Stochastic Variance Reduced Optimization for Nonconvex Sparse Learning

Xingguo Li[*] LIXX1661@UMN.EDU
Tuo Zhao[†] TZHAO5@JHU.EDU
Raman Arora[†] ARORA@CS.JHU.EDU
Han Liu[‡] HANLIU@PRINCETON.EDU
Jarvis Haupt[*] JDHAUPT@UMN.EDU

[*]Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455
[†]Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218
[‡]Department of Operations Research and Financial Engineering, Princeton University, NJ 08544
Xingguo Li and Tuo Zhao equally contributed.

Abstract

We propose a stochastic variance reduced optimization algorithm for solving a class of large-scale nonconvex optimization problems with cardinality constraints, and provide sufficient conditions under which the proposed algorithm enjoys strong linear convergence guarantees and optimal estimation accuracy in high dimensions. Numerical experiments demonstrate the efficiency of our method in terms of both parameter estimation and computational performance.

1. Introduction

High dimensionality is challenging from both the statistical and computational perspectives. To make the analysis manageable, we usually assume that only a small number of variables are relevant for modeling the response variable. In the past decade, a large family of ℓ_1 regularized or ℓ_1 constrained sparse estimators have been proposed, including Lasso (Tibshirani, 1996), Logistic Lasso (Van de Geer, 2008), Group Lasso (Yuan & Lin, 2006), Graphical Lasso (Banerjee et al., 2008; Friedman et al., 2008), and more. The ℓ_1 regularization serves as a convex surrogate for controlling the cardinality of the parameters, and a large family of algorithms such as proximal gradient algorithms (Nesterov, 2013) have been developed for finding ℓ_1 regularized estimators in polynomial time. However, techniques based on convex relaxation, using ℓ_1 norm as a surrogate for ℓ_0 constraint, often incur large estimation bias, and attain worse empirical performance than those based on the cardinality constraint (Fan & Li, 2001; Zhang, 2010; Zhao et al., 2014a; Zhao & Liu, 2016). This motivates us to study a family of cardinality constrained M-estimators. Formally, we consider the nonconvex problem:

$$\min_{\boldsymbol{w}\in\mathbb{R}^d} \mathcal{F}(\boldsymbol{w}) \quad \text{s.t. } \|\boldsymbol{w}\|_0 \leq k, \tag{1.1}$$

where $\mathcal{F}(\boldsymbol{w})$ is a smooth and nonstrongly convex loss function, and $\|\boldsymbol{w}\|_0$ denotes the number of nonzero entries in $\boldsymbol{w}$.

To solve (1.1), a gradient hard thresholding (GHT) algorithm has been studied in the statistics as well as the machine learning community over the past few years (Blumensath & Davies, 2009; Foucart, 2011; Yuan et al., 2013; Jain et al., 2014). GHT involves performing a gradient update followed by a hard thresholding operation. Let $\mathcal{H}_k(\boldsymbol{w})$ denote a hard thresholding operator that keeps the largest k entries in magnitude and sets the other entries equal to zero. Then, given a solution $\boldsymbol{w}^{(t+1)}$ at the t-th iteration, GHT performs the following update:

$$\boldsymbol{w}^{(t+1)} = \mathcal{H}_k\Big(\boldsymbol{w}^{(t+1)} - \eta\nabla\mathcal{F}(\boldsymbol{w}^{(t+1)})\Big),$$

where $\mathcal{H}_k(\cdot)$ is the hard thresholding operator, which keeps the largest k (in magnitude) entries and sets the other entries equal to zero, $\nabla\mathcal{F}(\boldsymbol{w}^{(t+1)})$ is the gradient of the objective at $\boldsymbol{w}^{(t+1)}$ and η is a step size. Existing literature has shown that under suitable conditions, the GHT algorithm attains linear convergence to an approximately global optimum with optimal estimation accuracy with high probability (Yuan et al., 2013; Jain et al., 2014).

