcal{O}\big(\sqrt{\kappa_{\widehat{s}}} \log\big(\frac{1}{\epsilon}\big)\big)$ | $\mathcal{O}\big(\|\nabla_{\widehat{\mathcal{I}}} f(\boldsymbol{x}^*)\|_2\big)$ |

[1] For general rank-constrained problem, the statistic error is not explicitly provided in FG-HT, SG-HT and SVRG-HT while is given in our Theorem 1 for HSG-HT and Theorem 2 for AHSG-HT.

each iteration, FG-HT first updates the variable $\boldsymbol{x}$ by using full gradient descent and then performs hard thresholding on the updated variable. Theoretical results show that FG-HT converges linearly towards a proper nominal solution with high estimation accuracy [9, 10, 12]. Besides, compared with the algorithms adopting $\ell_1$- or nuclear-norm convex relaxation (*e.g.*, [13, 14, 15, 16]), directly solving problem (1) via FG-HT often exhibits similar accuracy guarantee but is more computationally efficient. However, despite these desirable properties, FG-HT needs to compute the full gradient at each iteration which can be expensive in large-scale problems. If the restricted condition number is $\kappa_s$, then $\mathcal{O}\big(\kappa_s \log\big(\frac{1}{\epsilon}\big)\big)$ iterations are needed to attain an $\epsilon$-suboptimal solution (up to a statistical error), and thus the sample-wise gradient evaluation complexity, or incremental first order oracle (IFO, see Definition 1), is $\mathcal{O}\big(n\kappa_s \log\big(\frac{1}{\epsilon}\big)\big)$ which scales linearly with $n\kappa_s$.

**Stochastic gradient hard thresholding.** To improve computational efficiency, stochastic hard thresholding algorithms [17, 18, 19] have recently been developed via leveraging the finite-sum structure of problem (1). For instance, Nguyen *et al.* [17] proposed a stochastic gradient hard thresholding (SG-HT) algorithm for solving problem (1). At each iteration, SG-HT only evaluates gradient of one (or a mini-batch) randomly selected sample for variable update and hard thresholding. It was shown that the IFO complexity and hard thresholding complexity of SG-HT are both $\mathcal{O}\big(\kappa_s \log\big(\frac{1}{\epsilon}\big)\big)$ which is independent on $n$. However, SG-HT can only be shown to converge to a sub-optimal statistical estimation accuracy (see Table 1) which is inferior to that of the full-gradient methods. Another limitation of SG-HT is that it requires the restricted condition number $\kappa_s$ to be not larger than $4/3$ which is hard to meet in real high-dimensional sparse estimation problems such as sparse linear regression [10]. To overcome these issues, the stochastic variance reduced gradient hard thresholding (SVRG-HT) algorithm [18, 19] is developed as an adaptation of SVRG [20] to problem (1). Benefiting from the variance reduced technique, SVRG-HT can converge more stably and efficiently while having better estimation accuracy than SG-HT. Also different from SG-HT, the convergence analysis for SVRG-HT allows arbitrary bounded restricted condition number. As shown in Table 1, both the IFO complexity and hard thresholding complexity of SVRG-HT are $\mathcal{O}\big((n+\kappa_s)\log\big(\frac{1}{\epsilon}\big)\big)$. Although the IFO complexity of SVRG-HT substantially improves over FG-HT, the overall complexity still scale linearly with respect to the sample size $n$. Therefore, when the data-scale is huge (e.g., $n \gg \kappa_s$) and the per-iteration hard thresholding operation is expensive, SVRG-HT could still be computationally inefficient in practice.

**Overview of our approach.** The method we propose can be viewed as a simple yet efficient extension of the hybrid stochastic gradient descent (HSGD) method [21, 22] from unconstrained finite-sum minimization to the cardinality-constrained finite-sum problem (1). The core idea of HSGD is to iteratively sample an evolving mini-batch of terms in the finite-sum for gradient estimation. This style of incremental gradient method has been shown, both in theory and practice, to bridge smoothly the gap between deterministic and stochastic gradient methods [22]. Inspired by the success of HSGD, we propose the hybrid stochastic gradient hard thresholding (HSG-HT) method which has the following variable update form:

$$\boldsymbol{x}^{t+1} = \Phi_k\left(\boldsymbol{x}^t - \eta \boldsymbol{g}^t\right), \text{with } \boldsymbol{g}^t = \frac{1}{s_t}\sum\nolimits_{i_t \in \mathcal{S}_t} \nabla f_{i_t}(\boldsymbol{x}^t),$$

where $\eta$ is the learning rate and $\mathcal{S}_t$ is the set of $s_t$ selected samples. In early stage of iterations, HSG-HT selects a few samples to estimate the full gradient; and along with more iterations, $s_t$ increases, giving more accurate full gradient estimation. Such a mechanism allows it to enjoy the merits of both SG-HT and FG-HT, *i.e.* the low iteration complexity of SG-HT and the steady convergence rate of FG-HT with constant learning rate $\eta$. Given a $k^*$-sparse/low-rank target solution $\boldsymbol{x}^*$, for objective function with restricted condition number $\kappa_s$ and $s = 2k + k^*$, we show that $\mathcal{O}\big(\frac{\kappa_s}{\epsilon}\big)$ rounds of IFO update and $\mathcal{O}\big(\kappa_s \log\big(\frac{1}{\epsilon}\big)\big)$ steps of hard thresholding operation are sufficient for HSG-HT to find $\tilde{\boldsymbol{x}}$ such that $\|\tilde{\boldsymbol{x}} - \boldsymbol{x}^*\|^2 \leq \epsilon + \mathcal{O}\big(\|\nabla_{\widetilde{\mathcal{I}}} f(\boldsymbol{x}^*)\|^2\big)$. In this way, HSG-HT exhibits sample-size-independent IFO and hard thresholding complexity. Another attractiveness of HSG-HT is that it can be accelerated via applying the heavy ball acceleration technique [23, 24, 25]. To this end, we modify the iteration of HSG-HT by adding a small momentum $\nu(\boldsymbol{x}^t - \boldsymbol{x}^{t-1})$ for some $\nu > 0$ to the gradient descent step:

$$\boldsymbol{x}^{t+1} = \Phi_k\left(\boldsymbol{x}^t - \eta \boldsymbol{g}^t + \nu(\boldsymbol{x}^t - \boldsymbol{x}^{t-1})\right).$$

We call the above modified version as accelerated HSG-HT (AHSG-HT). For twice differentiable functions, we prove that such a simple momentum strategy boosts the IFO complexity of HSG-HT to $\mathcal{O}\big(\frac{\sqrt{\kappa_{\widehat{s}}}}{\epsilon}\big)$, and the hard thresholding complexity to $\mathcal{O}\big(\sqrt{\kappa_{\widehat{s}}} \log\big(\frac{1}{\epsilon}\big)\big)$, where $\widehat{s} = 3k + k^*$. To the best of our knowledge, AHSG-HT is the first momentum based algorithm that can be provably shown to have such an improved complexity for stochastic gradient hard thresholding.

**Highlight of results and contribution.** Table 1 summarizes our main results on computational complexity and statistical estimation accuracy of HSG-HT and AHSG-HT, along with the results for the above mentioned state-of-the-art gradient hard thresholding algorithms. From this table we can observe that our methods have several theoretical advantages over the considered prior methods, which are highlighted in the next few paragraphs.

*On sparsity/low-rankness level constraint condition.* AHSG-HT substantially improves the bounding condition on the sparsity/low-rankness level $k$: it only requires $k = \Omega(\kappa_{\widehat{s}} k^*)$, while the other considered algorithms with optimal statistical estimation accuracy all require $k = \Omega(\kappa_s^2 k^*)$. Moreover, both HSG-HT and AHSG-HT get rid of the restrictive condition $\kappa_s \leq 4/3$ required in SG-HT.

*On statistical estimation accuracy.* For sparsity-constrained problem, the statistical estimation accuracy of HSG-HT is comparable to that in FG-HT and is better than those in SVRG and SG-HT, as $\|\nabla_{\widetilde{\mathcal{I}}} f(\boldsymbol{x}^*)\|_2$ in HSG-HT is usually superior over the error $\sqrt{s}\|\nabla f(\boldsymbol{x}^*)\|_\infty$ in SVRG-HT and is much smaller than the one $\frac{1}{n}\sum_{i=1}^n \|\nabla f_i(\boldsymbol{x}^*)\|_2$ in SG-HT. AHSG-HT has better estimation error than HSG-HT and could be superior over FG-HT since it allows smaller sparsity/low-rankness level $k = \Omega(\kappa_{\widehat{s}} k^*)$ and thus has a smaller cardinality of its support set $\widehat{\mathcal{I}}$, especially for the problems with large restrictive condition number.

*On computational complexity.* Both HSG-HT and AHSG-HT enjoy sample-size-independent IFO and hard thresholding complexity. To compare the IFO complexity, our methods will be cheaper than FG-HT and SVRG-HT when $n$ dominates $\frac{1}{\epsilon}$. This suggests that HSG-HT and AHSG-HT are more suitable for handling large-scale data. SG-HT has the lowest IFO complexity, which however is obtained at the price of severely deteriorated statistical estimation accuracy. In terms of hard thresholding complexity, AHSG-HT is the best one and HSG-HT matches FG-HT and SG-HT.

Last but not least, we highlight that AHSG-HT, to our best knowledge, for the first time provides improved convergence guarantees for momentum based stochastic gradient hard thresholding methods. While in convex problems the momentum based methods such as heavy ball and Nesterov's methods have long been known to work favorably for accelerating full/stochastic gradient methods [23, 26, 27, 28], it still remains largely unknown if it is possible to accelerate gradient hard thresholding methods for solving the non-convex finite-sum problem (1). There is a recent attempt at understanding a Nesterov's momentum full gradient hard thresholding method [29]. Although showing linear rate of convergence for that method, the results are obtained under seemingly very restrictive assumptions, and more frustratingly, the acceleration achieved by that method exhibits no better dependence on restricted condition number than the plain FG-HT. In sharp contrast, under mild conditions, AHSG-HT can be shown to have much improved dependence on condition number than HSG-HT.

## 2 Preliminaries

Throughout this paper, we use $\|\boldsymbol{x}\|$ to denote the Euclidean norm for vector $\boldsymbol{x} \in \mathbb{R}^d$ and the Frobenius norm for matrix $\boldsymbol{x} \in \mathbb{R}^{d_1 \times d_2}$. $\|\boldsymbol{x}\|_\infty$ denotes the largest absolute entry in $\boldsymbol{x}$. The hard thresholding

operation $\Phi_k(\boldsymbol{x})$ preserves the $k$ largest entries of $\boldsymbol{x}$ in magnitude for vector $\boldsymbol{x}$, and for matrix $\boldsymbol{x}$ it only preserves the top $k$-top singular values. Namely, $\Phi_k(\boldsymbol{x}) = \boldsymbol{H}_k\boldsymbol{\Sigma}_k\boldsymbol{V}_k^T$, where $\boldsymbol{H}_k$ and $\boldsymbol{V}_k$ are respectively the top-$k$ left and right singular vectors of $\boldsymbol{x}$, $\boldsymbol{\Sigma}_k$ is the diagonal matrix of the top-$k$ singular values of $\boldsymbol{x}$. We use $\mathsf{supp}(\boldsymbol{x})$ to denote the support set of $\boldsymbol{x}$. Specifically, for vector $\boldsymbol{x}$, $\mathsf{supp}(\boldsymbol{x})$ indexes its nonzero entries; and for matrix $\boldsymbol{x} \in \mathbb{R}^{d_1\times d_2}$, it indexes the subspace $\boldsymbol{U}$ that is a set of singular vectors spanning the column space of $\boldsymbol{x}$. For vector variable $\boldsymbol{x}$, $\nabla_{\mathcal{I}} f(\boldsymbol{x})$ preserves the entries in $\nabla f(\boldsymbol{x})$ indexed by the support set $\mathcal{I}$ and sets the remaining entries to be zero; while for matrix variable $\boldsymbol{x}$, $\nabla_{\mathcal{I}} f(\boldsymbol{x})$ with $\mathcal{I} = \mathcal{I}_1 \cup \mathcal{I}_2$ projects $\nabla f(\boldsymbol{x})$ into the subspace indexed by $\mathcal{I}_1 \cup \mathcal{I}_2$, namely $\nabla_{\mathcal{I}} f(\boldsymbol{x}) = (\boldsymbol{U}_1\boldsymbol{U}_1^T + \boldsymbol{U}_2\boldsymbol{U}_2^T - \boldsymbol{U}_1\boldsymbol{U}_1^T\boldsymbol{U}_2\boldsymbol{U}_2^T)\nabla f(\boldsymbol{x})$, where $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ respectively span the subspaces indexed by $\mathcal{I}_1$ and $\mathcal{I}_2$.

We assume the objective function in (1) to have restricted strong convexity (RSC) and restricted strong smoothness (RSS). For both sparsity- and rank-constrained problems, the RSC and RSS conditions are commonly used in analyzing hard thresholding algorithms [9, 10, 18, 19, 17].

**Assumption 1** (Restricted strong convexity condition, RSC)**.** *A differentiable function $f(\boldsymbol{x})$ is restricted $\rho_s$-strongly convex with parameter $s$ if there exists a generic constant $\rho_s > 0$ such that for any $\boldsymbol{x}$, $\boldsymbol{x}'$ with $\|\boldsymbol{x} - \boldsymbol{x}'\|_0 \leq s$ or $\mathsf{rank}(\boldsymbol{x} - \boldsymbol{x}') \leq s$, we have*

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

**Assumption 2** (Restricted strong smoothness condition, RSS)**.** *For each $f_i(\boldsymbol{x})$, it is said to be restricted $\ell_s$-strongly smooth with parameter $s$ if there exists a generic constant $\ell_s > 0$ such that for any $\boldsymbol{x}$, $\boldsymbol{x}'$ with $\|\boldsymbol{x} - \boldsymbol{x}'\|_0 \leq s$ or $\mathsf{rank}(\boldsymbol{x} - \boldsymbol{x}') \leq s$, we have*

$$f_i(\boldsymbol{x}) - f_i(\boldsymbol{x}') - \langle \nabla f_i(\boldsymbol{x}'), \boldsymbol{x} - \boldsymbol{x}' \rangle \leq \frac{\ell_s}{2}\|\boldsymbol{x} - \boldsymbol{x}'\|^2.$$

We also need to impose the following boundness assumption on the variance of stochastic gradient.

**Assumption 3** (Bounded stochastic gradient variance)**.** *For any $\boldsymbol{x}$ and each loss $f_i(\boldsymbol{x})$, the distance between $\nabla f_i(\boldsymbol{x})$ and the full gradient $\nabla f(\boldsymbol{x})$ is upper bounded as $\max_i \|\nabla f_i(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\| \leq B$.*

Similar to [18, 20, 30, 31], the incremental first order oracle (IFO) complexity is adopted as the computational complexity metric for solving finite-sum problem (1). In high-dimensional sparse learning and low-rank matrix recovery problems, the per-iteration hard thresholding operation can be equally time-consuming or even more expensive than gradient evaluation. For instance, in rank-constrained problems, hard thresholding operation can be as expensive as top-$k$ SVD for a matrix. Therefore we also need to take the computational complexity of hard thresholding into our account.

**Definition 1** (IFO and Hard Thresholding Complexity)**.** *For $f(\boldsymbol{x})$ in problem (1), an IFO takes an index $i \in [n]$ and a point $\boldsymbol{x}$, and returns the pair $(f_i(\boldsymbol{x}), \nabla f_i(\boldsymbol{x}))$. In a hard thresholding operation, we feed $\boldsymbol{x}$ into $\Phi_k(\cdot)$ and obtain the output $\Phi_k(\boldsymbol{x})$.*

The IFO and hard thresholding complexity as a whole can more comprehensively reflect the overall computational performance of a first-order hard thresholding algorithm, as objective value, gradient evaluation and hard thresholding operation usually dominate the per-iteration computation.

## 3 Hybrid Stochastic Gradient Hard Thresholding

In this section, we first introduce the Hybrid Stochastic Gradient Hard Thresholding (HSG-HT) algorithm and then analyze its convergence performance for sparsity- and rank-constrained problems.

### 3.1 The HSG-HT Algorithm

The HSG-HT algorithm is outlined in Algorithm 1. At the $t$-th iteration, it first uniformly randomly selects $s_t$ samples $\mathcal{S}_t$ from all data and evaluates the approximated gradient $\boldsymbol{g}^t = \frac{1}{s_t}\sum_{i_t\in\mathcal{S}_t} \nabla f_{i_t}(\boldsymbol{x}^t)$. Then, there are two options for variable update. The first option is to update $\boldsymbol{x}^{t+1}$ with a standard local descent step along $\boldsymbol{g}^t$ followed by a hard threholding step, giving the plain update procedure $\boldsymbol{x}^{t+1} = \Phi_k\left(\boldsymbol{x}^t - \eta\boldsymbol{g}^t\right)$ in option **O1**. The other option **O2** is to update $\boldsymbol{x}^{t+1}$ based on a momentum formulation $\boldsymbol{x}^{t+1} = \Phi_k\left(\boldsymbol{x}^t - \eta\boldsymbol{g}^t + \nu(\boldsymbol{x}^t - \boldsymbol{x}^{t-1})\right)$, leading to an accelerated variant of HSG-HT.

**Algorithm 1:** (Accelerated) Hybrid Stochastic Gradient Hard Thresholding

**Input** : Initial point $\boldsymbol{x}^0$, sample index set $\mathcal{S} = \{1, \cdots, n\}$, learning rate $\eta$, momentum strength $\nu$, mini-batch sizes $\{s_t\}$.

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

Uniformly randomly select $s_t$ samples $\mathcal{S}_t$ from $\mathcal{S}$

Compute the approximate gradient $\boldsymbol{g}^t = \frac{1}{s_t}\sum_{i_t \in \mathcal{S}_t} \nabla f_{i_t}(\boldsymbol{x}^t)$

Update $\boldsymbol{x}^{t+1}$ using either of the following two options:

**(O1)** $\boldsymbol{x}^{t+1} = \Phi_k\left(\boldsymbol{x}^t - \eta \boldsymbol{g}^t\right)$; `/* for plain HSG-HT */`

**(O2)** $\boldsymbol{x}^{t+1} = \Phi_k\left(\boldsymbol{x}^t - \eta \boldsymbol{g}^t + \nu(\boldsymbol{x}^t - \boldsymbol{x}^{t-1})\right)$; `/* for accelerated HSG-HT */`

**end**

**Output** : $\boldsymbol{x}^T$.

The plain update in **O1** can be viewed as a special case of the momentum based update in **O2** with strength $\nu = 0$. In early stage of iteration when the mini-batch size $s_t$ is relatively small, HSG-HT performs more like SG-HT with low per-iteration gradient evaluation cost. Along with more iterations, $s_t$ increases and HSG-HT performs like full gradient hard thresholding methods. Next, we analyze the parameter estimation performance and the objective convergence of HSG-HT. The analysis of the accelerated version will be presented in Section 4.

### 3.2 Statistical Estimation Analysis

We first analyze the parameter estimation performance of HSG-HT by characterizing the distance between the output of Algorithm 1 and the optimum $\boldsymbol{x}^*$. Such an analysis is helpful in understanding the convergence behavior and the statistical estimation error of the computed solution. We summarize the main result for both sparsity- and rank-constrained problems in Theorem 1.