The GHT algorithm, though enjoying good convergence rates, is not suitable for solving large-scale problems. The computational bottleneck stems from the fact that the GHT algorithm evaluates the (full) gradient at each iteration; its computational complexity therefore depends linearly on the number of samples. The GHT algorithm, therefore, becomes computationally expensive for high-dimensional problems with large sample size.

To address the scalability issue, (Nguyen et al., 2014) considers a scenario that is a typical setting in machine learning wherein the objective function decomposes over samples, i.e. where the objective function $\mathcal{F}(\boldsymbol{w})$ takes an additive form over many smooth component functions:

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

$$\mathcal{F}(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{w}) \text{ and } \nabla\mathcal{F}(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\boldsymbol{w}),$$

and each $f_i(\boldsymbol{w})$ is associated with a few samples (i.e., the mini-batch setting). In such settings, we exploit the additive nature of $\mathcal{F}(\boldsymbol{w})$ and consider a stochastic gradient hard thresholding (SGHT) algorithm based on unbiased estimates of the full gradient rather than computing it. In particular, the SGHT algorithm estimates the full gradient $\nabla\mathcal{F}(\boldsymbol{w}^{(t+1)})$ by a stochastic gradient $\nabla f_i(\boldsymbol{w}^{(t+1)})$, where $f_i(\boldsymbol{w})$ is uniformly randomly sampled from all n component functions at each iteration. Though the SGHT algorithm greatly reduces the computational complexity in each iteration, it can only obtain an estimator with suboptimal estimation accuracy, owing to the variance of the stochastic gradient introduced by random sampling. Moreover, the theoretical analysis in (Nguyen et al., 2014) requires $\mathcal{F}(\boldsymbol{w})$ to satisfy the Restricted Isometry Property (RIP) with parameter $1/7$, i.e., the restricted condition number of the Hessian matrix $\nabla^2\mathcal{F}(\boldsymbol{w})$ cannot exceed $4/3$ (see more details in §3). Taking sparse linear regression as an example, such an RIP condition requires the design matrix to be nearly orthogonal, which is not satisfied by many simple random correlated Gaussian designs (Raskutti et al., 2010).

To address the suboptimal estimation accuracy and the restrictive requirement on $\mathcal{F}(\boldsymbol{w})$ in the stochastic setting, we propose a stochastic variance reduced gradient hard thresholding (SVR-GHT) algorithm. More specifically, we exploit a semi-stochastic optimization scheme to reduce the variance introduced by the random sampling (Johnson & Zhang, 2013; Konečnỳ & Richtárik, 2013; Zhao et al., 2014b). The SVR-GHT algorithm contains two nested loops: In each outer loop, SVR-GHT calculates the full gradient. In the subsequent inner loops, on each iteration the stochastic gradient update is adjusted by the full gradient followed by hard thresholding. This simple modification enables the algorithm to attain linear convergence to an approximately global optimum with optimal estimation accuracy, and meanwhile the amortized computational complexity remains similar to that of conventional stochastic optimization. Moreover, our theoretical analysis is applicable to an arbitrarily large restricted condition number of the Hessian matrix $\nabla^2\mathcal{F}(\boldsymbol{w})$.

Several existing algorithms are closely related to our proposed algorithm, including the proximal stochastic variance reduced gradient algorithm (Xiao & Zhang, 2014), stochastic averaging gradient algorithm (Roux et al., 2012) and stochastic dual coordinate ascent algorithm (Shalev-Shwartz & Zhang, 2013). However, these algorithms guarantee global linear convergence only for strongly convex optimization problems. Several existing statistical methods are also closely related to cardinality constrained M-estimators, including nonconvex constrained/regularized M-estimators (Shen et al., 2012; Loh & Wainwright, 2013). These methods usually require somewhat complicated computational formulation and often involve many tuning parameters (see more details in §6).