**Theorem 1.** *Suppose the objective function $f(\boldsymbol{x})$ is $\rho_s$-strongly convex and each individual $f_i(\boldsymbol{x})$ is $\ell_s$-strongly smooth with parameter $s = 2k + k^*$. Let $\kappa_s = \frac{\ell_s}{\rho_s}$ and $\alpha = 1 + \frac{2\sqrt{k^*}}{\sqrt{k-k^*}}$. Assume the sparsity/low-rankness level $k \geq \left(1 + 712\kappa_s^2\right)k^*$. Set the learning rate $\eta = \frac{1}{6\ell_s}$ and the mini-batch size $s_t = \frac{\tau}{\omega^t}$ with $\omega = 1 - \frac{1}{480\kappa_s}$ and $\tau \geq \frac{40\alpha B}{3\rho_s\ell_s\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|^2}$. Then the output $\boldsymbol{x}^T$ of HSG-HT satisfies*

$$\mathbb{E}\|\boldsymbol{x}^T - \boldsymbol{x}^*\| \leq \left(1 - \frac{1}{480\kappa_s}\right)^{T/2}\|\boldsymbol{x}^0 - \boldsymbol{x}^*\| + \frac{\sqrt{\alpha}}{\ell_s\sqrt{12(1-\beta)}}\|\nabla_{\widetilde{\mathcal{I}}} f(\boldsymbol{x}^*)\|,$$

*where $\beta = \alpha\left(1 - \frac{1}{12\kappa_s}\right)$, $\widetilde{\mathcal{I}} = \mathit{supp}(\Phi_{2k}(\nabla f(\boldsymbol{x}^*))) \cup \mathit{supp}(\boldsymbol{x}^*)$, and $T$ is the number of iterations.*

A proof of Theorem 1 is given in Appendix B.1. Theorem 1 shows that for both sparsity- and rank-constrained problem, if using sparsity/low-rankness level $k = \Omega(\kappa_s^2 k^*)$ and gradually expanding the mini-batch size at an exponential rate of $\frac{1}{\omega}$ with $\omega = 1 - \frac{1}{480\kappa_s}$, then the estimated solution $\{\boldsymbol{x}^t\}$ generated by HSG-HT converges linearly towards $\boldsymbol{x}^*$ at the rate of $(1 - \frac{1}{480\kappa_s})^{\frac{1}{2}}$. This indicates that HSG-HT enjoys a similar fast and steady convergence rate just like the deterministic FG-HT [10]. As the condition number $\kappa_s = \ell_s/\rho_s$ is usually large in realistic problems, the exponential rate $\frac{1}{\omega}$ is actually only a slightly larger than one. This means even a moderate-scale dataset allows a lot of HSGD iterations to decrease the loss sufficiently as shown in Figure 1 and 2 in Section 5.

One can also observe that the estimation error of $\mathbb{E}\|\boldsymbol{x}^t - \boldsymbol{x}^*\|$ is controlled by the multiplier of $\|\nabla_{\widetilde{\mathcal{I}}} f(\boldsymbol{x}^*)\|$ which usually represents the statistical error of model. For sparsity-constrained problem, such a statistical error bound matches that established in FG-HT [10], and is usually better than the error bound $\mathcal{O}\left(\sqrt{s}\|\nabla f(\boldsymbol{x}^*)\|_\infty + \|\nabla_{\widetilde{\mathcal{I}}} f(\boldsymbol{x}^*)\|_2\right)$ with $s = 2k + k^*$ in SVRG-HT [18] since $\|\nabla_{\widetilde{\mathcal{I}}} f(\boldsymbol{x}^*)\|_2 \leq \sqrt{s}\|\nabla f(\boldsymbol{x}^*)\|_\infty$. Compared with the error $\mathcal{O}\left(\frac{1}{n}\sum_{i=1}^n \|\nabla f_i(\boldsymbol{x}^*)\|_2\right)$ in SG-HT [17], the error in HSG-HT is much better. This is because the magnitude $\|\nabla f(\boldsymbol{x}^*)\|_2$ of the full gradient is usually small when sample size is large, while the individual (or small mini-batch) gradient norm $\|\nabla f_i(\boldsymbol{x}^*)\|_2$ could still have relatively large magnitude. For example, in sparse linear regression problems the difference could be as significant as $\mathcal{O}(\sqrt{\log(d)/n})$ (in HSG-HT) versus $\mathcal{O}(\sqrt{\log(d)})$ (in SG-HT). Notice, for the general rank-constrained problem, FG-HT, SG-HT and SVRG do not explicitly provide the statistical error given in HSG-HT. Moreover, to guarantee convergence, SG-HT

requires the restrictive condition $\kappa_s \leq 4/3$, while our analysis removes such a condition and allows an arbitrarily large $\kappa_s$ as long as it is bounded.

Based on Theorem 1, we can derive the IFO and hard thresholding complexity of HSG-HT for problem (1) in Corollary 1 with proof in Appendix B.2. For fairness, here we follow the convention in [10, 17, 18, 19] to use $\mathbb{E}\|\boldsymbol{x} - \boldsymbol{x}^*\| \leq \sqrt{\epsilon}$ + statistical error as the measure of $\epsilon$-suboptimality.

**Corollary 1.** *Suppose the conditions in Theorem 1 hold. To achieve* $\mathbb{E}\|\boldsymbol{x}^T - \boldsymbol{x}^*\| \leq \sqrt{\epsilon} + \frac{\sqrt{\alpha}\|\nabla_{\widetilde{\mathcal{I}}} f(\boldsymbol{x}^*)\|}{\ell_s\sqrt{12(1-\beta)}}$, *the IFO complexity of HSG-HT in Algorithm 1 is* $\mathcal{O}\big(\frac{\kappa_s}{\epsilon}\big)$ *and the hard thresholding complexity is* $\mathcal{O}\big(\kappa_s \log\big(\frac{1}{\epsilon}\big)\big)$.

Compared with FG-HT [10] and SVRG-HT [18, 19] whose IFO complexity are $\mathcal{O}\left(n\kappa_s \log\left(\frac{1}{\epsilon}\right)\right)$ and $\mathcal{O}\left((n+\kappa_s)\log\left(\frac{1}{\epsilon}\right)\right)$ respectively, HSG-HT is more computationally efficient in IFO than FG-HT and SVRG-HT when sample size $n$ dominates $\frac{1}{\epsilon}$. This is usually the case when the data scale is huge while the desired accuracy $\epsilon$ is moderately small. Concerning the hard thresholding complexity, HSG-HT shares the same complexity $\mathcal{O}\left(\kappa_s \log\left(\frac{1}{\epsilon}\right)\right)$ with FG-HT, which is considerably cheaper than the $\mathcal{O}\left((n+\kappa_s)\log\left(\frac{1}{\epsilon}\right)\right)$ hard thresholding complexity of SVRG-HT when data scale is large. Overall, HSG-HT better trades-off between IFO and hard thresholding complexity than FG-HT and SVRG-HT when $n$ is much larger than $\frac{1}{\epsilon}$ in large-scale learning problems.

### 3.3 Convergence Analysis

For sparsity-constrained problem, we further investigate the convergence behavior of HSG-HT in terms of the objective value $f(\boldsymbol{x})$ to the optimal loss $f(\boldsymbol{x}^*)$. The main result is summarized in the following theorem, whose proof is deferred to Appendix B.3.

**Theorem 2.** *Suppose* $f(\boldsymbol{x})$ *is* $\rho_s$*-strongly convex and each individual component* $f_i(\boldsymbol{x})$ *is* $\ell_s$*-strongly smooth with parameter* $s = 2k + k^*$. *Let* $\kappa_s = \frac{\ell_s}{\rho_s}$ *and the sparsity level* $k \geq \left(1 + 64\kappa_s^2\right)k^*$. *By setting the learning rate* $\eta = \frac{1}{2\ell_s}$ *and the mini-batch size* $s_t = \frac{\tau}{\omega^t}$ *with* $\omega = 1 - \frac{1}{16\kappa_s}$ *and* $\tau \geq \frac{148B\kappa_s^2}{\rho_s[f(\boldsymbol{x}^0) - f(\boldsymbol{x}^*)]}$, *then for sparsity-constrained problem, the output* $\boldsymbol{x}^T$ *of Algorithm 1 satisfies*

$$\mathbb{E}[f(\boldsymbol{x}^T) - f(\boldsymbol{x}^*)] \leq \big(1 - \frac{1}{16\kappa_s}\big)^T [f(\boldsymbol{x}^0) - f(\boldsymbol{x}^*)].$$

Theorem 2 shows that for sparsity-constrained problem, HSG-HT also enjoys linear convergence rate in terms of the objective value by gradually exponentially expanding the mini-batch size. The result in Theorem 2 also implies that the expected value of $f(\boldsymbol{x}^t)$ can be arbitrarily close to the $k^*$-sparse target value $f(\boldsymbol{x}^*)$ as long as the iteration number is sufficiently large. This property is important, since in realistic problems, such as classification or regression problems, if $f(\boldsymbol{x})$ is more close to the optimum $f(\boldsymbol{x}^*)$, then the prediction result can be better. FG-HT [10] also enjoys such a good property. In contrast, for SVRG-HT [18], the convergence bound is known to be $\mathbb{E}[f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*)] \leq \mathcal{O}\left(\zeta^t + \sqrt{s}\|\nabla f(\boldsymbol{x}^*)\|_\infty\right)$ for some shrinkage rate $\zeta \in (0,1)$. The result is inferior to ours due to the presence of a non-vanishing statistical barrier term $\sqrt{s}\|\nabla f(\boldsymbol{x}^*)\|_\infty$.

## 4 Acceleration via Heavy-Ball Method

In this section, we show that HSG-HT can be effectively accelerated by applying the heavy ball technique [23, 24]. As proposed in the option **O2** in Algorithm 1, the idea is to use the integration of the estimated gradient $\boldsymbol{g}^t$ and a small momentum $\nu(\boldsymbol{x}^t - \boldsymbol{x}^{t-1})$ to modify the update as $\boldsymbol{x}^{t+1} = \Phi_k\left(\boldsymbol{x}^t - \eta\boldsymbol{g}^t + \nu(\boldsymbol{x}^t - \boldsymbol{x}^{t-1})\right)$. The following result confirms that such an accelerated variant, *i.e.* AHSG-HT, can significantly improve the efficiency of HSG-HT for twice differentiable loss functions. A proof of this result can be found in Appendix C.1.

**Theorem 3.** *Suppose the objective function* $f(\boldsymbol{x})$ *is twice differentiable and it satisfies the RSC and RSS conditions with parameter* $\widehat{s} = 3k + k^*$. *Let* $\kappa_{\widehat{s}} = \frac{\ell_{\widehat{s}}}{\rho_{\widehat{s}}}$. *Assume the sparsity/low-rankness level* $k \geq (1 + 16\kappa_{\widehat{s}})\,k^*$. *Set the learning rate* $\eta = \frac{4}{(\sqrt{\ell_{\widehat{s}}} + \sqrt{\rho_{\widehat{s}}})^2}$, *the mini-batch size* $s_t = \frac{\tau}{\omega^t}$ *where* $\omega = (1 - \frac{1}{18\sqrt{\kappa_{\widehat{s}}}})^2$ *and* $\tau \geq \frac{81B\kappa_{\widehat{s}}}{4(\sqrt{\rho_{\widehat{s}}} + \sqrt{\ell_{\widehat{s}}})^4\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|^2}$, *the momentum parameter* $\nu = \big(\frac{\sqrt{\kappa_{\widehat{s}}} - 1}{\sqrt{\kappa_{\widehat{s}}} + 1}\big)^2$. *Then*

*the output $\boldsymbol{x}^T$ of AHSG-HT in Algorithm 1 satisfies*

$$\mathbb{E}\|\boldsymbol{x}^T - \boldsymbol{x}^*\| \leq 2\big(1 - \frac{1}{18\sqrt{\kappa_s}}\big)^T \|\boldsymbol{x}^0 - \boldsymbol{x}^*\| + \frac{8\sqrt{\kappa_{\widehat{s}}}}{(\sqrt{\rho_{\widehat{s}}} + \sqrt{\ell_{\widehat{s}}})^2}\|\nabla_{\widehat{\mathcal{I}}} f(\boldsymbol{x}^*)\|,$$

*where $\widehat{\mathcal{I}} = \mathit{supp}(\Phi_{3k}(\nabla f(\boldsymbol{x}^*))) \cup \mathit{supp}(\boldsymbol{x}^*)$ and $T$ is the number of iterations.*

From this result, we can observe that for both sparsity- and rank-constrained problems, AHSG-HT has a faster convergence rate $(1-\frac{1}{18\sqrt{\kappa_{\widehat{s}}}})$ than the rate $(1-\frac{1}{480\kappa_s})^{\frac{1}{2}}$ of HSG-HT. This is because the restricted condition number $\kappa_{\widehat{s}}$ are usually comparable to or even smaller than $\kappa_s$ since the factor $k$ in $\widehat{s} = 3k + k^*$ are allowed to be smaller than that in $s = 2k + k^*$ (explained below). Also, such an acceleration relaxes the restriction on the sparsity/low-rankness level $k$: AHSG-HT allows $k = \Omega(\kappa_{\widehat{s}} k^*)$ which is considerably superior to the condition of $k = \Omega(\kappa_s^2 k^*)$ as required in other hard thresholding algorithms including HSG-HT, FG-HT and SVRG-HT. Actually, this acceleration also benefits the statistical error, since the support set $\widehat{\mathcal{I}}$ in statistical error $\mathcal{O}\|\nabla_{\widehat{\mathcal{I}}} f(\boldsymbol{x}^*)\|$ has smaller cardinality $(3k + k^*)$. So comparing with other algorithms requiring larger truncation level $k = \Omega(\kappa_s^2 k^*)$, AHSG-HT can enjoy smaller statistical error, especially for the problems with large restricted condition number $\kappa_s$.

To better illustrate the boosted efficiency, we establish the computational complexity of AHSG-HT in IFO and hard thresholding in Corollary 2, whose proof is given in Appendix C.2.

**Corollary 2.** *Suppose the conditions in Theorem 3 hold. To achieve $\mathbb{E}\|\boldsymbol{x}^T - \boldsymbol{x}^*\| \leq \sqrt{\epsilon} + \frac{8\sqrt{\kappa_{\widehat{s}}}\|\nabla_{\widehat{\mathcal{I}}} f(\boldsymbol{x}^*)\|}{(\sqrt{\rho_{\widehat{s}}}+\sqrt{\ell_{\widehat{s}}})^2}$, the IFO complexity of AHSG-HT in Algorithm 1 is $\mathcal{O}\big(\frac{\sqrt{\kappa_{\widehat{s}}}}{\epsilon}\big)$ and the hard thresholding complexity is $\mathcal{O}\big(\sqrt{\kappa_{\widehat{s}}}\log\big(\frac{1}{\epsilon}\big)\big)$.*

Corollary 2 shows that equipped with heavy ball acceleration, the IFO complexity of HSG-HT can be reduced from $\mathcal{O}\big(\frac{\kappa_s}{\epsilon}\big)$ to $\mathcal{O}\big(\frac{\sqrt{\kappa_{\widehat{s}}}}{\epsilon}\big)$, and its hard thresholding complexity can be reduced from $\mathcal{O}\big(\kappa_s \log\big(\frac{1}{\epsilon}\big)\big)$ to $\mathcal{O}\big(\sqrt{\kappa_{\widehat{s}}}\log\big(\frac{1}{\epsilon}\big)\big)$. Such an improvement on the dependence of restricted condition number is noteworthy in large-scale ill-conditioned learning problems.

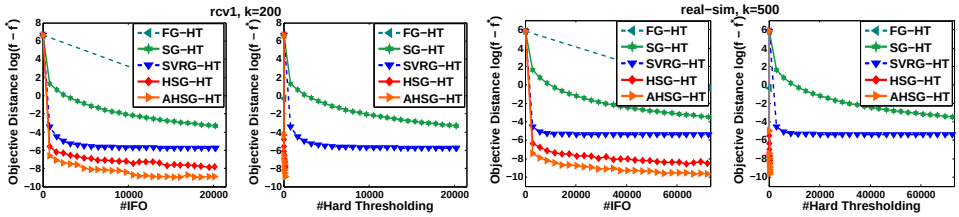



Figure 1: Single-epoch processing: comparison among hard thresholding algorithms for a single pass over data on sparse logistic regression with regularization parameter $\lambda = 10^{-5}$.

## 5 Experiments

We now compare the numerical performance of HSG-HT and AHSG-HT to several state-of-the-art algorithms, including FG-HT [10], SG-HT [17] and SVRG-HT [18, 19]. We evaluate all the considered algorithms on two sets of learning tasks. The first set contains two sparsity-constrained problems: logistic regression with $f_i(\boldsymbol{x}) = \log(1+\exp(-\boldsymbol{b}_i \boldsymbol{a}_i^\top \boldsymbol{x})) + \frac{\lambda}{2}\|\boldsymbol{x}\|_2^2$ and multi-class softmax regression with $f_i(\boldsymbol{x}) = \sum_{j=1}^c \big[\frac{\lambda}{2c}\|\boldsymbol{x}_j\|_2^2 - \mathbf{1}\{\boldsymbol{b}_i = j\}\log\frac{\exp(\boldsymbol{a}_i^\top \boldsymbol{x}_j)}{\sum_{l=1}^c \exp(\boldsymbol{a}_i^\top \boldsymbol{x}_l)}\big]$, where $\boldsymbol{b}_i$ is the target output of $\boldsymbol{a}_i$ and $c$ is the class number. The second one is a rank-constrained linear regression problem:

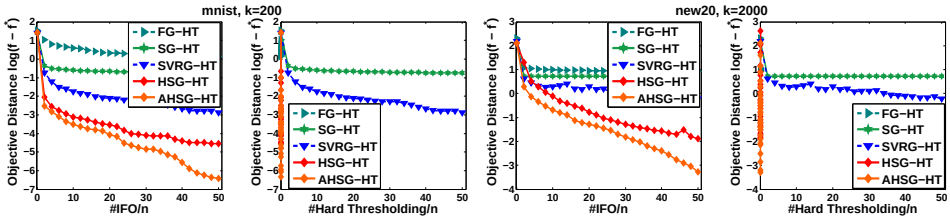
$$\min_{\boldsymbol{x}} \frac{1}{n}\sum_{i=1}^n \big[\|\boldsymbol{b}_i - \langle \boldsymbol{x}, \boldsymbol{a}_i\rangle\|_2^2 + \frac{\lambda}{2}\|\boldsymbol{x}\|_F^2\big], \quad \text{s.t. } \mathsf{rank}\,(\boldsymbol{x}) \leq k,$$

which has several important applications including multi-class classification and multi-task regression for simultaneously learning shared characteristics of all classes/tasks [32], as well as high dimensional image and financial data modeling [5, 6]. We run simulations on six datasets, including rcv1, real-sim, mnist, news20, coil100 and caltech256. The details of these data sets are described in Appendix E. For HSG-HT and AHSG-HT, we follow our theory to exponentially expand the mini-batch size $s_t$ but

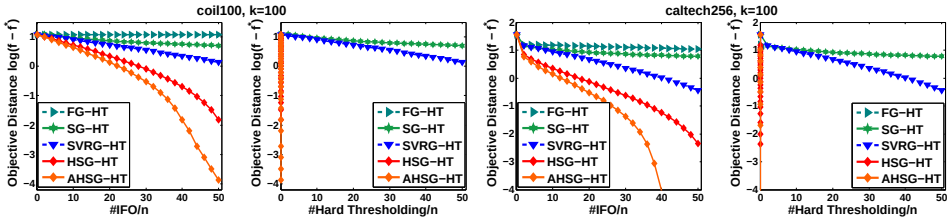
with small exponential rate, with $\tau = 1$. Since there is no ground truth on real data, we run FG-HT sufficiently long until $\|\boldsymbol{x}^t - \boldsymbol{x}^{t+1}\|/\|\boldsymbol{x}^t\| \leq 10^{-6}$, and then use the output $f(\boldsymbol{x}^t)$ as the approximate optimal value $f^*$ for sub-optimality estimation in Figure 1 and Figure 2.

**Single-epoch evaluation results.** We first consider the sparse logistic regression problem with single-epoch processing. As demonstrated in Figure 1 (more experiments in Appendix E) that HSG-HT and AHSG-HT converge significantly faster than the other considered algorithms in one pass over data. This confirms our theoretical predictions in Corollary 1 and 2 that HSG-HT and AHSG-HT are cheaper in IFO complexity than the sample-size-dependent algorithms when the desired accuracy is moderately small and data scale is large. In view of the hard thresholding complexity, AHSG-HT and HSG-HT are comparable to FG-HT and they all require much fewer hard thresholding operations than SG-HT and SVRG-HT to reach the same accuracy. This also well aligns with our theory: in one pass setting, roughly speaking, AHSG-HT and HSG-HT respectively need $\mathcal{O}\big(\sqrt{\kappa_{\widehat{s}}}\log\big(\frac{n}{\kappa_{\widehat{s}}}\big)\big)$ and $\mathcal{O}\big(\kappa_s \log\big(\frac{n}{\kappa_s}\big)\big)$ steps of hard thresholding which are both much less than the $\mathcal{O}(n)$ complexity of SG-HT and SVRG-HT. From Figure 1 and the magnifying figures in Appendix E for better displaying objective loss decrease along with hard thresholding iteration, one can observe that AHSG-HT has shaper convergence behavior than HSG-HT, which demonstrates the acceleration power of AHSG-HT.

**Multi-epoch evaluation results.** We further evaluate the considered algorithms on sparsity-constrained softmax regression and rank-constrained linear regression problems, for which an approach usually needs multiple cycles of data processing to reach high accuracy solution. In our implementation, HSG-HT (and AHSG-HT) degenerates to plain (and accelerated) FG-HT when the mini-batch size exceeds data size. The degeneration case, however, does not happen in our experiment with the specified small expanding rate. The corresponding results are illustrated in Figure 2, from which we can observe that HSG-HT and AHSG-HT exhibit much shaper convergence curves and lower hard thresholding complexity than other considered hard thresholding algorithms.




(a) Sparsity-constrained softmax problem




(b) Rank-constrained linear regression problem

Figure 2: Multi-epochs processing: comparison among hard thresholding algorithms for multiple passes over data on sparse softmax regression and rank-constrained linear regression with both regularization parameters $\lambda = 10^{-5}$.

## 6 Conclusions

In this paper, we proposed HSG-HT as a hybrid stochastic gradient hard thresholding method for sparsity/rank-constrained empirical risk minimization problems. We proved that HSG-HT enjoys the $\mathcal{O}\big(\kappa_s \log\big(\frac{1}{\epsilon}\big)\big)$ hard thresholding complexity like full gradient methods, while possessing sample-size-independent IFO complexity of $\mathcal{O}\big(\frac{\kappa_s}{\epsilon}\big)$. Compared to the existing variance-reduced hard thresholding algorithms, HSG-HT enjoys lower overall computational cost when sample size is large and the accuracy is moderately small. Furthermore, we showed that HSG-HT can be effectively accelerated via applying the heavy ball acceleration technique to attain improved dependence on restricted condition number. The provable efficiency of HSGT-HT and its accelerated variant has been confirmed by extensive numerical evaluation with comparison against the state-of-the-art algorithms.

## References


[1] D. Donoho. Compressed sensing. *IEEE Trans. on Information Theory*, 52(4):1289–1306, 2006. 1

[2] J. Tropp and A. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Trans. on Information Theory*, 53(12):4655–4666, 2007. 1

[3] S. Bahmani, B. Raj, and P. Boufounos. Greedy sparsity-constrained optimization. *Journal of Machine Learning Research*, 14:807–841, 2013. 1

[4] A. Jalali, C. Johnson, and P. Ravikumar. On learning discrete graphical models using greedy methods. In *Proc. Conf. Neutral Information Processing Systems*, pages 1935–1943, 2011. 1

[5] S. Negahban and M. Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, pages 1069–1097, 2011. 1, 7

[6] A. Rohde and A. Tsybakov. Estimation of high-dimensional low-rank matrices. *The Annals of Statistics*, 39(2):887–930, 2011. 1, 7

[7] T. Blumensath and M. Davies. Iterative hard thresholding for compressed sensing. *Applied and computational harmonic analysis*, 27(3):265–274, 2009. 1

[8] S. Foucart. Hard thresholding pursuit: an algorithm for compressive sensing. *SIAM Journal on Numerical Analysis*, 49(6):2543–2563, 2011. 1

[9] X. Yuan, P. Li, and T. Zhang. Gradient hard thresholding pursuit for sparsity-constrained optimization. In *Proc. Int'l Conf. Machine Learning*, pages 127–135, 2014. 1, 2, 4

[10] P. Jain, A. Tewari, and P. Kar. On iterative hard thresholding methods for high-dimensional M-estimation. In *Proc. Conf. Neutral Information Processing Systems*, pages 685–693, 2014. 1, 2, 4, 5, 6, 7

[11] R. Garg and Rohit R. Khandekar. Gradient descent with sparsification: an iterative algorithm for sparse recovery with restricted isometry property. In *Proc. Int'l Conf. Machine Learning*, pages 337–344. ACM, 2009. 1

[12] X. Yuan, P. Li, and T. Zhang. Exact recovery of hard thresholding pursuit. In *Proc. Conf. Neutral Information Processing Systems*, pages 3558–3566, 2016. 2

[13] R. Tibshirani. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996. 2

[14] S. Van de Geer. High-dimensional generalized linear models and the LASSO. *The Annals of Statistics*, 36(2):614–645, 2008. 2

[15] B. Recht, M. Fazel, and P. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010. 2

[16] N. Srebro, J. Rennie, and T. Jaakkola. Maximum-margin matrix factorization. In *Proc. Conf. Neutral Information Processing Systems*, pages 1329–1336, 2005. 2

[17] N. Nguyen, D. Needell, and T. Woolf. Linear convergence of stochastic iterative greedy algorithms with sparse constraints. *IEEE Trans. on Information Theory*, 63(11):6869–6895, 2017. 2, 4, 5, 6, 7

[18] X. Li, R. Arora, H. Liu, J. Haupt, and T. Zhao. Nonconvex sparse learning via stochastic optimization with progressive variance reduction. *Proc. Int'l Conf. Machine Learning*, 2016. 2, 4, 5, 6, 7

[19] J. Shen and P. Li. A tight bound of hard thresholding. *arXiv preprint arXiv:1605.01656*, 2016. 2, 4, 6, 7

[20] R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Proc. Conf. Neutral Information Processing Systems*, pages 315–323, 2013. 2, 4

[21] D. P. Bertsekas. A new class of incremental gradient methods for least squares problems. *SIAM Journal on Optimization*, 7(4):913–926, 1997. 2

[22] M. Friedlander and M. Schmidt. Hybrid deterministic-stochastic methods for data fitting. *SIAM Journal on Scientific Computing*, 34(3):A1380–A1405, 2012. 2

[23] B. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964. 3, 6

[24] A. Govan. Introduction to optimization. In *North Carolina State University, SAMSI NDHS, Undergraduate workshop*, 2006. 3, 6

[25] N. Qian. On the momentum term in gradient descent learning algorithms. *Neural networks*, 12(1):145–151, 1999. 3

[26] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013. 3

[27] P. Jain, S. Kakade, R. Kidambi, P. Netrapalli, and A. Sidford. Accelerating stochastic gradient descent. *arXiv preprint arXiv:1704.08227*, 2017. 3

[28] S. Reddi, S. Kale, and S. Kumar. On the convergence of adam and beyond. In *Proc. Int'l Conf. Learning Representations*, 2018. 3

[29] K. Rajiv and K. Anastasios. IHT dies hard: Provable accelerated iterative hard thresholding. In *Proc. Int'l Conf. Artificial Intelligence and Statistics*, volume 84, pages 188–198, 2018. 3

[30] Y. Zhang and L. Xiao. Stochastic primal-dual coordinate method for regularized empirical risk minimization. In *Proc. Int'l Conf. Machine Learning*, pages 353–361, 2015. 4

[31] Q. Lin, Z. Lu, and L. Xiao. An accelerated proximal coordinate gradient method. In *Proc. Conf. Neutral Information Processing Systems*, pages 3059–3067, 2014. 4

[32] Y. Amit, M. Fink, N. Srebro, and S. Ullman. Uncovering shared structures in multiclass classification. In *Proc. Int'l Conf. Machine Learning*, pages 17–24. ACM, 2007. 7

# Efficient Stochastic Gradient Hard Thresholding

**Pan Zhou**[*] **Xiaotong Yuan**[†] **Jiashi Feng**[*]
[*] Learning & Vision Lab, National University of Singapore, Singapore
[†] B-DAT Lab, Nanjing University of Information Science & Technology, Nanjing, China
pzhou@u.nus.edu xtyuan@nuist.edu.cn elefjia@nus.edu.sg

## Abstract

This supplementary document contains the technical proofs of convergence results and some additional numerical results of the NIPS'18 submission entitled "Efficient Stochastic Gradient Hard Thresholding". It is structured as follows. The proof of the results in Section 3, including Theorems 1 and 2 and Corollary 1, is presented in Appendix B. Then Appendix C provides the proof of the results in Section 4, including Theorems 3 and Corollary 2. Next, Appendix D gives the proof of auxiliary lemmas. Finally, the detailed descriptions of datasets and more experimental results are provided in Appendix E.

## A Auxiliary Lemmas

In this section, we introduce auxiliary lemmas which will be used for proving the results in the manuscript. For readability, we defer the proofs of some lemmas into Appendix D.

**Lemma 1.** *[1] When $\Phi_k(\boldsymbol{x}) : \mathbb{R}^d \to \mathbb{R}^d$ is a vector hard thresholding operator that keeps the largest $k$ entries in $\boldsymbol{x}$ and sets other entries to zero. For any $k > k^*$ where $\|\boldsymbol{x}^*\|_0 = k^*$, we have*

$$\|\Phi_k(\boldsymbol{x}) - \boldsymbol{x}^*\|^2 \leq \left(1 + \frac{2\sqrt{k^*}}{\sqrt{k - k^*}}\right)\|\boldsymbol{x} - \boldsymbol{x}^*\|^2. \tag{3}$$

*On the other hand, when $\Phi_k(\boldsymbol{x}) : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^{d_1 \times d_2}$ is a matrix hard thresholding operator that keeps the largest $k$ top singular values of matrix $\boldsymbol{x}$ and sets other singular values to zero. For any $k > k^*$ where $\mathsf{rank}(\boldsymbol{x}^*) = k^*$, the property (3) still holds.*

**Lemma 2.** *[2] When $\Phi_k(\boldsymbol{x}) : \mathbb{R}^d \to \mathbb{R}^d$ is a vector hard thresholding operator that keeps the largest $k$ entries in $\boldsymbol{x}$ and sets other entries to zero. Assume $k > k^*$, $\boldsymbol{y} = \Phi_k(\boldsymbol{x})$ and $\boldsymbol{y}^* = \Phi_k(\boldsymbol{x})$, where $\|\boldsymbol{x}^*\|_0 = k^*$. Then we have*

$$\|\boldsymbol{y} - \boldsymbol{x}\|^2 \leq \frac{d - k}{d - k^*}\|\boldsymbol{y}^* - \boldsymbol{x}\|^2.$$

*On the other hand, when $\Phi_k(\boldsymbol{x}) : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^{d_1 \times d_2}$ is a matrix hard thresholding operator that keeps the largest $k$ top singular values of matrix $\boldsymbol{x}$ and sets other singular values to zero. For any $k > k^*$ where $\mathsf{rank}(\boldsymbol{x}^*) = k^*$, we have*

$$\|\boldsymbol{y} - \boldsymbol{x}\|^2 \leq \frac{r - k}{r - k^*}\|\boldsymbol{y}^* - \boldsymbol{x}\|^2.$$

*where $r = \mathsf{rank}(\boldsymbol{x})$.*

**Lemma 3.** *[3] Assume that $\boldsymbol{g}^t$ is the sampled gradient in Algorithm 1 for sparsity- or rank-constrained problem. Then the gradient variance of the gradient estimation $\boldsymbol{g}^t$ can be bounded as follows*

$$\mathbb{E}\|\boldsymbol{g}^t - \nabla f(\boldsymbol{x}^t)\|^2 \leq \frac{n - s_t}{n}\frac{1}{s_t}B_t,$$

*where $B_t = \frac{1}{n-1}\sum_{i=1}^n \|\nabla f_i(\boldsymbol{x}^t) - \nabla f(\boldsymbol{x}^t)\|^2$ and $\boldsymbol{x}^t$ denotes the variable at the $t$-th iteration.*

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

**Lemma 4.** *Assume that $\boldsymbol{g}^t$ is the sampled gradient in Algorithm 1 for the sparsity- or rank-constrained problem. Then we can bound $\mathbb{E}\|\boldsymbol{g}_{\mathcal{I}}^t\|^2$ as follows*

$$\mathbb{E}\|\boldsymbol{g}_{\mathcal{I}}^t\|^2 \leq \frac{3}{s_t}B_t + 6\ell_s\left(f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*) + \langle\nabla_{\mathcal{I}} f(\boldsymbol{x}^*), \boldsymbol{x}^t - \boldsymbol{x}^*\rangle\right) + 3\|\nabla_{\mathcal{I}} f(\boldsymbol{x}^*)\|^2,$$

*where $B_t = \frac{1}{n-1}\sum_{i=1}^n \|\nabla f_i(\boldsymbol{x}^t) - \nabla f(\boldsymbol{x}^t)\|^2$, $\mathcal{I} = \mathit{supp}(\boldsymbol{x}^*)\bigcup \mathit{supp}(\boldsymbol{x}^t)\bigcup \mathit{supp}(\boldsymbol{x}^{t+1})$ and $\ell_s$ denotes the smooth parameter with $s = 2k + k^*$. Here $k^*$ denotes the cardinality of the support set $\mathit{supp}(\boldsymbol{x}^*)$.*

*Proof.* We defer the proof of Lemma 4 to Appendix D.1. □

**Lemma 5.** *For both vector variable $\boldsymbol{x}$ in sparsity-constrained problem or matrix variable $\boldsymbol{x}$ in rank-constrained problem, we have*

$$\|\boldsymbol{x}^t - \boldsymbol{x}^*\|^2 \leq \frac{4}{\rho_s}(f(\boldsymbol{x}^*) - f(\boldsymbol{x}^t)) + \frac{8}{\rho_s^2}\|\boldsymbol{g}_{\mathcal{I}^t\cup\mathcal{I}^*}^t\|^2 + \frac{8}{\rho_s^2}\|\nabla f(\boldsymbol{x}^t) - \boldsymbol{g}^t\|^2,$$

*where $\mathcal{I}^t = \mathit{supp}(\boldsymbol{x}^t)$ and $\mathcal{I}^* = \mathit{supp}(\boldsymbol{x}^*)$.*

*Proof.* See the proof of Lemma 5 in Appendix D.2. □

**Lemma 6.** *Let $\boldsymbol{A} \in \mathbb{R}^{d\times d}$. Assume that $\mu\boldsymbol{I} \preceq \boldsymbol{A} \preceq \ell\boldsymbol{I}$ for some $0 < \mu < \ell$. Then the following inequality holds*

$$\left\|\begin{bmatrix}(1+\nu)\boldsymbol{I} - \eta\boldsymbol{A} & -\nu\boldsymbol{I}\\ \boldsymbol{I} & 0\end{bmatrix}\right\| \leq \max\left\{|1-\sqrt{\eta\mu}|, |1-\sqrt{\eta\ell}|\right\}$$

*for $\nu = \max\left\{|1-\sqrt{\eta\mu}|^2, |1-\sqrt{\eta\ell}|^2\right\}$.*

*Proof.* We defer the proof of Lemma 6 to Appendix D.3. □

# B Proofs for Section 3

## B.1 Proof of Theorem 1

*Proof.* Here we also first consider the sparsity-constrained problem. Assume $\boldsymbol{v} = \boldsymbol{x}^t - \eta\boldsymbol{g}_{\mathcal{I}}^t$ and $\mathcal{I} = \mathcal{I}^*\cup\mathcal{I}^t\cup\mathcal{I}^{t+1}$, where $\mathcal{I}^* = \mathsf{supp}(\boldsymbol{x}^*)$, $\mathcal{I}^t = \mathsf{supp}(\boldsymbol{x}^t)$ and $\mathcal{I}^{t+1} = \mathsf{supp}(\boldsymbol{x}^{t+1})$. Then we have

$$\begin{aligned}
&\mathbb{E}\|\boldsymbol{v} - \boldsymbol{x}^*\|^2\\
=&\mathbb{E}\|\boldsymbol{x}^t - \eta\boldsymbol{g}_{\mathcal{I}}^t - \boldsymbol{x}^*\|^2\\
=&\mathbb{E}\|\boldsymbol{x}^t - \boldsymbol{x}^*\|^2 + \eta^2\mathbb{E}\|\boldsymbol{g}_{\mathcal{I}}^t\|^2 - 2\eta\mathbb{E}\langle\boldsymbol{x}^t - \boldsymbol{x}^*, \boldsymbol{g}_{\mathcal{I}}^t\rangle \qquad (4)\\
\overset{①}{\leq}&\mathbb{E}\|\boldsymbol{x}^t - \boldsymbol{x}^*\|^2 + \eta^2\mathbb{E}\|\boldsymbol{g}_{\mathcal{I}}^t\|^2 - 2\eta\mathbb{E}\left(f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*)\right)\\
\overset{②}{\leq}&\mathbb{E}\|\boldsymbol{x}^t - \boldsymbol{x}^*\|^2 + \eta^2\left[\frac{3}{s_t}B_t + 6\ell_s\left(f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*) + \langle\nabla_{\mathcal{I}} f(\boldsymbol{x}^*), \boldsymbol{x}^t - \boldsymbol{x}^*\rangle\right) + 3\|\nabla_{\mathcal{I}} f(\boldsymbol{x}^*)\|^2\right]\\
&- 2\eta\mathbb{E}\left[f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*)\right]\\
\overset{③}{=}&\mathbb{E}\|\boldsymbol{x}^t - \boldsymbol{x}^*\|^2 + 2\eta(3\eta\ell_s - 1)\left[f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*)\right] + 6\eta^2\ell_s\langle\nabla_{\mathcal{I}} f(\boldsymbol{x}^*), \boldsymbol{x}^t - \boldsymbol{x}^*\rangle + \frac{3\eta^2}{s_t}B_t + 3\eta^2\|\nabla_{\mathcal{I}} f(\boldsymbol{x}^*)\|^2,
\end{aligned}$$

where ① holds by using $\|(\boldsymbol{x}^t - \boldsymbol{x}^*)_{\mathcal{I}^c}\| = 0$ and the convexity of $f(\boldsymbol{x})$, namely, $f(\boldsymbol{x}^*) - f(\boldsymbol{x}^t) \geq \langle\boldsymbol{x}^* - \boldsymbol{x}^t, \nabla f(\boldsymbol{x}^t)\rangle = \langle\boldsymbol{x}^* - \boldsymbol{x}^t, \nabla_{\mathcal{I}} f(\boldsymbol{x}^t)\rangle$, and ② holds by using Lemma 4, and ③ holds by using $\|\nabla_{\mathcal{I}} f(\boldsymbol{x}^*)\|_2^2 \leq \|\nabla_{\widetilde{\mathcal{I}}} f(\boldsymbol{x}^*)\|_2^2$ where $\widetilde{\mathcal{I}} = \mathsf{supp}(\Phi_{2k}(\nabla f(\boldsymbol{x}^*)))\cup\mathsf{supp}(\boldsymbol{x}^*)$.

Next, we apply Lemma 1 and obtain

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\|^2 \leq&\alpha\mathbb{E}\|\boldsymbol{v} - \boldsymbol{x}^*\|^2 \qquad (5)\\
\leq&\alpha\left[\mathbb{E}\|\boldsymbol{x}^t - \boldsymbol{x}^*\|^2 + 2\eta(3\eta\ell - 1)\left[f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*)\right] + 6\eta^2\ell_s\langle\nabla_{\mathcal{I}} f(\boldsymbol{x}^*), \boldsymbol{x}^t - \boldsymbol{x}^*\rangle\right]\\
&+ \alpha\left[\frac{3\eta^2}{s_t}B_t + 3\eta^2\|\nabla_{\mathcal{I}} f(\boldsymbol{x}^*)\|^2\right],
\end{aligned}$$

where $\alpha = 1 + \frac{2\sqrt{k^*}}{\sqrt{k-k^*}}$. On the other hand, we have

$$f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*) \geq \langle \nabla f(\boldsymbol{x}^*), \boldsymbol{x}^t - \boldsymbol{x}^* \rangle + \frac{\rho_s}{2}\|\boldsymbol{x}^t - \boldsymbol{x}^*\|^2. \tag{6}$$

By setting $\eta \leq \frac{1}{3\ell_s}$ and plugging Eqn. (6) into Eqn. (5), we can obtain

$$\begin{aligned}
&\mathbb{E}\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\|^2 \\
\leq &\alpha\left[1+\rho\eta(3\eta\ell_s-1)\right]\mathbb{E}\|\boldsymbol{x}^t-\boldsymbol{x}^*\|^2+2\alpha\eta(6\eta\ell_s-1)\langle\nabla f(\boldsymbol{x}^*), \boldsymbol{x}^t - \boldsymbol{x}^*\rangle+\frac{3\alpha\eta^2}{s_t}B_t+3\alpha\eta^2\|\nabla_{\mathcal{I}} f(\boldsymbol{x}^*)\|^2 \\
\overset{①}{\leq} &\alpha\left[1+\rho\eta(3\eta\ell_s-1)\right]\mathbb{E}\|\boldsymbol{x}^t-\boldsymbol{x}^*\|^2+\frac{3\alpha\eta^2}{s_t}B+3\alpha\eta^2\|\nabla_{\mathcal{I}} f(\boldsymbol{x}^*)\|^2,
\end{aligned} \tag{7}$$

where ① holds since we set $\eta = \frac{1}{6\ell_s}$ and $B = \max_t B_t$. Then we let

$$\beta := \alpha\left[1+\rho\eta(3\eta\ell_s-1)\right] = \alpha\left(1 - \frac{1}{12\kappa_s}\right) \in (0,1) \quad \text{where} \quad \kappa_s = \frac{\ell_s}{\rho_s}.$$

We further set $s_t = \tau/\omega^t$ and assume that $\tau$ is large enough such that

$$\gamma := \frac{3\alpha\eta^2 B}{\tau} = \frac{\alpha B}{12\tau\ell_s^2} \leq \delta\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|^2, \tag{8}$$

where $\delta$ is a positive constant and will be discussed later. Now we use mathematical induction to prove

$$\mathbb{E}\|\boldsymbol{x}^t - \boldsymbol{x}^*\|^2 \leq \theta^t\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|^2 + \frac{\alpha}{12(1-\beta)\ell_s^2}\|\nabla_{\mathcal{I}} f(\boldsymbol{x}^*)\|^2, \ (\forall t), \tag{9}$$

where $\theta < 1$ is a constant and will be given below.