2. Algorithm

Before we proceed with the proposed algorithm, we introduce some notation. Given an integer $n \geq 1$, we define $[n] = \{1, \ldots, n\}$. Given a vector $\boldsymbol{v} = (v_1, \ldots, v_d)^\top \in \mathbb{R}^d$, we define vector norms: $\|\boldsymbol{v}\|_1 = \sum_j |v_j|$, $\|\boldsymbol{v}\|_2^2 = \sum_j v_j^2$, and $\|\boldsymbol{v}\|_\infty = \max_j |v_j|$. Given an index set $\mathcal{I} \subseteq [d]$, we define $\mathcal{I}^C$ as the complement set of $\mathcal{I}$, and $\boldsymbol{v}_\mathcal{I} \in \mathbb{R}^d$, where $[\boldsymbol{v}_\mathcal{I}]_j = v_j$ if $j \in \mathcal{I}$ and $[\boldsymbol{v}_\mathcal{I}]_j = 0$ if $j \notin \mathcal{I}$. We use $\mathrm{supp}(\boldsymbol{v})$ to denote the index set of nonzero entries of $\boldsymbol{v}$. Given two vectors $\boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^d$, we use $\langle \boldsymbol{v}, \boldsymbol{w}\rangle = \sum_{i=1}^d v_i w_i$ to denote the inner product. Given a matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$, we use $\mathbf{A}^\top$ to denote the transpose, and $\mathbf{A}_{i*}$ (or $\mathbf{A}_{*j}$) to denote the i-th row (or j-th column) of $\mathbf{A}$. Given an index set $\mathcal{I} \subseteq [d]$, we denote the submatrix of $\mathbf{A}$ with all row (or column) indices in $\mathcal{I}$ by $\mathbf{A}_{\mathcal{I}*}$ (or $\mathbf{A}_{*\mathcal{I}}$). Moreover, we use the common notations of $\Omega(\cdot)$ and $\mathcal{O}(\cdot)$ to characterize the asymptotics of two real sequences. We denote $\log(\cdot)$ as the natural logarithm when we do not specify the base.

We summarize the proposed stochastic variance reduced gradient hard thresholding (SVR-GHT) algorithm in Algorithm 1. Different from the SGHT algorithm proposed in (Nguyen et al., 2014), our SVR-GHT algorithm adopts the semi-stochastic optimization scheme proposed in (Johnson & Zhang, 2013), which can guarantee that the variance introduced by stochastic sampling over component functions diminishes with the optimization error.

Algorithm 1 Stochastic Variance Reduced Gradient Hard Thresholding Algorithm.

Input: update frequency m, step size parameter η, sparsity k, and initial solution $\widetilde{\boldsymbol{w}}^{(0)}$
for $r = 1, 2, \ldots$
  $\widetilde{\boldsymbol{w}} = \widetilde{\boldsymbol{w}}^{(r-1)}$, $\widetilde{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^n \nabla f_i(\widetilde{\boldsymbol{w}})$, $\boldsymbol{w}^{(0)} = \widetilde{\boldsymbol{w}}$
  for $t = 0, 1, \ldots, m-1$
    (S1) Randomly sample i_t from $[n]$
    (S2) $\bar{\boldsymbol{w}}^{(t+1)} = \boldsymbol{w}^{(t)} - \eta\left(\nabla f_{i_t}(\boldsymbol{w}^{(t)}) - \nabla f_{i_t}(\widetilde{\boldsymbol{w}}) + \widetilde{\boldsymbol{\mu}}\right)$
    (S3) $\boldsymbol{w}^{(t+1)} = \mathcal{H}_k(\bar{\boldsymbol{w}}^{(t+1)})$
  end for
  $\widetilde{\boldsymbol{w}}^{(r)} = \boldsymbol{w}^{(m)}$
end for

3. Analysis

Throughout the analysis, we assume that the objective function $\mathcal{F}(\boldsymbol{w})$ satisfies the restricted strong convexity (RSC) condition, and the component functions $\{f_i(\boldsymbol{w})\}_{i=1}^n$ satisfy the restricted strong smoothness (RSS) condition.