Obviously, when $t = 0$, Eqn. (9) holds. Now assume that for all $k \leq t$, Eqn. (9) holds. Then for $k = t + 1$, we have

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\|^2 \leq &\beta\mathbb{E}\|\boldsymbol{x}^t - \boldsymbol{x}^*\|^2 + \delta\omega^t\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|^2 + \frac{\alpha}{12\ell_s^2}\|\nabla_{\mathcal{I}} f(\boldsymbol{x}^*)\|^2 \\
\leq &(\beta\theta^t + \delta\omega^t)\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|^2 + \frac{\alpha}{12\ell_s^2}\left[\frac{\beta}{1-\beta} + 1\right]\|\nabla_{\mathcal{I}} f(\boldsymbol{x}^*)\|_2^2 \\
\overset{①}{\leq} &\theta^{t+1}\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|^2 + \frac{\alpha}{12(1-\beta)\ell_s^2}\|\nabla_{\mathcal{I}} f(\boldsymbol{x}^*)\|_2^2
\end{aligned}$$

where ① holds since we let

$$\omega = \theta = \beta + \delta. \tag{10}$$

This means that if Eqn. (18) holds, then Eqn. (9) always holds. So the conclusion holds.

Finally, we discuss the value of $k$ such that $\beta \in (0,1)$. It is easily to check that if $k \geq \left(\frac{6400}{9}\kappa_s^2 + 1\right)k^*$, where $\kappa_s = \ell_s/\rho_s$, then we have $\beta = \alpha\left(1 - \frac{1}{12\kappa_s}\right) < 1 - \frac{1}{120\kappa_s} - \frac{1}{160\kappa_s^2}$. So we just let

$$k \geq \left(712\kappa_s^2 + 1\right)k^*, \quad \text{and} \quad \beta = \alpha\left(1 - \frac{1}{12\kappa_s}\right) \leq 1 - \frac{1}{120\kappa_s}.$$

Then let $\tau \geq \frac{40\alpha B}{3\ell_s\rho_s\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|^2}$. We have

$$\omega = \theta = \beta + \delta \leq 1 - \frac{1}{120\kappa_s} + \frac{3}{480\kappa_s} = 1 - \frac{1}{480\kappa_s}.$$

Therefore, we have

$$\mathbb{E}\|\boldsymbol{x}^t - \boldsymbol{x}^*\|^2 \leq \theta^t\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|^2 + \frac{\alpha}{12(1-\beta)\ell_s^2}\|\nabla_{\mathcal{I}} f(\boldsymbol{x}^*)\|^2, \ (\forall t), \tag{11}$$

where $\beta = \alpha\left(1 - \frac{1}{12\kappa_s}\right) \leq 1 - \frac{1}{120\kappa_s}$ and $\theta \leq 1 - \frac{1}{480\kappa_s}$. Then we can further derive

$$\begin{aligned}\mathbb{E}\|\boldsymbol{x}^T - \boldsymbol{x}^*\| \leq & \sqrt{\mathbb{E}\|\boldsymbol{x}^T - \boldsymbol{x}^*\|^2} \leq \sqrt{\theta^T\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|^2 + \frac{\alpha}{12(1-\beta)\ell_s^2}\|\nabla_{\widetilde{\mathcal{I}}} f(\boldsymbol{x}^*)\|^2} \\ \leq & \theta^{\frac{T}{2}}\|\boldsymbol{x}^0 - \boldsymbol{x}^*\| + \frac{\sqrt{\alpha}}{\ell_s\sqrt{12(1-\beta)}}\|\nabla_{\widetilde{\mathcal{I}}} f(\boldsymbol{x}^*)\|,\end{aligned}$$

where $\theta \leq 1 - \frac{1}{480\kappa_s}$. So we obtain the result on sparsity-constrained problem.

For rank-constrained problem, its proof is very similar to the proof for sparsity-constrained problem. Firstly, assume that the skinny SVDs of $\boldsymbol{x}^t$, $\boldsymbol{x}^{t+1}$, $\boldsymbol{x}^*$ are respectively $\boldsymbol{x}^t = \boldsymbol{U}_t\boldsymbol{\Sigma}_t\boldsymbol{V}_t^T$, $\boldsymbol{x}^{t+1} = \boldsymbol{U}_{t+1}\boldsymbol{\Sigma}_{t+1}\boldsymbol{V}_{t+1}^T$ and $\boldsymbol{x}^* = \boldsymbol{U}_*\boldsymbol{\Sigma}_*\boldsymbol{V}_*^T$. Then we have $\boldsymbol{x}^t = \boldsymbol{U}_t\boldsymbol{U}_t^T\boldsymbol{x}^t$, $\boldsymbol{x}^{t+1} = \boldsymbol{U}_{t+1}\boldsymbol{U}_{t+1}^T\boldsymbol{x}^{t+1}$ and $\boldsymbol{x}^* = \boldsymbol{U}_*\boldsymbol{U}_*^T\boldsymbol{x}^*$. We further define the projection operation $\mathcal{P}_{\boldsymbol{U}}(\boldsymbol{x}) = \boldsymbol{U}\boldsymbol{U}^T\boldsymbol{x}$ which projects $\boldsymbol{x}$ in the subspace spanned by $\boldsymbol{U}$. Then let $\boldsymbol{v} = \boldsymbol{x}^t - \eta\boldsymbol{g}_{\mathcal{I}}^t$ and $\mathcal{I} = \mathcal{I}^* \cup \mathcal{I}^t \cup \mathcal{I}^{t+1}$, where $\mathcal{I}^* = \mathsf{supp}(\boldsymbol{x}^*)$, $\mathcal{I}^t = \mathsf{supp}(\boldsymbol{x}^t)$ and $\mathcal{I}^{t+1} = \mathsf{supp}(\boldsymbol{x}^{t+1})$. The notation $\boldsymbol{x}_{\mathcal{I}^t}^t = \boldsymbol{U}_t\boldsymbol{U}_t^T\boldsymbol{x}^t = \mathcal{P}_{\boldsymbol{U}^t}(\boldsymbol{x}^t)$. For the specifical notation $\boldsymbol{g}_{\mathcal{I}}^t$, it denotes $\boldsymbol{g}_{\mathcal{I}}^t = \mathcal{P}_{\mathcal{I}}(\boldsymbol{g}^t) = (\boldsymbol{U}_t\boldsymbol{U}_t^T + \boldsymbol{U}_{t+1}\boldsymbol{U}_{t+1}^T + \boldsymbol{U}_*\boldsymbol{U}_*^T - \boldsymbol{U}_t\boldsymbol{U}_t^T\boldsymbol{U}_{t+1}\boldsymbol{U}_{t+1}^T - \boldsymbol{U}_t\boldsymbol{U}_t^T\boldsymbol{U}_*\boldsymbol{U}_*^T - \boldsymbol{U}_{t+1}\boldsymbol{U}_{t+1}^T\boldsymbol{U}_*\boldsymbol{U}_*^T + \boldsymbol{U}_t\boldsymbol{U}_t^T\boldsymbol{U}_{t+1}\boldsymbol{U}_{t+1}^T\boldsymbol{U}_*\boldsymbol{U}_*^T)\boldsymbol{g}^t$ where $\mathcal{P}_{\mathcal{I}}$ is a projection.

Then we prove the result similar to Eqn. (4) as follows:

$$\begin{aligned}&\mathbb{E}\|\boldsymbol{v} - \boldsymbol{x}^*\|^2 \\ =&\mathbb{E}\|\boldsymbol{x}^t - \eta\boldsymbol{g}_{\mathcal{I}}^t - \boldsymbol{x}^*\|^2 \\ =&\mathbb{E}\|\boldsymbol{x}^t - \boldsymbol{x}^*\|^2 + \eta^2\mathbb{E}\|\boldsymbol{g}_{\mathcal{I}}^t\|^2 - 2\eta\mathbb{E}\langle\boldsymbol{x}^t - \boldsymbol{x}^*, \boldsymbol{g}_{\mathcal{I}}^t\rangle \\ \overset{①}{\leq}&\mathbb{E}\|\boldsymbol{x}^t - \boldsymbol{x}^*\|^2 + \eta^2\mathbb{E}\|\boldsymbol{g}_{\mathcal{I}}^t\|^2 - 2\eta\mathbb{E}\left(f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*)\right) \\ \overset{②}{\leq}&\mathbb{E}\|\boldsymbol{x}^t - \boldsymbol{x}^*\|^2 + \eta^2\left[\frac{3}{s_t}B_t + 6\ell_s\left(f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*) + \langle\nabla_{\mathcal{I}} f(\boldsymbol{x}^*), \boldsymbol{x}^t - \boldsymbol{x}^*\rangle\right) + 3\|\nabla_{\mathcal{I}} f(\boldsymbol{x}^*)\|^2\right] \\ &- 2\eta\mathbb{E}\left[f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*)\right] \\ \overset{③}{=}&\mathbb{E}\|\boldsymbol{x}^t - \boldsymbol{x}^*\|^2 + 2\eta(3\eta\ell_s - 1)\left[f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*)\right] + 6\eta^2\ell_s\langle\nabla_{\mathcal{I}} f(\boldsymbol{x}^*), \boldsymbol{x}^t - \boldsymbol{x}^*\rangle + \frac{3\eta^2}{s_t}B_t + 3\eta^2\|\nabla_{\mathcal{I}} f(\boldsymbol{x}^*)\|^2,\end{aligned}$$

where ① holds since we have $\mathcal{P}_{\mathcal{I}}(\boldsymbol{x}^t - \boldsymbol{x}^*) = (\boldsymbol{U}_t\boldsymbol{U}_t^T + \boldsymbol{U}_{t+1}\boldsymbol{U}_{t+1}^T + \boldsymbol{U}_*\boldsymbol{U}_*^T - \boldsymbol{U}_t\boldsymbol{U}_t^T\boldsymbol{U}_{t+1}\boldsymbol{U}_{t+1}^T - \boldsymbol{U}_t\boldsymbol{U}_t^T\boldsymbol{U}_*\boldsymbol{U}_*^T - \boldsymbol{U}_{t+1}\boldsymbol{U}_{t+1}^T\boldsymbol{U}_*\boldsymbol{U}_*^T + \boldsymbol{U}_t\boldsymbol{U}_t^T\boldsymbol{U}_{t+1}\boldsymbol{U}_{t+1}^T\boldsymbol{U}_*\boldsymbol{U}_*^T)(\boldsymbol{U}_t\boldsymbol{U}_t^T\boldsymbol{x}^t - \boldsymbol{U}_*\boldsymbol{U}_*^T\boldsymbol{x}^*) = \boldsymbol{x}^t - \boldsymbol{x}^*$ which gives $\|(\boldsymbol{x}^t - \boldsymbol{x}^*)_{\mathcal{I}^c}\| = \|(\boldsymbol{I} - \mathcal{P}_{\mathcal{I}})(\boldsymbol{x}^t - \boldsymbol{x}^*)\| = 0$ and $\mathbb{E}\langle\boldsymbol{x}^t - \boldsymbol{x}^*, \boldsymbol{g}_{\mathcal{I}}^t\rangle = \mathbb{E}\langle\boldsymbol{x}^t - \boldsymbol{x}^*, \mathcal{P}_{\mathcal{I}}(\boldsymbol{g}^t)\rangle = \mathbb{E}\langle\mathcal{P}_{\mathcal{I}}(\boldsymbol{x}^t - \boldsymbol{x}^*), \mathcal{P}_{\mathcal{I}}(\boldsymbol{g}^t)\rangle = \mathbb{E}\langle\boldsymbol{x}^t - \boldsymbol{x}^*, \boldsymbol{g}^t\rangle = \mathbb{E}\langle\boldsymbol{x}^t - \boldsymbol{x}^*, \nabla f(\boldsymbol{x}^t)\rangle \leq f(\boldsymbol{x}^*) - f(\boldsymbol{x}^t)$. ② still uses the results in Lemma 4. ③ holds by using $\|\nabla_{\mathcal{I}} f(\boldsymbol{x}^*)\|_2^2 \leq \|\nabla_{\widetilde{\mathcal{I}}} f(\boldsymbol{x}^*)\|_2^2$ where $\widetilde{\mathcal{I}} = \mathsf{supp}(\Phi_{2k}(\nabla f(\boldsymbol{x}^*))) \cup \mathsf{supp}(\boldsymbol{x}^*)$.

Before we apply Lemma 1 to obtain similar result in Eqn. (5). We first establish

$$\begin{aligned}\Phi_k(\boldsymbol{v}) =&\boldsymbol{U}_{t+1}\boldsymbol{U}_{t+1}^T(\boldsymbol{x}^t - \eta\boldsymbol{g}_{\mathcal{I}}^t) = \boldsymbol{U}_{t+1}\boldsymbol{U}_{t+1}^T\boldsymbol{x}^t - \eta\boldsymbol{U}_{t+1}\boldsymbol{U}_{t+1}^T[(\boldsymbol{U}_t\boldsymbol{U}_t^T + \boldsymbol{U}_{t+1}\boldsymbol{U}_{t+1}^T + \boldsymbol{U}_*\boldsymbol{U}_*^T \\ &- \boldsymbol{U}_t\boldsymbol{U}_t^T\boldsymbol{U}_{t+1}\boldsymbol{U}_{t+1}^T - \boldsymbol{U}_t\boldsymbol{U}_t^T\boldsymbol{U}_*\boldsymbol{U}_*^T - \boldsymbol{U}_{t+1}\boldsymbol{U}_{t+1}^T\boldsymbol{U}_*\boldsymbol{U}_*^T + \boldsymbol{U}_t\boldsymbol{U}_t^T\boldsymbol{U}_{t+1}\boldsymbol{U}_{t+1}^T\boldsymbol{U}_*\boldsymbol{U}_*^T)]\boldsymbol{g}^t \\ =&\boldsymbol{U}_{t+1}\boldsymbol{U}_{t+1}^T\boldsymbol{x}^t - \eta\boldsymbol{U}_{t+1}\boldsymbol{U}_{t+1}^T\boldsymbol{g}^t = \boldsymbol{U}_{t+1}\boldsymbol{U}_{t+1}^T(\boldsymbol{x}^t - \eta\boldsymbol{g}^t) = \Phi_k(\boldsymbol{x}^t - \eta\boldsymbol{g}^t) = \boldsymbol{x}^{t+1}.\end{aligned} \tag{12}$$

Then we apply Lemma 1 to establish

$$\begin{aligned}\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\|^2 =&\|\Phi_k(\boldsymbol{v}) - \boldsymbol{x}^*\|^2 \leq \left(1 + \frac{2\sqrt{k^*}}{\sqrt{k - k^*}}\right)\|\boldsymbol{v} - \boldsymbol{x}^*\|^2 \\ \leq&\alpha\left[\mathbb{E}\|\boldsymbol{x}^t - \boldsymbol{x}^*\|^2 + 2\eta(3\eta\ell - 1)\left[f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*)\right] + 6\eta^2\ell_s\langle\nabla_{\mathcal{I}} f(\boldsymbol{x}^*), \boldsymbol{x}^t - \boldsymbol{x}^*\rangle\right] \\ &+ \alpha\left[\frac{3\eta^2}{s_t}B_t + 3\eta^2\|\nabla_{\mathcal{I}} f(\boldsymbol{x}^*)\|^2\right],\end{aligned}$$

where $\alpha = 1 + \frac{2\sqrt{k^*}}{\sqrt{k - k^*}}$.

Next by using Assumption in which we have $f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*) \geq \langle \nabla f(\boldsymbol{x}^*), \boldsymbol{x}^t - \boldsymbol{x}^* \rangle + \frac{\rho_s}{2}\|\boldsymbol{x}^t - \boldsymbol{x}^*\|^2$, we can establish the result in Eqn. (7). Finally, since the deduction after Eqn. (7) does not rely on the rank-constrained problem, we can just follow it and obtain the desired result on rank-constrained problem. The proof is completed. □

### B.2 Proof of Corollary 1

*Proof.* Here we use the result in Eqn. (11) in Section B.1:

$$\mathbb{E}\|\boldsymbol{x}^t - \boldsymbol{x}^*\|^2 \leq \theta^t \|\boldsymbol{x}^0 - \boldsymbol{x}^*\|^2 + \frac{\alpha}{12(1-\beta)\ell_s^2}\|\nabla_{\mathcal{I}} f(\boldsymbol{x}^*)\|^2, \ (\forall t),$$

where $\theta \leq 1 - \frac{1}{480\kappa_s}$ and $\beta = \alpha\left(1 - \frac{1}{12\kappa_s}\right)$. Then to achieve $\theta^T\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|^2 \leq \epsilon$, we have

$$T \geq \log_{1/\theta}\left(\frac{\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|^2}{\epsilon}\right)$$

In this way, we have

$$\begin{aligned}\mathbb{E}\|\boldsymbol{x}^T - \boldsymbol{x}^*\|_2 \leq & \sqrt{\mathbb{E}\|\boldsymbol{x}^T - \boldsymbol{x}^*\|_2^2} \leq \sqrt{\theta^T\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|^2 + \frac{\alpha}{12(1-\beta)\ell_s^2}\|\nabla_{\widetilde{\mathcal{I}}} f(\boldsymbol{x}^*)\|^2} \\ \leq & \sqrt{\epsilon + \frac{\alpha}{12(1-\beta)\ell_s^2}\|\nabla_{\widetilde{\mathcal{I}}} f(\boldsymbol{x}^*)\|_2^2} \leq \sqrt{\epsilon} + \frac{\sqrt{\alpha}}{\ell_s\sqrt{12(1-\beta)}}\|\nabla_{\widetilde{\mathcal{I}}} f(\boldsymbol{x}^*)\|.\end{aligned}$$

Therefore, the IFO complexity is

$$\begin{aligned}\tau\left[1 + \frac{1}{\omega} + \cdots + \frac{1}{\omega^{T-1}}\right] = & \tau\frac{(1/\omega)^{\log_{1/\theta}\left(\frac{\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|^2}{\epsilon}\right)} - 1}{1/\omega - 1} \overset{①}{=} \frac{\tau}{1/\omega - 1}\left[\frac{\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|^2}{\epsilon} - 1\right] \\ \overset{②}{\leq} & \frac{6400\alpha B}{\rho_s \ell_s} \cdot \frac{\kappa_s}{\epsilon} \overset{③}{=} \mathcal{O}\left(\frac{\kappa_s}{\epsilon}\right),\end{aligned}$$

where ① uses $\omega = \theta = 1 - \frac{1}{480\kappa_s}$; ② uses $\tau \geq \frac{40\alpha B}{3\rho_s\ell_s\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|^2}$; ③ holds since (1) the parameter $\rho_s$ is the strong convex parameter at sparsity/low-rank level $s$ and thus is not very small since $s$ is much smaller than the feature dimension, (2) we can always scale the problem such that $\rho_s$ is not small. Notice, such a scale does not affect the ratio $B/\ell_s$ since they always scale at the same order. Thus, we have the IFO complexity $\mathcal{O}\left(\frac{\kappa_s}{\epsilon}\right)$.

On the other hand, we have

$$\log_{1/\theta}\left(\frac{1}{\epsilon}\right) = \frac{\log\left(\frac{1}{\epsilon}\right)}{\log\left(\frac{1}{\theta}\right)} = \frac{\log\left(\frac{1}{\epsilon}\right)}{\log\left(1 + \frac{1}{480\kappa_s - 1}\right)} = \frac{\log\left(\frac{1}{\epsilon}\right)}{\log\left(1 + \frac{1}{480\kappa_s - 1}\right)} \overset{①}{\leq} \mathcal{O}\left(\kappa_s \log\left(\frac{1}{\epsilon}\right)\right),$$

where ① holds since we have $\log(1 + x) \geq \log(2) \cdot x$ for $x \in [0, 1]$. The proof is completed. □

### B.3 Proof of Theorem 2

*Proof.* We first prove the result for sparsity-constrained problem. In this case, the variable $\boldsymbol{x}$ is vector. Let $\mathcal{I} = \mathcal{I}^{t+1} \cup \mathcal{I}^t \cup \mathcal{I}^*$, where $\mathcal{I}^* = \mathsf{supp}(\boldsymbol{x}^*)$, $\mathcal{I}^t = \mathsf{supp}(\boldsymbol{x}^t)$ and $\mathcal{I}^{t+1} = \mathsf{supp}(\boldsymbol{x}^{t+1})$. Recall

that $\boldsymbol{x}^{t+1} = \Phi_k(\boldsymbol{x}^t - \eta \boldsymbol{g}^t)$. Then we have

$$
\begin{aligned}
& f(\boldsymbol{x}^{t+1}) \\
\leq & f(\boldsymbol{x}^t) + \langle \nabla f(\boldsymbol{x}^t), \boldsymbol{x}^{t+1} - \boldsymbol{x}^t \rangle + \frac{\ell_s}{2} \|\boldsymbol{x}^{t+1} - \boldsymbol{x}^t\|^2 \\
\leq & f(\boldsymbol{x}^t) + \langle \boldsymbol{g}^t, \boldsymbol{x}^{t+1} - \boldsymbol{x}^t \rangle + \frac{\ell_s}{2} \|\boldsymbol{x}^{t+1} - \boldsymbol{x}^t\|^2 + \|\nabla f(\boldsymbol{x}^t) - \boldsymbol{g}^t\| \|\boldsymbol{x}^{t+1} - \boldsymbol{x}^t\| \\
= & f(\boldsymbol{x}^t) + \frac{1}{2\eta} \|\boldsymbol{x}_{\mathcal{I}}^{t+1} - \boldsymbol{x}_{\mathcal{I}}^t + \eta \boldsymbol{g}_{\mathcal{I}}^t\|^2 - \frac{\eta \|\boldsymbol{g}_{\mathcal{I}}^t\|^2}{2} - \frac{1 - \eta \ell_s}{2\eta} \|\boldsymbol{x}^{t+1} - \boldsymbol{x}^t\|^2 + \|\nabla f(\boldsymbol{x}^t) - \boldsymbol{g}^t\| \|\boldsymbol{x}^{t+1} - \boldsymbol{x}^t\| \\
= & f(\boldsymbol{x}^t) + \frac{1}{2\eta} \|\boldsymbol{x}_{\mathcal{I}}^{t+1} - \boldsymbol{x}_{\mathcal{I}}^t + \eta \boldsymbol{g}_{\mathcal{I}}^t\|^2 - \frac{\eta \|\boldsymbol{g}_{\mathcal{I} \setminus (\mathcal{I}^t \cup \mathcal{I}^*)}^t\|^2}{2} - \frac{\eta \|\boldsymbol{g}_{\mathcal{I}^t \cup \mathcal{I}^*}^t\|^2}{2} - \frac{1 - \eta \ell_s}{2\eta} \|\boldsymbol{x}^{t+1} - \boldsymbol{x}^t\|^2 \\
& + \|\nabla f(\boldsymbol{x}^t) - \boldsymbol{g}^t\| \|\boldsymbol{x}^{t+1} - \boldsymbol{x}^t\| \\
= & f(\boldsymbol{x}^t) + \frac{1}{2\eta} \left( \|\boldsymbol{x}_{\mathcal{I}}^{t+1} - \boldsymbol{x}_{\mathcal{I}}^t + \eta \boldsymbol{g}_{\mathcal{I}}^t\|^2 - \eta^2 \|\boldsymbol{g}_{\mathcal{I} \setminus (\mathcal{I}^t \cup \mathcal{I}^*)}^t\|^2 \right) - \frac{\eta \|\boldsymbol{g}_{\mathcal{I}^t \cup \mathcal{I}^*}^t\|^2}{2} - \frac{1 - \eta \ell_s}{2\eta} \|\boldsymbol{x}^{t+1} - \boldsymbol{x}^t\|^2 \\
& + \|\nabla f(\boldsymbol{x}^t) - \boldsymbol{g}^t\| \|\boldsymbol{x}^{t+1} - \boldsymbol{x}^t\|.
\end{aligned} \tag{13}
$$

Now we bound the second term $\|\boldsymbol{x}_{\mathcal{I}}^{t+1} - \boldsymbol{x}_{\mathcal{I}}^t + \eta \boldsymbol{g}_{\mathcal{I}}^t\|^2 - \eta^2 \|\boldsymbol{g}_{\mathcal{I} \setminus (\mathcal{I}^t \cup \mathcal{I}^*)}^t\|^2$. Here we adopt similar strategy in [2]. Since we have $\mathcal{I} \setminus (\mathcal{I}^t \cup \mathcal{I}^*) = \mathcal{I}^{t+1} \setminus (\mathcal{I}^t \cup \mathcal{I}^*) \subseteq \mathcal{I}^{t+1}$, then we can establish $\boldsymbol{x}_{\mathcal{I} \setminus (\mathcal{I}^t \cup \mathcal{I}^*)}^{t+1} = \boldsymbol{x}_{\mathcal{I} \setminus (\mathcal{I}^t \cup \mathcal{I}^*)}^t - \eta \boldsymbol{g}_{\mathcal{I} \setminus (\mathcal{I}^t \cup \mathcal{I}^*)}^t$. On the other hand, we have $\boldsymbol{x}_{\mathcal{I} \setminus (\mathcal{I}^t \cup \mathcal{I}^*)}^t = 0$ which further yields $\boldsymbol{x}_{\mathcal{I} \setminus (\mathcal{I}^t \cup \mathcal{I}^*)}^{t+1} = -\eta \boldsymbol{g}_{\mathcal{I} \setminus (\mathcal{I}^t \cup \mathcal{I}^*)}^t$. Next, we choose a set $\mathcal{R} \subseteq \mathcal{I}^t \setminus \mathcal{I}^{t+1}$ such that $|\mathcal{R}| = |\mathcal{I}^{t+1} \setminus (\mathcal{I}^t \cup \mathcal{I}^*)|$. We can find such a set $\mathcal{R}$ because we have $|\mathcal{I}^{t+1} \setminus (\mathcal{I}^t \cup \mathcal{I}^*)| = |\mathcal{I}^t \setminus \mathcal{I}^{t+1}| - |(\mathcal{I}^{t+1} \cap \mathcal{I}^*) \setminus \mathcal{I}^t|$. Besides, since $\boldsymbol{x}^{t+1} = \Phi_k(\boldsymbol{x}^t - \eta \boldsymbol{g}^t)$, we can establish:

$$
\eta^2 \|\boldsymbol{g}_{\mathcal{I} \setminus (\mathcal{I}^t \cup \mathcal{I}^*)}^t\|^2 = \|\boldsymbol{x}_{\mathcal{I} \setminus (\mathcal{I}^t \cup \mathcal{I}^*)}^{t+1}\|^2 \geq \|\boldsymbol{x}_{\mathcal{R}}^t - \eta \boldsymbol{g}_{\mathcal{R}}^t\|^2. \tag{14}
$$

Then combining Eqn. (14) and the fact that $\boldsymbol{x}_{\mathcal{R}}^{t+1} = 0$, we have

$$
\begin{aligned}
\|\boldsymbol{x}_{\mathcal{I}}^{t+1} - \boldsymbol{x}_{\mathcal{I}}^t + \eta \boldsymbol{g}_{\mathcal{I}}^t\|^2 - \eta^2 \|\boldsymbol{g}_{\mathcal{I} \setminus (\mathcal{I}^t \cup \mathcal{I}^*)}^t\|^2 \leq & \|\boldsymbol{x}_{\mathcal{I}}^{t+1} - \boldsymbol{x}_{\mathcal{I}}^t + \eta \boldsymbol{g}_{\mathcal{I}}^t\|^2 - \|\boldsymbol{x}_{\mathcal{R}}^{t+1} - \boldsymbol{x}_{\mathcal{R}}^t + \eta \boldsymbol{g}_{\mathcal{R}}^t\|^2 \\
= & \|\boldsymbol{x}_{\mathcal{I} \setminus \mathcal{R}}^{t+1} - \boldsymbol{x}_{\mathcal{I} \setminus \mathcal{R}}^t + \eta \boldsymbol{g}_{\mathcal{I} \setminus \mathcal{R}}^t\|^2.
\end{aligned} \tag{15}
$$

Next, we bound the size of $\mathcal{I} \setminus \mathcal{R}$ as $|\mathcal{I} \setminus \mathcal{R}| \leq |\mathcal{I}^{t+1}| + |(\mathcal{I}^t \setminus \mathcal{I}^{t+1}) \setminus \mathcal{R}| + |\mathcal{I}^*| \leq k + |(\mathcal{I}^{t+1} \cap \mathcal{I}^*) \setminus \mathcal{I}^t| + k^* \leq k + 2k^*$. Also, since $\mathcal{I}^{t+1} \subseteq (\mathcal{I} \setminus \mathcal{R})$, we have $\boldsymbol{x}_{\mathcal{I} \setminus \mathcal{R}}^{t+1} = \Phi_k(\boldsymbol{x}_{\mathcal{I} \setminus \mathcal{R}}^t - \eta \boldsymbol{g}_{\mathcal{I} \setminus \mathcal{R}}^t)$. By combining Eqn. (15) and Lemma 2, we can obtain

$$
\begin{aligned}
& \|\boldsymbol{x}_{\mathcal{I}}^{t+1} - \boldsymbol{x}_{\mathcal{I}}^t + \eta \boldsymbol{g}_{\mathcal{I}}^t\|^2 - \eta^2 \|\boldsymbol{g}_{\mathcal{I} \setminus (\mathcal{I}^t \cup \mathcal{I}^*)}^t\|^2 \\
\leq & \frac{2k^*}{k + k^*} \|\boldsymbol{x}_{\mathcal{I}^* \setminus \mathcal{R}}^* - \boldsymbol{x}_{\mathcal{I} \setminus \mathcal{R}}^t + \eta \boldsymbol{g}_{\mathcal{I} \setminus \mathcal{R}}^t\|^2 \\
\leq & \frac{2k^*}{k + k^*} \|\boldsymbol{x}_{\mathcal{I}^*}^* - \boldsymbol{x}_{\mathcal{I}}^t + \eta \boldsymbol{g}_{\mathcal{I}}^t\|^2 \\
= & \frac{2k^*}{k + k^*} \left( \|\boldsymbol{x}^* - \boldsymbol{x}^t\|^2 + 2\eta \langle \boldsymbol{g}^t, \boldsymbol{x}^* - \boldsymbol{x}^t \rangle + \eta^2 \|\boldsymbol{g}_{\mathcal{I}}^t\|^2 \right) \\
\overset{①}{=} & \frac{2k^*}{k + k^*} \left( \|\boldsymbol{x}^* - \boldsymbol{x}^t\|^2 + 2\eta \langle \nabla f(\boldsymbol{x}^t), \boldsymbol{x}^* - \boldsymbol{x}^t \rangle + \eta^2 \|\boldsymbol{g}_{\mathcal{I}}^t\|^2 \right) + \xi \langle \boldsymbol{g}^t - \nabla f(\boldsymbol{x}^t), \boldsymbol{x}^* - \boldsymbol{x}^t \rangle \\
\leq & \frac{2k^*}{k + k^*} \left[ \|\boldsymbol{x}^* - \boldsymbol{x}^t\|^2 + 2\eta \left( f(\boldsymbol{x}^*) - f(\boldsymbol{x}^t) - \frac{\rho_s}{2} \|\boldsymbol{x}^* - \boldsymbol{x}^t\|^2 \right) + \eta^2 \|\boldsymbol{g}_{\mathcal{I}}^t\|^2 \right] + \xi \langle \boldsymbol{g}^t - \nabla f(\boldsymbol{x}^t), \boldsymbol{x}^* - \boldsymbol{x}^t \rangle \\
= & \frac{4\eta k^*}{k + k^*} (f(\boldsymbol{x}^*) - f(\boldsymbol{x}^t)) + \frac{2(1 - \eta \rho_s) k^*}{k + k^*} \|\boldsymbol{x}^* - \boldsymbol{x}^t\|^2 + \frac{2\eta^2 k^*}{k + k^*} \|\boldsymbol{g}_{\mathcal{I}}^t\|^2 + \xi \langle \boldsymbol{g}^t - \nabla f(\boldsymbol{x}^t), \boldsymbol{x}^* - \boldsymbol{x}^t \rangle \\
= & \frac{4\eta k^*}{k + k^*} (f(\boldsymbol{x}^*) - f(\boldsymbol{x}^t)) + \frac{2(1 - \eta \rho_s) k^*}{k + k^*} \|\boldsymbol{x}^* - \boldsymbol{x}^t\|^2 + \frac{2\eta^2 k^*}{k + k^*} \|\boldsymbol{g}_{\mathcal{I} \setminus (\mathcal{I}^t \cup \mathcal{I}^*)}^t\|^2 + \frac{2\eta^2 k^*}{k + k^*} \|\boldsymbol{g}_{\mathcal{I}^t \cup \mathcal{I}^*}^t\|^2 \\
& + \xi \langle \boldsymbol{g}^t - \nabla f(\boldsymbol{x}^t), \boldsymbol{x}^* - \boldsymbol{x}^t \rangle,
\end{aligned}
$$

where $\xi = \frac{4\eta k^*}{k+k^*}$ in ①. By substituting the above inequality into Eqn. (13), we can further obtain

$$
\begin{aligned}
& f(\boldsymbol{x}^{t+1}) \\
\leq & f(\boldsymbol{x}^t) + \frac{2k^*}{k+k^*}(f(\boldsymbol{x}^*) - f(\boldsymbol{x}^t)) + \frac{(1-\eta\rho_s)k^*}{\eta(k+k^*)}\|\boldsymbol{x}^* - \boldsymbol{x}^t\|^2 + \frac{\eta k^*}{k+k^*}\|\boldsymbol{g}^t_{\mathcal{I}\setminus(\mathcal{I}^t\cup\mathcal{I}^*)}\|^2 \\
& + \left(\frac{\eta k^*}{k+k^*} - \frac{\eta}{2}\right)\|\boldsymbol{g}^t_{\mathcal{I}^t\cup\mathcal{I}^*}\|^2 - \frac{1-\eta\ell_s}{2\eta}\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^t\|^2 + \frac{\xi}{2\eta}\langle \boldsymbol{g}^t - \nabla f(\boldsymbol{x}^t), \boldsymbol{x}^* - \boldsymbol{x}^t\rangle \\
& + \|\nabla f(\boldsymbol{x}^t) - \boldsymbol{g}^t\|\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^t\| \\
\overset{①}{\leq} & f(\boldsymbol{x}^t) + \frac{2k^*}{k+k^*}(f(\boldsymbol{x}^*) - f(\boldsymbol{x}^t)) + \frac{(1-\eta\rho_s)k^*}{\eta(k+k^*)}\|\boldsymbol{x}^* - \boldsymbol{x}^t\|^2 - \left[\frac{1-\eta\ell_s}{2\eta} - \frac{k^*}{\eta(k+k^*)}\right]\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^t\|^2 \\
& + \left(\frac{\eta k^*}{k+k^*} - \frac{\eta}{2}\right)\|\boldsymbol{g}^t_{\mathcal{I}^t\cup\mathcal{I}^*}\|^2 + \frac{\xi}{2\eta}\langle \boldsymbol{g}^t - \nabla f(\boldsymbol{x}^t), \boldsymbol{x}^* - \boldsymbol{x}^t\rangle + \|\nabla f(\boldsymbol{x}^t) - \boldsymbol{g}^t\|\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^t\| \\
\overset{②}{\leq} & f(\boldsymbol{x}^t) + \frac{2k^*}{k+k^*}(f(\boldsymbol{x}^*) - f(\boldsymbol{x}^t)) + \frac{(1-\eta\rho_s)k^*}{\eta(k+k^*)}\|\boldsymbol{x}^* - \boldsymbol{x}^t\|^2 - \left(\frac{\eta}{2} - \frac{\eta k^*}{k+k^*}\right)\|\boldsymbol{g}^t_{\mathcal{I}^t\cup\mathcal{I}^*}\|^2 \\
& + \frac{\xi}{2\eta}\langle \boldsymbol{g}^t - \nabla f(\boldsymbol{x}^t), \boldsymbol{x}^* - \boldsymbol{x}^t\rangle + \frac{\eta(k+k^*)}{2\left((1-\eta\ell_s)k - (1+\eta\ell_s)k^*\right)}\|\nabla f(\boldsymbol{x}^t) - \boldsymbol{g}^t\|^2,
\end{aligned}
$$

where in ① we used $\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^t\| \geq \eta\|\boldsymbol{g}^t_{\mathcal{I}\setminus(\mathcal{I}^t\cup\mathcal{I}^*)}\|$, and in ② we have used the basic inequality $ab \leq \frac{a^2}{4c} + cb^2, \forall c > 0$. By invoking Lemma 5 in the above we further get

$$
\begin{aligned}
& f(\boldsymbol{x}^{t+1}) \\
\leq & f(\boldsymbol{x}^t) + \frac{2k^*}{k+k^*}\left(1 + \frac{2(1-\eta\rho_s)}{\eta\rho_s}\right)(f(\boldsymbol{x}^*) - f(\boldsymbol{x}^t)) - \left(\frac{\eta}{2} - \frac{(\eta^2\rho_s^2 + 8(1-\eta\rho_s))k^*}{\eta\rho_s^2(k+k^*)}\right)\|\boldsymbol{g}^t_{\mathcal{I}^t\cup\mathcal{I}^*}\|^2 \\
& + \frac{\xi}{2\eta}\langle \boldsymbol{g}^t - \nabla f(\boldsymbol{x}^t), \boldsymbol{x}^* - \boldsymbol{x}^t\rangle + \left(\frac{\eta(k+k^*)}{2\left((1-\eta\ell_s)k - (1+\eta\ell_s)k^*\right)} + \frac{8(1-\eta\rho_s)k^*}{\eta\rho_s^2(k+k^*)}\right)\|\nabla f(\boldsymbol{x}^t) - \boldsymbol{g}^t\|^2.
\end{aligned}
$$

Let us now consider $\eta = \frac{1}{2\ell_s}$ in the above inequality, which leads to

$$
\begin{aligned}
f(\boldsymbol{x}^{t+1}) \leq & f(\boldsymbol{x}^t) + \frac{2(4\ell_s - \rho_s)k^*}{\rho_s(k+k^*)}(f(\boldsymbol{x}^*) - f(\boldsymbol{x}^t)) - \left(\frac{1}{4\ell_s} - \frac{(\rho_s^2 - 16\rho_s\ell_s + 32\ell_s^2)k^*}{2\ell_s\rho_s^2(k+k^*)}\right)\|\boldsymbol{g}^t_{\mathcal{I}^t\cup\mathcal{I}^*}\|^2 \\
& + \ell_s\xi\langle \boldsymbol{g}^t - \nabla f(\boldsymbol{x}^t), \boldsymbol{x}^* - \boldsymbol{x}^t\rangle + \left(\frac{k+k^*}{2\ell_s(k-3k^*)} + \frac{8(2\ell_s - \rho_s)k^*}{\rho_s^2(k+k^*)}\right)\|\nabla f(\boldsymbol{x}^t) - \boldsymbol{g}^t\|^2.
\end{aligned}
$$

Since $k \geq \left(1 + \frac{64\ell_s^2}{\rho_s^2}\right)k^*$, with algebra manipulation we can further show that

$$
\begin{aligned}
f(\boldsymbol{x}^{t+1}) \leq & f(\boldsymbol{x}^t) + \frac{\rho_s}{8\ell_s}(f(\boldsymbol{x}^*) - f(\boldsymbol{x}^t)) + \ell_s\xi\langle \boldsymbol{g}^t - \nabla f(\boldsymbol{x}^t), \boldsymbol{x}^* - \boldsymbol{x}^t\rangle + \left(\frac{5}{4\ell_s} + \frac{8\ell_s}{\rho_s^2}\right)\|\nabla f(\boldsymbol{x}^t) - \boldsymbol{g}^t\|^2 \\
\leq & f(\boldsymbol{x}^t) + \frac{\rho_s}{8\ell_s}(f(\boldsymbol{x}^*) - f(\boldsymbol{x}^t)) + \ell_s\xi\langle \boldsymbol{g}^t - \nabla f(\boldsymbol{x}^t), \boldsymbol{x}^* - \boldsymbol{x}^t\rangle + \frac{37\ell_s}{4\rho_s^2}\|\nabla f(\boldsymbol{x}^t) - \boldsymbol{g}^t\|^2.
\end{aligned}
$$

Taking expectation (conditioned on $\boldsymbol{x}^t$) on both sides of the above we arrive at

$$
\begin{aligned}
\mathbb{E}[f(\boldsymbol{x}^{t+1}) \mid \boldsymbol{x}^t] \leq & f(\boldsymbol{x}^t) + \frac{\rho_s}{8\ell_s}(f(\boldsymbol{x}^*) - f(\boldsymbol{x}^t)) + \frac{37\ell_s}{4\rho_s^2}\mathbb{E}[\|\nabla f(\boldsymbol{x}^t) - \boldsymbol{g}^t\|^2 \mid \boldsymbol{x}^t] \\
\overset{①}{\leq} & f(\boldsymbol{x}^t) + \frac{\rho_s}{8\ell_s}(f(\boldsymbol{x}^*) - f(\boldsymbol{x}^t)) + \frac{37\ell_s}{4\rho_s^2 s_t}B,
\end{aligned}
$$

where ① uses Lemma 3, in which $B = \max_t B_t$ and $B_t = \frac{1}{n-1}\sum_{i=1}^n \|\nabla f_i(\boldsymbol{x}^t) - \nabla f(\boldsymbol{x}^t)\|^2$. By further taking expectation on $\boldsymbol{x}^t$ we obtain

$$
\mathbb{E}[f(\boldsymbol{x}^{t+1}) - f(\boldsymbol{x}^*)] \leq \left(1 - \frac{\rho_s}{8\ell_s}\right)\mathbb{E}[f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*)] + \frac{37\ell_s}{4\rho_s^2 s_t}B.
$$

We further set $\beta = 1 - \frac{1}{8\kappa_s}$ and $s_t = \tau/\omega^t$, and then assume that $\tau$ is large enough such that

$$\gamma := \frac{37\ell_s B}{4\tau\rho_s^2} \leq \delta[f(\boldsymbol{x}^0) - f(\boldsymbol{x}^*)], \tag{16}$$

where $\delta$ is a positive constant and will be discussed later. Now we use mathematical induction to prove

$$\mathbb{E}[f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*)] \leq \theta^t[f(\boldsymbol{x}^0) - f(\boldsymbol{x}^*)], \ (\forall t), \tag{17}$$

where $\theta < 1$ is a constant and will be given below.