Definition 3.1 (Restricted Strong Convexity Condition)**.** A differentiable function $\mathcal{F}$ is restricted ρ_s^--strongly convex

at sparsity level s if there exists a constant $\rho_s^- > 0$ such that for any $\boldsymbol{w}, \boldsymbol{w}' \in \mathbb{R}^d$ with $\|\boldsymbol{w} - \boldsymbol{w}'\|_0 \leq s$, we have

$$\mathcal{F}(\boldsymbol{w}) - \mathcal{F}(\boldsymbol{w}') - \langle \nabla \mathcal{F}(\boldsymbol{w}'), \boldsymbol{w} - \boldsymbol{w}' \rangle \geq \tfrac{\rho_s^-}{2} \|\boldsymbol{w} - \boldsymbol{w}'\|_2^2. \quad (3.1)$$

Definition 3.2 (Restricted Strong Smoothness Condition)**.** For any $i \in [n]$, a differentiable function f_i is restricted ρ_s^+-strongly smooth at sparsity level s if there exists a uniform constant $\rho_s^+ > 0$ such that for any $\boldsymbol{w}, \boldsymbol{w}' \in \mathbb{R}^d$ with $\|\boldsymbol{w} - \boldsymbol{w}'\|_0 \leq s$, we have

$$f_i(\boldsymbol{w}) - f_i(\boldsymbol{w}') - \langle \nabla f_i(\boldsymbol{w}'), \boldsymbol{w} - \boldsymbol{w}' \rangle \leq \tfrac{\rho_s^+}{2} \|\boldsymbol{w} - \boldsymbol{w}'\|_2^2. \quad (3.2)$$

We also define the restricted condition number as $\kappa_s = \frac{\rho_s^+}{\rho_s^-}$.

(I) Computational Theory: We first present our main result characterizing the error of the objective value and estimation error of parameters.

Theorem 3.3. Let $\boldsymbol{w}^*$ be a sparse vector of the true model parameter such that $\|\boldsymbol{w}^*\|_0 \leq k^*$. Suppose $\mathcal{F}(\boldsymbol{w})$ satisfies RSC condition and $\{f_i(\boldsymbol{w})\}_{i=1}^n$ satisfy RSS condition with $s = 2k + k^*$, and Algorithm 1 is invoked with $k \geq C_1 \kappa_s^2 k^*$, $\eta \rho_s^+ \in [C_2, C_3]$ and $m \geq C_4 \kappa_s$ for some constants C_1, C_2, C_3 and C_4. Define $\widetilde{\mathcal{I}} = \mathrm{supp}\left(\mathcal{H}_{2k}(\nabla \mathcal{F}(\boldsymbol{w}^*))\right) \cup \mathrm{supp}(\boldsymbol{w}^*)$. Then, given some $\alpha = 1 + \frac{2\sqrt{k^*}}{\sqrt{k - k^*}}$ satisfying $\frac{\alpha^m(\alpha-1)}{\eta\rho_s^-(1-6\eta\rho_s^+)(\alpha^m - 1)} + \frac{6\eta\rho_s^+}{1-6\eta\rho_s^+} \leq \frac{3}{4}$, Algorithm 1 returns

$$\mathbb{E}\left[\mathcal{F}(\widetilde{\boldsymbol{w}}^{(r)}) - \mathcal{F}(\boldsymbol{w}^*)\right] \leq \left(\tfrac{3}{4}\right)^r \left[\mathcal{F}(\widetilde{\boldsymbol{w}}^{(0)}) - \mathcal{F}(\boldsymbol{w}^*)\right] + g_1(\boldsymbol{w}^*), \quad (3.3)$$

$$\mathbb{E}\|\widetilde{\boldsymbol{w}}^{(r)} - \boldsymbol{w}^*\|_2 \leq \sqrt{\frac{2\left(\frac{3}{4}\right)^r \left[\mathcal{F}(\widetilde{\boldsymbol{w}}^{(0)}) - \mathcal{F}(\boldsymbol{w}^*)\right]}{\rho_s^-}} + g_2(\boldsymbol{w}^*), \quad (3.4)$$

where $g_1(\boldsymbol{w}^*) = \frac{6\eta}{\left(1-6\eta\rho_s^+\right)} \|\nabla_{\widetilde{\mathcal{I}}} \mathcal{F}(\boldsymbol{w}^*)\|_2^2$ and $g_2(\boldsymbol{w}^*) = \frac{2\sqrt{s}\|\nabla \mathcal{F}(\boldsymbol{w}^*)\|_\infty}{\rho_s^-} + \|\nabla_{\widetilde{\mathcal{I}}} \mathcal{F}(\boldsymbol{w}^*)\|_2 \sqrt{\frac{12\eta}{(1-6\eta\rho_s^+)\rho_s^-}}$ are perturbations depending on $\boldsymbol{w}^*$. Moreover, given a constant $\delta \in (0,1)$ and a pre-specified accuracy $\varepsilon > 0$, we need at most $r = \left\lceil 4 \log\left(\frac{\mathcal{F}(\widetilde{\boldsymbol{w}}^{(0)}) - \mathcal{F}(\boldsymbol{w}^*)}{\varepsilon\delta}\right)\right\rceil$ outer iterations such that with probability at least $1 - \delta$, we have simultaneously

$$\mathcal{F}(\widetilde{\boldsymbol{w}}^{(r)}) - \mathcal{F}(\boldsymbol{w}^*) \leq \varepsilon + g_1(\boldsymbol{w}^*), \quad (3.5)$$

$$\|\widetilde{\boldsymbol{w}}^{(r)} - \boldsymbol{w}^*\|_2 \leq \sqrt{2\varepsilon/\rho_s^-} + g_2(\boldsymbol{w}^*). \quad (3.6)$$

Theorem 3.3 has three important implications: **(I)** Our analysis for SVR-GHT allows κ_s to increase with (n, d, k^*) as long as $\mathcal{F}(\boldsymbol{w})$ and all $f_i(\boldsymbol{w})$'s satisfy RSC and RSS at sparsity level $s = \Omega(\kappa_s^2 k^*)$. In contrast, the analysis for SGHT in (Nguyen et al., 2014) requires $\kappa \leq \frac{4}{3}$, which can be restrictive; **(II)** Existing literature shows that directly calculating $\min_{\|\boldsymbol{w}\|_0 \leq k^*} \mathcal{F}(\boldsymbol{w})$ is NP-hard (Natarajan, 1995). But with a suitably chosen $k = \Omega(\kappa_s^2 k^*)$, we can obtain a good approximation of $\boldsymbol{w}^*$ by SVR-GHT; **(III)** To get $\widetilde{\boldsymbol{w}}^{(r)}$ satisfying (3.5) and (3.6), we need $\mathcal{O}(\log(\frac{1}{\epsilon}))$ outer iterations. Since within each outer iteration, we need to calculate a full gradient and m stochastic variance reduced gradients, the overall computational complexity is $\mathcal{O}\left([n + \kappa_s] \cdot \log(\frac{1}{\epsilon})\right)$. In contrast, the overall computational complexity of GHT is $\mathcal{O}(\kappa_s n \log(\frac{1}{\epsilon}))$. Thus SVR-GHT gains a significant improvement over GHT when κ_s and n are large.

(II) Statistical Theory: We next present the statistical theory of constrained M-estimators obtained by the proposed SVR-GHT algorithm. Our analysis is applicable to a large family of statistical learning problems, including generalized linear models, low-rank matrix estimation (where the cardinality constraint would be replaced by a rank constraint) and sparse precision matrix estimation. We focus here on the most popular example: sparse linear regression, and leave the exploration of other models to future investigation.