Obviously, when $t = 0$, Eqn. (17) holds. Now assume that for all $k \leq t$, Eqn. (17) holds. Then for $k = t + 1$, we have

$$\begin{aligned}\mathbb{E}[f(\boldsymbol{x}^{t+1}) - f(\boldsymbol{x}^*)] \leq & \beta\mathbb{E}[f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*)] + \delta\omega^t[f(\boldsymbol{x}^0) - f(\boldsymbol{x}^*)] \\ \leq & (\beta\theta^t + \delta\omega^t)[f(\boldsymbol{x}^0) - f(\boldsymbol{x}^*)] \overset{①}{\leq} \theta^{t+1}[f(\boldsymbol{x}^0) - f(\boldsymbol{x}^*)],\end{aligned}$$

where ① holds since we let

$$\omega = \theta = \beta + \delta. \tag{18}$$

This means that if Eqn. (16) holds, then Eqn. (17) always holds. So the conclusion holds.

Then let $\tau \geq \frac{148B\kappa_s^2}{\rho_s[f(\boldsymbol{x}^0) - f(\boldsymbol{x}^*)]}$ which gives $\delta \leq \frac{1}{16\kappa_s}$. We have

$$\omega = \theta = \beta + \delta \leq 1 - \frac{1}{8\kappa_s} + \frac{1}{16\kappa_s} = 1 - \frac{1}{16\kappa_s}.$$

Therefore, we have

$$\mathbb{E}[f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*)] \leq \left(1 - \frac{1}{16\kappa_s}\right)^t [f(\boldsymbol{x}^0) - f(\boldsymbol{x}^*)].$$

The proof is completed. □

# C Proofs for Section 4

## C.1 Proofs of Theorem 3

*Proof.* Here we also first consider the sparsity-constrained problem. Let us consider $\mathcal{I} = \mathcal{I}^{t+1} \cup \mathcal{I}^t \cup \mathcal{I}^{t-1} \cup \mathcal{I}^*$ and $\boldsymbol{y}^{t+1} = \boldsymbol{x}^t - \eta\boldsymbol{g}_{\mathcal{I}}^t + \nu(\boldsymbol{x}^t - \boldsymbol{x}^{t-1})$. Since $f(\boldsymbol{x})$ is twice differentiable, we know that there exists $\boldsymbol{z}$ in the line between $\boldsymbol{x}^t$ and $\boldsymbol{x}^*$ such that

$$\begin{aligned}\boldsymbol{y}_{\mathcal{I}}^{t+1} - \boldsymbol{x}^* = & \boldsymbol{x}^t - \boldsymbol{x}^* - \eta\nabla_{\mathcal{I}} f(\boldsymbol{x}^t) + \eta\nabla_{\mathcal{I}} f(\boldsymbol{x}^*) + \nu(\boldsymbol{x}^t - \boldsymbol{x}^{t-1}) + \eta(\nabla_{\mathcal{I}} f(\boldsymbol{x}^t) - \boldsymbol{g}_{\mathcal{I}}^t) - \eta\nabla_{\mathcal{I}} f(\boldsymbol{x}^*) \\ = & ((1+\nu)\boldsymbol{I} - \eta\nabla_{\mathcal{I}\mathcal{I}}^2 f(\boldsymbol{z}))(\boldsymbol{x}^t - \boldsymbol{x}^*) - \nu(\boldsymbol{x}^{t-1} - \boldsymbol{x}^*) + \eta(\nabla_{\mathcal{I}} f(\boldsymbol{x}^t) - \boldsymbol{g}_{\mathcal{I}}^t) - \eta\nabla_{\mathcal{I}} f(\boldsymbol{x}^*).\end{aligned} \tag{19}$$

Since the above iterate depends on the previous two iterates, we consider the following three-term recurrence in matrix form:

$$\begin{bmatrix} \boldsymbol{y}_{\mathcal{I}}^{t+1} - \boldsymbol{x}^* \\ \boldsymbol{x}^t - \boldsymbol{x}^* \end{bmatrix} = \begin{bmatrix} (1+\nu)\boldsymbol{I} - \eta\nabla_{\mathcal{I}\mathcal{I}}^2 f(\boldsymbol{z}) & -\nu\boldsymbol{I} \\ \boldsymbol{I} & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{x}^t - \boldsymbol{x}^* \\ \boldsymbol{x}^{t-1} - \boldsymbol{x}^* \end{bmatrix} + \eta \begin{bmatrix} \nabla_{\mathcal{I}} f(\boldsymbol{x}^t) - \boldsymbol{g}_{\mathcal{I}}^t - \nabla_{\mathcal{I}} f(\boldsymbol{x}^*) \\ 0 \end{bmatrix}.$$

Since $\boldsymbol{x}^{t+1}=\Phi_k(\boldsymbol{y}^{t+1})=\Phi_k(\boldsymbol{y}_{\mathcal{I}}^{t+1})$, based on Lemma 1 we get $\|\boldsymbol{x}^{t+1}-\boldsymbol{x}^*\|\leq\alpha\|\boldsymbol{y}_{\mathcal{I}}^{t+1}-\boldsymbol{x}^*\|$, where $\alpha=1+\frac{2\sqrt{k^*}}{\sqrt{k-k^*}}$. Let $\widehat{\mathcal{I}}=\mathsf{supp}(\Phi_{3k}(\nabla f(\boldsymbol{x}^*)))\cup\mathsf{supp}(\boldsymbol{x}^*)$ and $\widehat{s}=3k+k^*$. Then

$$
\begin{aligned}
&\mathbb{E}\left\|\begin{bmatrix}\boldsymbol{x}^{t+1}-\boldsymbol{x}^*\\ \boldsymbol{x}^t-\boldsymbol{x}^*\end{bmatrix}\right\|\\
\leq&\alpha\mathbb{E}\left\|\begin{bmatrix}\boldsymbol{y}_{\mathcal{I}}^{t+1}-\boldsymbol{x}^*\\ \boldsymbol{x}^t-\boldsymbol{x}^*\end{bmatrix}\right\|\\
\leq&\alpha\mathbb{E}\left\|\begin{bmatrix}(1+\nu)\boldsymbol{I}-\eta\nabla^2_{\mathcal{I}\mathcal{I}}f(\boldsymbol{z}) & -\nu\boldsymbol{I}\\ \boldsymbol{I} & 0\end{bmatrix}\right\|\left\|\begin{bmatrix}\boldsymbol{x}^t-\boldsymbol{x}^*\\ \boldsymbol{x}^{t-1}-\boldsymbol{x}^*\end{bmatrix}\right\|+\eta\mathbb{E}\left\|\begin{bmatrix}\nabla_{\mathcal{I}}f(\boldsymbol{x}^t)-\boldsymbol{g}_{\mathcal{I}}^t-\nabla_{\mathcal{I}}f(\boldsymbol{x}^*)\\ 0\end{bmatrix}\right\|\\
\leq&\alpha\mathbb{E}\left\|\begin{bmatrix}(1+\nu)\boldsymbol{I}_{\mathcal{I}\mathcal{I}}-\eta\nabla^2_{\mathcal{I}\mathcal{I}}f(\boldsymbol{z}) & -\nu\boldsymbol{I}_{\mathcal{I}\mathcal{I}}\\ \boldsymbol{I}_{\mathcal{I}\mathcal{I}} & 0\end{bmatrix}\right\|\left\|\begin{bmatrix}\boldsymbol{x}^t-\boldsymbol{x}^*\\ \boldsymbol{x}^{t-1}-\boldsymbol{x}^*\end{bmatrix}\right\|+\eta\mathbb{E}\|\nabla_{\mathcal{I}}f(\boldsymbol{x}^t)-\boldsymbol{g}_{\mathcal{I}}^t\|+\eta\|\nabla_{\mathcal{I}}f(\boldsymbol{x}^*)\|\\
\overset{①}{\leq}&\alpha\frac{\sqrt{\ell_{\widehat{s}}}-\sqrt{\rho_{\widehat{s}}}}{\sqrt{\ell_{\widehat{s}}}+\sqrt{\rho_{\widehat{s}}}}\mathbb{E}\left\|\begin{bmatrix}\boldsymbol{x}^t-\boldsymbol{x}^*\\ \boldsymbol{x}^{t-1}-\boldsymbol{x}^*\end{bmatrix}\right\|+\eta\mathbb{E}\|\nabla_{\mathcal{I}}f(\boldsymbol{x}^t)-\boldsymbol{g}_{\mathcal{I}}^t\|+\eta\|\nabla_{\mathcal{I}}f(\boldsymbol{x}^*)\|\\
\overset{②}{\leq}&\alpha\left(1-\sqrt{\frac{\rho_{\widehat{s}}}{\ell_{\widehat{s}}}}\right)\mathbb{E}\left\|\begin{bmatrix}\boldsymbol{x}^t-\boldsymbol{x}^*\\ \boldsymbol{x}^{t-1}-\boldsymbol{x}^*\end{bmatrix}\right\|+\eta\mathbb{E}\|\nabla_{\mathcal{I}}f(\boldsymbol{x}^t)-\boldsymbol{g}_{\mathcal{I}}^t\|+\eta\|\nabla_{\mathcal{I}}f(\boldsymbol{x}^*)\|\\
\overset{③}{\leq}&\left(1-\frac{1}{2}\sqrt{\frac{\rho_{\widehat{s}}}{\ell_{\widehat{s}}}}\right)\mathbb{E}\left\|\begin{bmatrix}\boldsymbol{x}^t-\boldsymbol{x}^*\\ \boldsymbol{x}^{t-1}-\boldsymbol{x}^*\end{bmatrix}\right\|+\eta\mathbb{E}\|\nabla_{\mathcal{I}}f(\boldsymbol{x}^t)-\boldsymbol{g}_{\mathcal{I}}^t\|+\eta\|\nabla_{\mathcal{I}}f(\boldsymbol{x}^*)\|\\
\overset{④}{\leq}&\left(1-\frac{1}{2}\sqrt{\frac{\rho_{\widehat{s}}}{\ell_{\widehat{s}}}}\right)\mathbb{E}\left\|\begin{bmatrix}\boldsymbol{x}^t-\boldsymbol{x}^*\\ \boldsymbol{x}^{t-1}-\boldsymbol{x}^*\end{bmatrix}\right\|+\eta\sqrt{\frac{B}{s_t}}+\eta\|\nabla_{\widehat{\mathcal{I}}}f(\boldsymbol{x}^*)\|,
\end{aligned}
\tag{20}
$$

where ① follows from Lemma 6 with $\eta=\frac{4}{(\sqrt{\rho_{\widehat{s}}}+\sqrt{\ell_{\widehat{s}}})^2}$ and $\nu=\max\left\{|1-\sqrt{\eta\rho_{\widehat{s}}}|^2,|1-\sqrt{\eta\ell_{\widehat{s}}}|^2\right\}$, ② follows from the fact $\ell_{\widehat{s}}\geq\rho_{\widehat{s}}$, ③ follows from the condition $k\geq\left(1+\frac{16\ell_{\widehat{s}}}{\rho_{\widehat{s}}}\right)k^*$ which implies $\alpha\leq 1+\frac{1}{2}\sqrt{\frac{\rho_{\widehat{s}}}{\ell_{\widehat{s}}}}$, ④ uses Lemma 3 with $B=\max_t B_t$ and $\|\nabla_{\mathcal{I}}f(\boldsymbol{x}^*)\|\leq\|\nabla_{\widehat{\mathcal{I}}}f(\boldsymbol{x}^*)\|$ where $\widehat{\mathcal{I}}=\mathsf{supp}(\Phi_{3k}(\nabla f(\boldsymbol{x}^*)))\cup\mathsf{supp}(\boldsymbol{x}^*)$.

Then we let

$$\beta:=1-\frac{1}{2\sqrt{\kappa_s}}\quad\text{where}\quad\kappa_s=\frac{\ell_s}{\rho_s}.$$

We further set $s_t=\tau/\omega^t$ and assume that $\tau$ is large enough such that

$$\gamma:=\frac{\eta\sqrt{B}}{\sqrt{\tau}}=\frac{4\sqrt{B}}{\sqrt{\tau}(\sqrt{\rho_{\widehat{s}}}+\sqrt{\ell_{\widehat{s}}})^2}\leq\delta(\|\boldsymbol{x}^0-\boldsymbol{x}^*\|+\|\boldsymbol{x}^{-1}-\boldsymbol{x}^*\|),\tag{21}$$

where $\delta$ is a positive constant and will be discussed later. Now we use mathematical induction to prove

$$\mathbb{E}\left\|\begin{bmatrix}\boldsymbol{x}^t-\boldsymbol{x}^*\\ \boldsymbol{x}^{t-1}-\boldsymbol{x}^*\end{bmatrix}\right\|\leq\theta^t\left\|\begin{bmatrix}\boldsymbol{x}^0-\boldsymbol{x}^*\\ \boldsymbol{x}^{-1}-\boldsymbol{x}^*\end{bmatrix}\right\|+\frac{4}{(1-\beta)(\sqrt{\rho_{\widehat{s}}}+\sqrt{\ell_{\widehat{s}}})^2}\|\nabla_{\widehat{\mathcal{I}}}f(\boldsymbol{x}^*)\|,\ (\forall t),\tag{22}$$

where $\theta<1$ is a constant and will be given below.

Obviously, when $t=0$, Eqn. (22) holds. Now assume that for all $k\leq t$, Eqn. (22) holds. Then for $k=t+1$, we have

$$
\begin{aligned}
\mathbb{E}\left\|\begin{bmatrix}\boldsymbol{x}^t-\boldsymbol{x}^*\\ \boldsymbol{x}^{t-1}-\boldsymbol{x}^*\end{bmatrix}\right\|\leq&\beta\mathbb{E}\left\|\begin{bmatrix}\boldsymbol{x}^{t-1}-\boldsymbol{x}^*\\ \boldsymbol{x}^{t-2}-\boldsymbol{x}^*\end{bmatrix}\right\|+\delta\omega^{\frac{t}{2}}\left\|\begin{bmatrix}\boldsymbol{x}^0-\boldsymbol{x}^*\\ \boldsymbol{x}^{-1}-\boldsymbol{x}^*\end{bmatrix}\right\|+\frac{4}{(\sqrt{\rho_{\widehat{s}}}+\sqrt{\ell_{\widehat{s}}})^2}\|\nabla_{\widehat{\mathcal{I}}}f(\boldsymbol{x}^*)\|\\
\leq&(\beta\theta^t+\delta\omega^{\frac{t}{2}})\mathbb{E}\left\|\begin{bmatrix}\boldsymbol{x}^0-\boldsymbol{x}^*\\ \boldsymbol{x}^{-1}-\boldsymbol{x}^*\end{bmatrix}\right\|+\frac{4}{(\sqrt{\rho_{\widehat{s}}}+\sqrt{\ell_{\widehat{s}}})^2}\left[\frac{\beta}{1-\beta}+1\right]\|\nabla_{\widehat{\mathcal{I}}}f(\boldsymbol{x}^*)\|\\
\overset{①}{\leq}&\theta^t\mathbb{E}\left\|\begin{bmatrix}\boldsymbol{x}^0-\boldsymbol{x}^*\\ \boldsymbol{x}^{-1}-\boldsymbol{x}^*\end{bmatrix}\right\|+\frac{4}{(1-\beta)(\sqrt{\rho_{\widehat{s}}}+\sqrt{\ell_{\widehat{s}}})^2}\|\nabla_{\widehat{\mathcal{I}}}f(\boldsymbol{x}^*)\|
\end{aligned}
$$

where ① holds since we let

$$\omega = \theta^2 \quad \text{and} \quad \theta = \beta + \delta. \tag{23}$$

This means that if Eqn. (23) holds, then Eqn. (22) always holds. So the conclusion holds.

Then let $\tau \geq \frac{81B\kappa_s}{4(\sqrt{\rho_s}+\sqrt{\ell_s})^4\|\boldsymbol{x}^0-\boldsymbol{x}^*\|^2}$. Recall $\boldsymbol{x}^{-1} = \boldsymbol{x}^0$, we have

$$\theta = \beta + \delta \leq 1 - \frac{1}{2\sqrt{\kappa_s}} + \frac{4}{9\sqrt{\kappa_s}} = 1 - \frac{1}{18\sqrt{\kappa_s}}.$$

Therefore, we have

$$\begin{aligned}\mathbb{E}\|\boldsymbol{x}^t - \boldsymbol{x}^*\| \leq & 2\theta^t\|\boldsymbol{x}^0 - \boldsymbol{x}^*\| + \frac{4}{(1-\beta)(\sqrt{\rho_{\widehat{s}}} + \sqrt{\ell_{\widehat{s}}})^2}\|\nabla_{\widehat{\mathcal{I}}} f(\boldsymbol{x}^*)\| \\ = & 2\theta^t\|\boldsymbol{x}^0 - \boldsymbol{x}^*\| + \frac{8\sqrt{\kappa_{\widehat{s}}}}{(\sqrt{\rho_{\widehat{s}}} + \sqrt{\ell_{\widehat{s}}})^2}\|\nabla_{\widehat{\mathcal{I}}} f(\boldsymbol{x}^*)\|, \ (\forall t)\end{aligned}$$

where $\beta = 1 - \frac{1}{2\sqrt{\kappa_s}}$ and $\theta \leq 1 - \frac{1}{18\sqrt{\kappa_s}}$ with $\omega = \theta^2$. So the result on sparsity-constrained problem holds.

Now we consider the rank-constrained problem. Since the proof accesses the Hessian, here we need to vectorize the matrix variable $\boldsymbol{x}$. For notation simplicity, we use $\widetilde{\boldsymbol{x}} \in \mathbb{R}^{d_1 d_2}$ to denote the vectorization of $\boldsymbol{x} \in \mathbb{R}^{d_1 \times d_2}$. Assume that the skinny SVDs of $\boldsymbol{x}^t$, and $\boldsymbol{x}^*$ are respectively $\boldsymbol{x}^t = \boldsymbol{U}_t\boldsymbol{\Sigma}_t\boldsymbol{V}_t^T$ and $\boldsymbol{x}^* = \boldsymbol{U}_*\boldsymbol{\Sigma}_*\boldsymbol{V}_*^T$. Then we have $\boldsymbol{x}^t = \boldsymbol{U}_t\boldsymbol{U}_t^T\boldsymbol{x}^t$ and $\boldsymbol{x}^* = \boldsymbol{U}_*\boldsymbol{U}_*^T\boldsymbol{x}^*$. We further define the projection operation $\mathcal{P}_{\boldsymbol{U}}(\boldsymbol{x}) = \boldsymbol{U}\boldsymbol{U}^T\boldsymbol{x}$ which projects $\boldsymbol{x}$ in the subspace spanned by $\boldsymbol{U}$. Then let $\mathcal{I} = \mathcal{I}^* \cup \mathcal{I}^{t-1} \cup \mathcal{I}^t \cup \mathcal{I}^{t+1}$, where $\mathcal{I}^* = \mathsf{supp}(\boldsymbol{x}^*)$, $\mathcal{I}^{t-1} = \mathsf{supp}(\boldsymbol{x}^{t-1})$, $\mathcal{I}^t = \mathsf{supp}(\boldsymbol{x}^t)$ and $\mathcal{I}^{t+1} = \mathsf{supp}(\boldsymbol{x}^{t+1})$. The notation $\boldsymbol{x}^t_{\mathcal{I}^t} = \boldsymbol{U}_t\boldsymbol{U}_t^T\boldsymbol{x}^t = \mathcal{P}_{\boldsymbol{U}^t}(\boldsymbol{x}^t)$. For the specifical notation $\boldsymbol{y}_{\mathcal{I}}$, it denotes $\boldsymbol{y}_{\mathcal{I}} = \mathcal{P}_{\mathcal{I}}(\boldsymbol{y}_{\mathcal{I}}) = (\boldsymbol{U}_{t-1}\boldsymbol{U}_{t-1}^T + \boldsymbol{U}_t\boldsymbol{U}_t^T + \boldsymbol{U}_{t+1}\boldsymbol{U}_{t+1}^T + \boldsymbol{U}_*\boldsymbol{U}_*^T - \boldsymbol{U}_{t-1}\boldsymbol{U}_{t-1}^T\boldsymbol{U}_t\boldsymbol{U}_t^T\boldsymbol{U}_{t+1}\boldsymbol{U}_{t+1}^T - \boldsymbol{U}_{t-1}\boldsymbol{U}_{t-1}^T\boldsymbol{U}_t\boldsymbol{U}_t^T\boldsymbol{U}_*\boldsymbol{U}_*^T - \boldsymbol{U}_{t-1}\boldsymbol{U}_{t-1}^T\boldsymbol{U}_{t+1}\boldsymbol{U}_{t+1}^T\boldsymbol{U}_*\boldsymbol{U}_*^T - \boldsymbol{U}_t\boldsymbol{U}_t^T\boldsymbol{U}_{t+1}\boldsymbol{U}_{t+1}^T\boldsymbol{U}_*\boldsymbol{U}_*^T + \boldsymbol{U}_{t-1}\boldsymbol{U}_{t-1}^T\boldsymbol{U}_t\boldsymbol{U}_t^T\boldsymbol{U}_{t+1}\boldsymbol{U}_{t+1}^T\boldsymbol{U}_*\boldsymbol{U}_*^T)\boldsymbol{g}^t$ where $\mathcal{P}_{\mathcal{I}}$ is a projection. Here $\boldsymbol{y}$ can be $\boldsymbol{x}$, $\boldsymbol{g}^t$ and $\nabla f(\boldsymbol{x})$.

Then we also prove Eqn. (19) holds. Let $\widetilde{\boldsymbol{y}}^t$, $\widetilde{\boldsymbol{x}}^*$, $\widetilde{\boldsymbol{x}}^t$ and $\nabla\widetilde{f}(\boldsymbol{x})$ respectively denotes the vectorization of $\boldsymbol{y}^t$, $\boldsymbol{x}^*$, $\boldsymbol{x}^t$ and $\nabla f(\boldsymbol{x})$. The notation $\widetilde{\boldsymbol{x}}_{\mathcal{I}}$ denotes the vectorization of $\boldsymbol{U}_{\mathcal{I}}\boldsymbol{U}_{\mathcal{I}}^T\widetilde{\boldsymbol{x}}$. Then we have

$$\begin{aligned}\widetilde{\boldsymbol{y}}^{t+1}_{\mathcal{I}} - \widetilde{\boldsymbol{x}}^* = & \widetilde{\boldsymbol{x}}^t - \widetilde{\boldsymbol{x}}^* - \eta\nabla_{\mathcal{I}}\widetilde{f}(\boldsymbol{x}^t) + \eta\nabla_{\mathcal{I}}\widetilde{f}(\boldsymbol{x}^*) + \nu(\widetilde{\boldsymbol{x}}^t - \widetilde{\boldsymbol{x}}^{t-1}) + \eta(\nabla_{\mathcal{I}}\widetilde{f}(\boldsymbol{x}^t) - \widetilde{\boldsymbol{g}}^t_{\mathcal{I}}) - \eta\nabla_{\mathcal{I}}\widetilde{f}(\boldsymbol{x}^*) \\ = & ((1+\nu)\boldsymbol{I} - \eta\boldsymbol{H})(\widetilde{\boldsymbol{x}}^t - \widetilde{\boldsymbol{x}}^*) - \nu(\widetilde{\boldsymbol{x}}^{t-1} - \widetilde{\boldsymbol{x}}^*) + \eta(\nabla_{\mathcal{I}}\widetilde{f}(\widetilde{\boldsymbol{x}}^t) - \widetilde{\boldsymbol{g}}^t_{\mathcal{I}}) - \eta\nabla_{\mathcal{I}}\widetilde{f}(\boldsymbol{x}^*),\end{aligned}$$

where $\boldsymbol{H} = \mathcal{P}_{\mathcal{I}}(\nabla^2\widetilde{f}(\widetilde{\boldsymbol{z}}))$. Here $\nabla^2\widetilde{f}(\widetilde{\boldsymbol{z}})$ comes from the fact that $\nabla_{\mathcal{I}}\widetilde{f}(\boldsymbol{x}^t) - \nabla_{\mathcal{I}}\widetilde{f}(\boldsymbol{x}^*) = \mathcal{P}_{\mathcal{I}}(\nabla\widetilde{f}(\boldsymbol{x}^t) - \nabla\widetilde{f}(\boldsymbol{x}^*)) \overset{①}{=} \mathcal{P}_{\mathcal{I}}\nabla^2\widetilde{f}(\widetilde{\boldsymbol{z}})(\widetilde{\boldsymbol{x}}^t - \widetilde{\boldsymbol{x}}^*)$ in which ① uses the second differential property of $f(\boldsymbol{x})$ and thus there exists a matrix $\boldsymbol{z}$ such that $\nabla\widetilde{f}(\boldsymbol{x}^t) - \nabla\widetilde{f}(\boldsymbol{x}^*) = \nabla^2\widetilde{f}(\widetilde{\boldsymbol{z}})(\widetilde{\boldsymbol{x}}^t - \widetilde{\boldsymbol{x}}^*)$. Then by Assumptions 1 and 2, we have $\rho_{\widehat{s}}\boldsymbol{I} \preceq \mathcal{P}_{\mathcal{I}}(\boldsymbol{H}) \preceq \ell_{\widehat{s}}\boldsymbol{I}$ since $\|\mathcal{P}_{\mathcal{I}}\|_2 \leq 1$. We obtain $\|\mathcal{P}_{\mathcal{I}}\|_2 \leq 1$ since we have $\mathcal{P}_{\mathcal{I}}^T\mathcal{P}_{\mathcal{I}} = \mathcal{P}_{\mathcal{I}}$.

On the other hand, we can follow Eqn. (12) in Section B.1 to prove

$$\begin{aligned}\Phi_k(\boldsymbol{y}^{t+1}_{\mathcal{I}}) = & \boldsymbol{U}_{t+1}\boldsymbol{U}_{t+1}^T[\boldsymbol{x}^t - \eta\boldsymbol{g}^t_{\mathcal{I}} + \nu(\boldsymbol{x}^t - \boldsymbol{x}^{t-1})] \\ = & \boldsymbol{U}_{t+1}\boldsymbol{U}_{t+1}^T[\boldsymbol{x}^t + \nu(\boldsymbol{x}^t - \boldsymbol{x}^{t-1})] + \eta\boldsymbol{U}_{t+1}\boldsymbol{U}_{t+1}^T\mathcal{P}_{\mathcal{I}}(\boldsymbol{g}^t) \\ \overset{①}{=} & \boldsymbol{U}_{t+1}\boldsymbol{U}_{t+1}^T[\boldsymbol{x}^t + \nu(\boldsymbol{x}^t - \boldsymbol{x}^{t-1})] + \eta\boldsymbol{U}_{t+1}\boldsymbol{U}_{t+1}^T\boldsymbol{g}^t = \boldsymbol{x}^{t+1},\end{aligned}$$

where ① plugs $\mathcal{P}_{\mathcal{I}}$ defined above and obtains $\boldsymbol{U}_{t+1}\boldsymbol{U}_{t+1}^T\mathcal{P}_{\mathcal{I}} = \boldsymbol{U}_{t+1}\boldsymbol{U}_{t+1}^T$. Then we apply Lemma 1 to establish

$$\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\|^2 = \|\Phi_k(\boldsymbol{y}^{t+1}_{\mathcal{I}}) - \boldsymbol{x}^*\|^2 \leq \left(1 + \frac{2\sqrt{k^*}}{\sqrt{k-k^*}}\right)\|\boldsymbol{y}^{t+1}_{\mathcal{I}} - \boldsymbol{x}^*\|^2,$$

where $\alpha = 1 + \frac{2\sqrt{k^*}}{\sqrt{k-k^*}}$. Therefore, we can establish

$$\mathbb{E}\left\|\begin{bmatrix}\boldsymbol{x}^{t+1} - \boldsymbol{x}^* \\ \boldsymbol{x}^t - \boldsymbol{x}^*\end{bmatrix}\right\| \leq \alpha\mathbb{E}\left\|\begin{bmatrix}\boldsymbol{y}^{t+1}_{\mathcal{I}} - \boldsymbol{x}^* \\ \boldsymbol{x}^t - \boldsymbol{x}^*\end{bmatrix}\right\|.$$

Then we can establish the first inequality in Eqn. (20). The following proof does not depend the property on rank-constrained problem. So we can just follow the above proof sketch for sparsity-constrained problem to prove the result on the rank-constrained problem. The proof is completed. □

### C.2 Proof of Corollary 2

*Proof.* To achieve $\epsilon$-accurate solution, let

$$2\theta^t\|\boldsymbol{x}^0-\boldsymbol{x}^*\|\leq\sqrt{\epsilon}$$

where $\widetilde{\beta}=1-\frac{1}{2\sqrt{\kappa_{\widehat{s}}}}$ and $\theta\leq 1-\frac{1}{18\sqrt{\kappa_{\widehat{s}}}}$ with $\omega=\theta^2$, we have

$$T\geq\log_{1/\theta}\left(\frac{2\|\boldsymbol{x}^0-\boldsymbol{x}^*\|}{\sqrt{\epsilon}}\right).$$

Therefore, the IFO complexity is

$$\tau\left[1+\frac{1}{\omega}+\cdots+\frac{1}{\omega^{T-1}}\right]=\tau\frac{(1/\omega)^{\log_{1/\theta}\left(\frac{2\|\boldsymbol{x}^0-\boldsymbol{x}^*\|}{\sqrt{\epsilon}}\right)}-1}{1/\omega-1}\overset{①}{=}\frac{\tau}{1/\omega-1}\left[\frac{4\|\boldsymbol{x}^0-\boldsymbol{x}^*\|^2}{\epsilon}-1\right]$$
$$\leq\frac{\tau}{1/\omega-1}\left[\frac{\|\boldsymbol{x}^0-\boldsymbol{x}^*\|^2}{\epsilon}\right]\overset{②}{\leq}\frac{81B\kappa_{\widehat{s}}}{4(\sqrt{\rho_{\widehat{s}}}+\sqrt{\ell_{\widehat{s}}})^4\epsilon}\frac{1}{1/\sqrt{1-\frac{1}{18\sqrt{\kappa_{\widehat{s}}}}}-1}\overset{③}{\leq}\frac{81B\cdot 36\sqrt{\kappa_{\widehat{s}}}}{\rho_{\widehat{s}}(\sqrt{\rho_{\widehat{s}}}+\sqrt{\ell_{\widehat{s}}})^2\epsilon}\frac{\ell_s}{(\sqrt{\rho_{\widehat{s}}}+\sqrt{\ell_{\widehat{s}}})^2}$$
$$\leq\frac{81B\cdot 36\sqrt{\kappa_{\widehat{s}}}}{\rho_{\widehat{s}}(\sqrt{\rho_{\widehat{s}}}+\sqrt{\ell_{\widehat{s}}})^2\epsilon}\frac{\ell_{\widehat{s}}}{(\sqrt{\rho_{\widehat{s}}}+\sqrt{\ell_{\widehat{s}}})^2}=\mathcal{O}\left(\frac{\sqrt{\kappa_{\widehat{s}}}B}{\rho_s\ell_{\widehat{s}}\epsilon}\right)\overset{④}{=}\mathcal{O}\left(\frac{\sqrt{\kappa_{\widehat{s}}}}{\epsilon}\right)$$

where ① uses $\omega=\theta^2$; ② uses $\theta\leq 1-\frac{1}{18\sqrt{\kappa_{\widehat{s}}}}$ and $\tau\geq\frac{81B\kappa_{\widehat{s}}}{4(\sqrt{\rho_{\widehat{s}}}+\sqrt{\ell_{\widehat{s}}})^4\|\boldsymbol{x}^0-\boldsymbol{x}^*\|^2}$; ③ uses $\frac{1}{1/\sqrt{1-\frac{1}{18\sqrt{\kappa_{\widehat{s}}}}}-1}\leq\frac{1}{1-\sqrt{1-\frac{1}{18\sqrt{\kappa_{\widehat{s}}}}}}\leq 36\sqrt{\kappa_{\widehat{s}}}$ since we have $1-\sqrt{1-a}\geq\frac{1}{2}a$ for $a\in(0,1)$; ④ holds since (1) the parameter $\rho_{\widehat{s}}$ is the strong convex parameter at sparsity/low-rank level $s$ and thus is not very small since $\widehat{s}$ is much smaller than the feature dimension, (2) we can always scale the problem such that $\rho_{\widehat{s}}$ is not small. Notice, such a scale does not affect the ratio $B/\ell_{\widehat{s}}$ since they always scale at the same order. Thus, we have the IFO complexity $\mathcal{O}\left(\frac{\sqrt{\kappa_{\widehat{s}}}}{\epsilon}\right)$.

On the other hand, we have

$$\log_{1/\theta}\left(\frac{1}{\sqrt{\epsilon}}\right)=\frac{\log\left(\frac{1}{\sqrt{\epsilon}}\right)}{\log\left(\frac{1}{\theta}\right)}=\frac{\log\left(\frac{1}{\sqrt{\epsilon}}\right)}{\log\left(1+\frac{1}{18\sqrt{\kappa_{\widehat{s}}}-1}\right)}=\frac{\log\left(\frac{1}{\sqrt{\epsilon}}\right)}{\log\left(1+\frac{1}{18\sqrt{\kappa_{\widehat{s}}}-1}\right)}\overset{①}{\leq}\mathcal{O}\left(\sqrt{\kappa_{\widehat{s}}}\log\left(\frac{1}{\sqrt{\epsilon}}\right)\right),$$

where  holds since we have $\log(1+x)\geq\log(2)\cdot x$ for $x\in[0,1]$. The proof is completed. □

## D Proof of Auxiliary Lemmas

### D.1 Proof of Lemma 4

*Proof.* Firstly, for both vector $\boldsymbol{x}$ and matrix variable $\boldsymbol{x}$ we can decompose $\mathbb{E}\|\boldsymbol{g}^t\|^2$ and bound it as follows:

$$\begin{aligned}\mathbb{E}\|\boldsymbol{g}_{\mathcal{I}}^t\|^2=&\mathbb{E}\|\boldsymbol{g}_{\mathcal{I}}^t-\nabla_{\mathcal{I}}f(\boldsymbol{x}^t)+\nabla_{\mathcal{I}}f(\boldsymbol{x}^t)-\nabla_{\mathcal{I}}f(\boldsymbol{x}^*)+\nabla_{\mathcal{I}}f(\boldsymbol{x}^*)\|^2\\ \leq&3\mathbb{E}\|\boldsymbol{g}_{\mathcal{I}}^t-\nabla_{\mathcal{I}}f(\boldsymbol{x}^t)\|^2+3\mathbb{E}\|\nabla_{\mathcal{I}}f(\boldsymbol{x}^t)-\nabla_{\mathcal{I}}f(\boldsymbol{x}^*)\|_2^2+3\|\nabla_{\mathcal{I}}f(\boldsymbol{x}^*)\|^2\\ \overset{①}{\leq}&\frac{3}{s_t}B_t+3\mathbb{E}\|\nabla_{\mathcal{I}}f(\boldsymbol{x}^t)-\nabla_{\mathcal{I}}f(\boldsymbol{x}^*)\|^2+3\|\nabla_{\mathcal{I}}f(\boldsymbol{x}^*)\|^2,\end{aligned}$$

where ① use Lemma 3. Now we bound the second term. We define a function

$$h_i(\boldsymbol{x})=f_i(\boldsymbol{x})-f_i(\boldsymbol{x}^*)-\langle\nabla_{\mathcal{I}}f_i(\boldsymbol{x}^*),\boldsymbol{x}-\boldsymbol{x}^*\rangle.$$

It is easy to check that $\nabla h_i(\boldsymbol{x}^*)=0$, which implies $h_i(\boldsymbol{x}^*)=\min_{\boldsymbol{x}}h_i(\boldsymbol{x})$. In this way, for vector variable $\boldsymbol{x}$, we have

$$\begin{aligned}0=h_i(\boldsymbol{x}^*)\leq\min_{\eta}h_i(\boldsymbol{x}-\eta\nabla_{\mathcal{I}}h_i(\boldsymbol{x}))\leq&\min_{\eta}h_i(\boldsymbol{x})-\eta\langle\nabla h_i(\boldsymbol{x}),\nabla_{\mathcal{I}}h_i(\boldsymbol{x})\rangle+\frac{\eta^2\ell_s}{2}\|\nabla_{\mathcal{I}}h_i(\boldsymbol{x})\|_2^2\\ \overset{①}{=}&\min_{\eta}h_i(\boldsymbol{x})-\eta\|\nabla_{\mathcal{I}}h_i(\boldsymbol{x})\|^2+\frac{\eta^2\ell_s}{2}\|\nabla_{\mathcal{I}}h_i(\boldsymbol{x})\|^2\\ \overset{②}{=}&h_i(\boldsymbol{x})-\frac{1}{2\ell_s}\|\nabla_{\mathcal{I}}h_i(\boldsymbol{x})\|_2^2,\end{aligned}\tag{24}$$

where ① holds since for vector $\boldsymbol{x}$, we have $\langle \nabla h_i(\boldsymbol{x}), \nabla_{\mathcal{I}} h_i(\boldsymbol{x})\rangle = \|\nabla_{\mathcal{I}} h_i(\boldsymbol{x})\|^2$ and ② holds by optimizing $\eta = \frac{1}{\ell_s}$.