We estimate the unknown sparse regression coefficient vector $\boldsymbol{w}^* \in \mathbb{R}^d$ from a noisy observation vector $\boldsymbol{y} \in \mathbb{R}^n$ of linear measurements: $\boldsymbol{y} = \mathbf{A}\boldsymbol{w}^* + \boldsymbol{z}$, where $\mathbf{A} \in \mathbb{R}^{nb \times d}$ is a design matrix, and $\boldsymbol{z} \in \mathbb{R}^{nb} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$. Here, we divide $\mathbf{A}$ into n submatrices, each of which contains b rows of $\mathbf{A}$. We denote the i-th submatrix as $\mathbf{A}_{\mathcal{S}_i *}$, where $\mathcal{S}_i$ is the corresponding row indices of $\mathbf{A}$ with $|\mathcal{S}_i| = b$ for all $i = 1, ..., n$. Then, the corresponding problem is

$$\min_{\boldsymbol{w} \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n \frac{1}{b} \|\boldsymbol{y}_{\mathcal{S}_i} - \mathbf{A}_{\mathcal{S}_i *} \boldsymbol{w}\|_2^2 \quad \text{s.t. } \|\boldsymbol{w}\|_0 \leq k. \quad (3.7)$$

We assume that $\|\boldsymbol{v}\|_0 \leq s$, $\boldsymbol{v} \in \mathbb{R}^d$, the design matrix $\mathbf{A}$ satisfies $\frac{\max_j \|\mathbf{A}_{*j}\|_2}{\sqrt{nb}} \leq 1$ and

$$\begin{aligned} \frac{\|\mathbf{A}\boldsymbol{v}\|_2^2}{nb} &\geq \psi_1 \|\boldsymbol{v}\|_2^2 - \varphi_1 \frac{\log d}{nb} \|\boldsymbol{v}\|_1^2, \\ \frac{\|\mathbf{A}_{\mathcal{S}_i *} \boldsymbol{v}\|_2^2}{b} &\leq \psi_2 \|\boldsymbol{v}\|_2^2 + \varphi_2 \frac{\log d}{b} \|\boldsymbol{v}\|_1^2, \ \forall i \in [n], \end{aligned} \quad (3.8)$$

where ψ_1, ψ_2, φ_1 and φ_2 are constants that do not scale with (n, b, k^*, d). Existing literature has shown that (3.8) is satisfied by many common examples of sub-Gaussian random design (Raskutti et al., 2010; Agarwal et al., 2012). The next lemma shows (3.8) implies RSC and RSS.

Lemma 3.4. Assume that the design matrix $\mathbf{A}$ satisfies (3.8). Given large enough n and b, there exist a constant C_5 and an integer k such that $\mathcal{F}(\boldsymbol{w})$ and $\{f_i(\boldsymbol{w})\}_{i=1}^n$ satisfy the RSC and RSS conditions with $s = 2k + k^*$, where

$$k = C_5 k^* \geq C_1 \kappa_s^2 k^*, \ \rho_s^- \geq \psi_1/2, \text{ and } \rho_s^+ \leq 2\psi_2.$$

Proof Sketch. For any $\boldsymbol{w}, \boldsymbol{w}' \in \mathbb{R}^d$ in sparse linear model, we have $\nabla^2 \mathcal{F}(\boldsymbol{w}) = \mathbf{A}^\top \mathbf{A}$ and

$$\mathcal{F}(\boldsymbol{w}) - \mathcal{F}(\boldsymbol{w}') - \langle \nabla \mathcal{F}(\boldsymbol{w}'), \boldsymbol{w} - \boldsymbol{w}' \rangle = \tfrac{1}{2} \|\mathbf{A}(\boldsymbol{w} - \boldsymbol{w}')\|_2^2.$$

By (3.8), if $b \geq \frac{\varphi_2 s \log d}{\psi_2}$ and $n \geq \frac{2\varphi_1\psi_2}{\psi_1\varphi_2}$, then we have $nb \geq \frac{2\varphi_1 s \log d}{\psi_1}$. Combining these with (3.8), we have $\rho_s^- \geq \frac{1}{2}\psi_1$, and $\rho_s^+ \leq 2\psi_2$. By the definition of κ, this indicates $\kappa_s = \frac{\rho_s^+}{\rho_s^-} \leq \frac{4\psi_2}{\psi_1}$. Then for some $C_5 \geq \frac{16 C_1 \psi_2^2}{\psi_1^2}$, we have $k = C_5 k^* \geq C_1 \kappa_s^2 k^*$. □

See detailed proof in Appendix A. Since Lemma 3.4 guarantees that $\mathcal{F}(\boldsymbol{w})$ and $\{f_i(\boldsymbol{w})\}_{i=1}^n$ satisfy the RSC and RSS conditions, Theorem 3.3 is applicable to the SVR-GHT algorithm for solving (3.7). This allows us to establish the following statistical guarantee for the obtained estimator.