Now we consider the matrix variable $\boldsymbol{x}$. Firstly, assume that the skinny SVDs of $\boldsymbol{x}^t$, $\boldsymbol{x}^{t+1}$, $\boldsymbol{x}^*$ are respectively $\boldsymbol{x}^t = \boldsymbol{U}_t\boldsymbol{\Sigma}_t\boldsymbol{V}_t^T$, $\boldsymbol{x}^{t+1} = \boldsymbol{U}_{t+1}\boldsymbol{\Sigma}_{t+1}\boldsymbol{V}_{t+1}^T$ and $\boldsymbol{x}^* = \boldsymbol{U}_*\boldsymbol{\Sigma}_*\boldsymbol{V}_*^T$. Then we have $\boldsymbol{x}^t = \boldsymbol{U}_t\boldsymbol{U}_t^T\boldsymbol{x}^t$, $\boldsymbol{x}^{t+1} = \boldsymbol{U}_{t+1}\boldsymbol{U}_{t+1}^T\boldsymbol{x}^{t+1}$ and $\boldsymbol{x}^* = \boldsymbol{U}_*\boldsymbol{U}_*^T\boldsymbol{x}^*$. We further define the projection operation $\mathcal{P}_{\boldsymbol{U}}(\boldsymbol{x}) = \boldsymbol{U}\boldsymbol{U}^T\boldsymbol{x}$ which projects $\boldsymbol{x}$ in the subspace spanned by $\boldsymbol{U}$. Then let $\boldsymbol{v} = \boldsymbol{x}^t - \eta\boldsymbol{g}_{\mathcal{I}}^t$ and $\mathcal{I} = \mathcal{I}^* \cup \mathcal{I}^t \cup \mathcal{I}^{t+1}$, where $\mathcal{I}^* = \mathsf{supp}(\boldsymbol{x}^*)$, $\mathcal{I}^t = \mathsf{supp}(\boldsymbol{x}^t)$ and $\mathcal{I}^{t+1} = \mathsf{supp}(\boldsymbol{x}^{t+1})$. The notation $\boldsymbol{x}_{\mathcal{I}^t}^t = \boldsymbol{U}_t\boldsymbol{U}_t^T\boldsymbol{x}^t = \mathcal{P}_{\boldsymbol{U}^t}(\boldsymbol{x}^t)$. For the specifical notation $\boldsymbol{g}_{\mathcal{I}}^t$, it denotes $\boldsymbol{g}_{\mathcal{I}}^t = \mathcal{P}_{\mathcal{I}}(\boldsymbol{g}^t) = (\boldsymbol{U}_t\boldsymbol{U}_t^T + \boldsymbol{U}_{t+1}\boldsymbol{U}_{t+1}^T + \boldsymbol{U}_*\boldsymbol{U}_*^T - \boldsymbol{U}_t\boldsymbol{U}_t^T\boldsymbol{U}_{t+1}\boldsymbol{U}_{t+1}^T - \boldsymbol{U}_t\boldsymbol{U}_t^T\boldsymbol{U}_*\boldsymbol{U}_*^T - \boldsymbol{U}_{t+1}\boldsymbol{U}_{t+1}^T\boldsymbol{U}_*\boldsymbol{U}_*^T + \boldsymbol{U}_t\boldsymbol{U}_t^T\boldsymbol{U}_{t+1}\boldsymbol{U}_{t+1}^T\boldsymbol{U}_*\boldsymbol{U}_*^T)\boldsymbol{g}^t$ where $\mathcal{P}_{\mathcal{I}}$ is a projection. Then we also have

$$\langle \nabla h_i(\boldsymbol{x}), \nabla_{\mathcal{I}} h_i(\boldsymbol{x})\rangle = \langle \nabla h_i(\boldsymbol{x}), \mathcal{P}_{\mathcal{I}}(\nabla h_i(\boldsymbol{x}))\rangle \overset{①}{=} \langle \mathcal{P}_{\mathcal{I}}(\nabla h_i(\boldsymbol{x})), \mathcal{P}_{\mathcal{I}}(\nabla h_i(\boldsymbol{x}))\rangle = \|\nabla_{\mathcal{I}} h_i(\boldsymbol{x})\|^2,$$

where ① holds since $\langle \mathcal{P}_{\mathcal{I}}(\nabla_{\mathcal{I}} h_i(\boldsymbol{x})), \mathcal{P}_{\mathcal{I}}(\nabla_{\mathcal{I}} h_i(\boldsymbol{x}))\rangle = \langle \nabla h_i(\boldsymbol{x}), \mathcal{P}_{\mathcal{I}}^T\mathcal{P}_{\mathcal{I}}(\nabla_{\mathcal{I}} h_i(\boldsymbol{x}))\rangle = \langle \nabla h_i(\boldsymbol{x}), \mathcal{P}_{\mathcal{I}}(\nabla_{\mathcal{I}} h_i(\boldsymbol{x}))\rangle$ due to $\mathcal{P}_{\mathcal{I}}^T\mathcal{P}_{\mathcal{I}} = \mathcal{P}_{\mathcal{I}}$.

Thus, Eqn. (24) holds for both vector variable $\boldsymbol{x}$ and matrix variable $\boldsymbol{x}$. It further yields

$$\|\nabla_{\mathcal{I}} f_i(\boldsymbol{x}) - \nabla_{\mathcal{I}} f_i(\boldsymbol{x}^*)\|_2^2 \le 2\ell_s\left(f_i(\boldsymbol{x}) - f_i(\boldsymbol{x}^*) - \langle \nabla_{\mathcal{I}} f_i(\boldsymbol{x}^*), \boldsymbol{x} - \boldsymbol{x}^*\rangle\right).$$

Then we are ready to bound the second term:

$$\begin{aligned}
\mathbb{E}\|\nabla_{\mathcal{I}} f(\boldsymbol{x}^t) - \nabla_{\mathcal{I}} f(\boldsymbol{x}^*)\|^2 =& \frac{1}{n}\sum_{i=1}^n \|\nabla_{\mathcal{I}} f_i(\boldsymbol{x}^t) - \nabla_{\mathcal{I}} f_i(\boldsymbol{x}^*)\|^2 \\
\le& \frac{1}{n}\sum_{i=1}^n 2\ell_s\left(f_i(\boldsymbol{x}^t) - f_i(\boldsymbol{x}^*) - \langle \nabla_{\mathcal{I}} f_i(\boldsymbol{x}^*), \boldsymbol{x}^t - \boldsymbol{x}^*\rangle\right) \\
=& 2\ell_s\left(f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*) + \langle \nabla_{\mathcal{I}} f(\boldsymbol{x}^*), \boldsymbol{x}^t - \boldsymbol{x}^*\rangle\right).
\end{aligned}$$

Therefore, for both vector variable $\boldsymbol{x}$ and matrix variable $\boldsymbol{x}$ we have

$$\mathbb{E}\|\boldsymbol{g}_{\mathcal{I}}^t\|^2 \le \frac{3}{s_t}B_t + 6\ell_s\left(f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*) + \langle \nabla_{\mathcal{I}} f(\boldsymbol{x}^*), \boldsymbol{x}^t - \boldsymbol{x}^*\rangle\right) + 3\|\nabla_{\mathcal{I}} f(\boldsymbol{x}^*)\|^2.$$

This completes the proof. □

### D.2 Proof of Lemma 5

*Proof.* We first consider vector variable $\boldsymbol{x}$. From the strong convexity we get

$$\begin{aligned}
f(\boldsymbol{x}^*) \ge& f(\boldsymbol{x}^t) + \langle \nabla f(\boldsymbol{x}^t), \boldsymbol{x}^* - \boldsymbol{x}^t\rangle + \frac{\rho_s}{2}\|\boldsymbol{x}^* - \boldsymbol{x}^t\|^2 \\
\overset{①}{=}& f(\boldsymbol{x}^t) + \langle \nabla_{\mathcal{I}^t\cup\mathcal{I}^*} f(\boldsymbol{x}^t) - \boldsymbol{g}_{\mathcal{I}^t\cup\mathcal{I}^*}^t + \boldsymbol{g}_{\mathcal{I}^t\cup\mathcal{I}^*}^t, \boldsymbol{x}^* - \boldsymbol{x}^t\rangle + \frac{\rho_s}{2}\|\boldsymbol{x}^* - \boldsymbol{x}^t\|^2 \\
\overset{②}{\ge}& f(\boldsymbol{x}^t) - \frac{2}{\rho_s}\|\nabla f(\boldsymbol{x}^t) - \boldsymbol{g}^t\|^2 - \frac{2}{\rho_s}\|\boldsymbol{g}_{\mathcal{I}^t\cup\mathcal{I}^*}^t\|^2 - \frac{\rho_s}{4}\|\boldsymbol{x}^* - \boldsymbol{x}^t\|^2 + \frac{\rho_s}{2}\|\boldsymbol{x}^* - \boldsymbol{x}^t\|^2 \\
=& f(\boldsymbol{x}^t) - \frac{2}{\rho_s}\|f(\boldsymbol{x}^t) - \boldsymbol{g}^t\|^2 - \frac{2}{\rho_s}\|\boldsymbol{g}_{\mathcal{I}^t\cup\mathcal{I}^*}^t\|^2 + \frac{\rho_s}{4}\|\boldsymbol{x}^* - \boldsymbol{x}^t\|^2,
\end{aligned} \tag{25}$$

where ② holds since we use $\langle \boldsymbol{x}, \boldsymbol{y}\rangle \ge -\left(\frac{1}{2c}\|\boldsymbol{x}\|_2^2 + \frac{c}{2}\|\boldsymbol{y}\|_2^2\right)$ for arbitrary $c \ge 0$. By rearranging both sides of the above we get the desired bound.

Then we consider the matrix variable $\boldsymbol{x} \in \mathbb{R}^{d_1\times d_2}$ for rank-constrained problem. Firstly, assume that the skinny SVDs of $\boldsymbol{x}^t$ and $\boldsymbol{x}^*$ are respectively $\boldsymbol{x}^t = \boldsymbol{U}_t\boldsymbol{\Sigma}_t\boldsymbol{V}_t^T$ and $\boldsymbol{x}^* = \boldsymbol{U}_*\boldsymbol{\Sigma}_*\boldsymbol{V}_*^T$. Then we have $\boldsymbol{x}^t = \boldsymbol{U}_t\boldsymbol{U}_t^T\boldsymbol{x}^t$ and $\boldsymbol{x}^* = \boldsymbol{U}_*\boldsymbol{U}_*^T\boldsymbol{x}^*$. We further define the projection operation $\mathcal{P}_{\boldsymbol{U}}(\boldsymbol{x}) = \boldsymbol{U}\boldsymbol{U}^T\boldsymbol{x}$ which projects $\boldsymbol{x}$ in the subspace spanned by $\boldsymbol{U}$. Then let $\mathcal{I} = \mathcal{I}^* \cup \mathcal{I}^t$, where $\mathcal{I}^* = \mathsf{supp}(\boldsymbol{x}^*)$ and $\mathcal{I}^t = \mathsf{supp}(\boldsymbol{x}^t)$. The notation $\boldsymbol{x}_{\mathcal{I}^t}^t = \boldsymbol{U}_t\boldsymbol{U}_t^T\boldsymbol{x}^t = \mathcal{P}_{\boldsymbol{U}^t}(\boldsymbol{x}^t)$. For the specifical notation $\nabla_{\mathcal{I}} f(\boldsymbol{x}^t)$,

it denotes $\nabla_{\mathcal{I}} f(\boldsymbol{x}^t) = \mathcal{P}_{\mathcal{I}}\left(\nabla f(\boldsymbol{x}^t)\right) = (\boldsymbol{U}_t\boldsymbol{U}_t^T + \boldsymbol{U}_*\boldsymbol{U}_*^T - \boldsymbol{U}_t\boldsymbol{U}_t^T\boldsymbol{U}_*\boldsymbol{U}_*^T)\nabla f(\boldsymbol{x}^t)$ where $\mathcal{P}_{\mathcal{I}}$ is a projection. Then we have

$$\mathcal{P}_{\mathcal{I}}(\boldsymbol{x}^* - \boldsymbol{x}^t) = (\boldsymbol{U}_t\boldsymbol{U}_t^T + \boldsymbol{U}_*\boldsymbol{U}_*^T - \boldsymbol{U}_t\boldsymbol{U}_t^T\boldsymbol{U}_*\boldsymbol{U}_*^T)(\boldsymbol{U}_*\boldsymbol{U}_*^T\boldsymbol{x}^* - \boldsymbol{U}_t\boldsymbol{U}_t^T\boldsymbol{x}^t) = \boldsymbol{x}^* - \boldsymbol{x}^t,$$

which further gives

$$\begin{aligned}\langle \nabla f(\boldsymbol{x}^t), \boldsymbol{x}^* - \boldsymbol{x}^t\rangle =& \langle \nabla f(\boldsymbol{x}^t), \mathcal{P}_{\mathcal{I}}(\boldsymbol{x}^* - \boldsymbol{x}^t)\rangle \overset{①}{=} \langle \mathcal{P}_{\mathcal{I}}(\nabla f(\boldsymbol{x}^t)), \mathcal{P}_{\mathcal{I}}(\boldsymbol{x}^* - \boldsymbol{x}^t)\rangle \\ =& \langle \nabla_{\mathcal{I}} f(\boldsymbol{x}^t), \boldsymbol{x}^* - \boldsymbol{x}^t\rangle,\end{aligned}$$

where ① holds since $\langle \mathcal{P}_{\mathcal{I}}(\nabla f(\boldsymbol{x}^t)), \mathcal{P}_{\mathcal{I}}(\boldsymbol{x}^* - \boldsymbol{x}^t)\rangle = \langle \nabla f(\boldsymbol{x}^t), \mathcal{P}_{\mathcal{I}}^T\mathcal{P}_{\mathcal{I}}(\boldsymbol{x}^* - \boldsymbol{x}^t)\rangle = \langle \nabla f(\boldsymbol{x}^t), \mathcal{P}_{\mathcal{I}}(\boldsymbol{x}^* - \boldsymbol{x}^t)\rangle$ due to $\mathcal{P}_{\mathcal{I}}^T\mathcal{P}_{\mathcal{I}} = \mathcal{P}_{\mathcal{I}}$.

In this way, ① in Eqn. (25) holds. Thus, the above result also hold for matrix variable in rank-constrained problem. The proof is completed. □

### D.3 Proof of Lemma 6

*Proof.* Let $\lambda_1 \le \lambda_2 \le \cdots \le \lambda_d$ be the eigenvalues of $\boldsymbol{A}$ and $\boldsymbol{\Lambda}$ be a diagonal matrix whose diagonal entries are $\{\lambda_i\}$ in a non-decreasing order. By proper manipulation we get

$$\left\|\begin{bmatrix}(1+\nu)\boldsymbol{I} - \eta\boldsymbol{A} & -\nu\boldsymbol{I}\\ \boldsymbol{I} & 0\end{bmatrix}\right\| = \left\|\begin{bmatrix}(1+\nu)\boldsymbol{I} - \eta\boldsymbol{\Lambda} & -\nu\boldsymbol{I}\\ \boldsymbol{I} & 0\end{bmatrix}\right\| = \max_{i\in[d]}\left\|\begin{bmatrix}1+\nu-\eta\lambda_i & -\nu\\ 1 & 0\end{bmatrix}\right\|,$$

where in the second equality we have used the fact that it is possible to permute the matrix to a block diagonal matrix with $2\times 2$ blocks. For each $i \in [d]$, the eigenvalues of the $2\times 2$ matrices are given by the roots of

$$\lambda^2 - (1+\nu-\eta\lambda_i)\lambda + \nu = 0.$$

Given that $\nu \ge |1 - \sqrt{\eta\lambda_i}|^2$, the roots of the above equation are imaginary and both have magnitude $\sqrt{\nu}$. Since $\nu = \max\{|1-\sqrt{\eta\mu}|^2, |1-\sqrt{\eta\ell}|^2\}$, the magnitude of each root is at most $\max\{|1-\sqrt{\eta\mu}|, |1-\sqrt{\eta\ell}|\}$. This proves the desired spectral norm bound. □

## E Additional Experimental Results

### E.1 Descriptions of Testing Datasets

We briefly introduce the seven testing datasets in the manuscript. Among them, three datasets are provided in the LibSVM website[1], including rcv1, real-sim and epsilon. We also evaluate our algorithms on mnist[2] for handwritting recognition, news20[3] for news classification, coil100[4] and caltech256[5] for image classification. Their detailed information is summarized in Table 2. We can observe that these datasets are different from each other in feature dimension, training samples, and class numbers, *etc*. It should be mentioned that for caltech256 including 256 kinds of objects and one background class, we use its OverFeat feature, while for other datasets, we all use their raw data.

Table 2: Descriptions of the ten testing datasets.

| | #class | #sample | #feature | | #class | #sample | #feature |
|---|---|---|---|---|---|---|---|
| rcv1 | 2 | 20,242 | 47,236 | news20 | 20 | 62,061 | 15,935 |
| real-sim | 2 | 72,309 | 20,958 | coil100 | 100 | 7,200 | 1,024 |
| epsilon | 2 | 100,000 | 2,000 | caltech256 | 257 | 5,140 | 2,000 |
| mnist | 10 | 60,000 | 784 | | | | |

[1] https://www.csie.ntu.edu.tw/ cjlin/libsvmtools/datasets/
[2] http://yann.lecun.com/exdb/mnist/
[3] http://qwone.com/ jason/20Newsgroups/
[4] http://www1.cs.columbia.edu/CAVE/software/softlib/coil-100.php
[5] https://authors.library.caltech.edu/7694/

### E.2 More Experiments on A Single Pass over Data

Finally, we give more experimental results on a single pass over data. Following the setting in Section 5 in manuscript, here we test the considered algorithms on logistic regression with regularization parameters $\lambda = 10^{-5}$. We follow our theoretical results to exponentially expand the mini-batch size $s_k$ in HSG-HT and AHSG-HT and set $\tau = 1$. Figure 3 summarizes the numerical results in this setting. One can observe that on these optimization problems, most algorithms still achieve high accuracy after one pass over data, while HSG-HT and AHSG-HT also converge significantly faster than the other algorithms. These observations are consistent with the results in Figure 1 in the manuscript. All these results demonstrate the high efficiency of HSG-HT and AHSG-HT and also confirm the theoretical implication of Corollary 1 and 2 that HSG-HT and AHSG-HT always have lowest hard thresholding complexity than the compared algorithms and have lower in IFO complexity than other considered variance-reduced algorithms linearly depending on the sample size $n$, when the desired accuracy is moderately small and data scale is large.

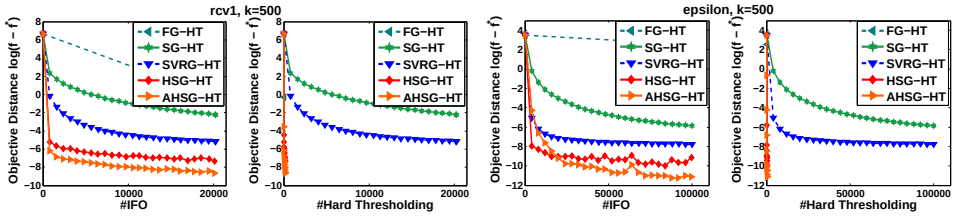
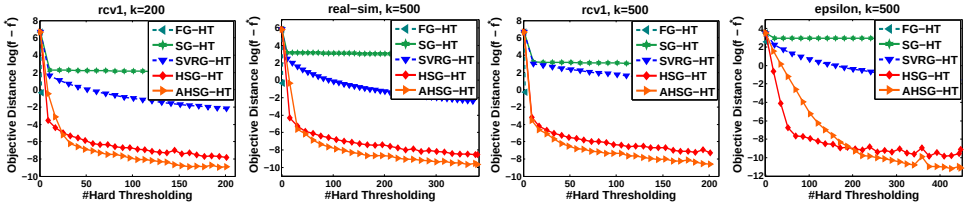



Figure 3: Single-epoch processing: comparison among hard thresholding algorithms for a single pass over data on sparse logistic regression with regularization parameter $\lambda = 10^{-5}$.

Since the objective loss decreases fast along with the hard thresholding iteration, here we magnify the subfigures in Figure 1 in the manuscript and the above Figure 3 which display the objective loss decrease along with the hard thresholding iteration. In this way, the objective loss decrease along with the hard thresholding iteration can be viewed better. From Figure 4, one can easily observe that AHSG-HT and HSG-HT converge much faster than the compared algorithms. Moreover, AHSG-HT achieves higher optimization accuracy which demonstrates that AHSG-HT is superior over HSG-HT in hard thresholding complexity. All these results confirm the our theoretical implication of Corollary 1 and 2.




(a) (b)

Figure 4: Comparison of hard thresholding complexity in single-epoch processing. (a) magnifies the hard thresholding iterations in Figure 1 in the manuscript. (b) magnifies the hard thresholding iterations in Figure 3 above.

## References


[1] X. Li, R. Arora, H. Liu, J. Haupt, and T. Zhao. Nonconvex sparse learning via stochastic optimization with progressive variance reduction. *Proc. Int'l Conf. Machine Learning*, 2016. 1

[2] P. Jain, A. Tewari, and P. Kar. On iterative hard thresholding methods for high-dimensional M-estimation. In *Proc. Conf. Neutral Information Processing Systems*, pages 685–693, 2014. 1, 6

[3] M. Friedlander and M. Schmidt. Hybrid deterministic-stochastic methods for data fitting. *SIAM Journal on Scientific Computing*, 34(3):A1380–A1405, 2012. 1