Theorem 3.5. Suppose that the design matrix $\mathbf{A}$ satisfies (3.8), and k, η and m are specified as in Theorem 3.3. Then given a constant $\delta \in (0,1)$, a sufficiently small accuracy $\varepsilon > 0$, and large enough n and b, we need at most $r = \left\lceil 4\log\left(\frac{\mathcal{F}(\widetilde{\boldsymbol{w}}^{(0)})-\mathcal{F}(\boldsymbol{w}^*)}{\varepsilon\delta}\right)\right\rceil$ outer iterations such that with high probability, we have

$$\|\widetilde{\boldsymbol{w}}^{(r)} - \boldsymbol{w}^*\|_2 = \mathcal{O}\left(\sigma\sqrt{k^*\log d/(nb)}\right). \tag{3.9}$$

Proof Sketch. For sparse linear model, we have $\nabla\mathcal{F}(\boldsymbol{w}^*) = \mathbf{A}^\top\mathbf{z}/(nb)$. Since $\mathbf{z}$ has i.i.d. $\mathcal{N}(0,\sigma^2)$ entries, then $\mathbf{A}_{*j}^\top\mathbf{z}/(nb) \sim \mathcal{N}(0,\sigma^2\|\mathbf{A}_{*j}\|_2^2/(nb)^2)$ for any $j \in [d]$. Using the Mill's inequality for tail bounds of Normal distribution, we have

$$\mathbb{P}\left(\left|\frac{\mathbf{A}_{*j}^\top\mathbf{z}}{nb}\right| > 2\sigma\sqrt{\frac{\log d}{nb}}\right) = \mathbb{P}\left(\left|\frac{\mathbf{A}_{*j}^\top\mathbf{z}}{\sigma\|\mathbf{A}_{*j}\|_2}\right| > \frac{2\sqrt{nb\log d}}{\|\mathbf{A}_{*j}\|_2}\right)$$
$$\leq \frac{\|\mathbf{A}_{*j}\|_2}{\sqrt{2\pi nb\log d}}\cdot\exp\left(-\frac{4nb\log d}{\|\mathbf{A}_{*j}\|_2^2}\right) \overset{(i)}{\leq} \frac{d^{-4}}{\sqrt{2\pi\log d}},$$

where (i) is from the assumption $\frac{\max_j\|\mathbf{A}_{*j}\|_2}{\sqrt{nb}} \leq 1$. Then with probability at least $1 - \frac{1}{\sqrt{2\pi\log d}}\cdot d^{-4}$

$$\|\nabla_{\widetilde{\mathcal{I}}}\mathcal{F}(\boldsymbol{w}^*)\|_2^2 \leq s\|\nabla\mathcal{F}(\boldsymbol{w}^*)\|_\infty^2 = s\left\|\frac{\mathbf{A}^\top\mathbf{z}}{nb}\right\|_\infty^2 \leq \frac{4\sigma^2 s\log d}{nb}.$$

When r is as specified, $s = 2k + k^* = (2C_5+1)k^*$ from Lemma 3.4 and (3.4) holds from Theorem 3.3, we have (3.9) with probability at least $1 - \delta - \frac{1}{\sqrt{2\pi\log d}}\cdot d^{-4}$. □

See detailed proof in Appendix B. Theorem 3.5 guarantees that the obtained estimator attains the optimal rate of convergence in parameter estimation (Raskutti et al., 2011), regardless whether n is allowed to scale with (b, k^*, d) or not. In contrast (Nguyen et al., 2014) only considers a fixed n setting, and shows that the estimator obtained by the SGHT algorithm only attains $\mathcal{O}(\sigma\sqrt{k^*\log d/b})$ with high probability (See Corollary 5 in (Nguyen et al., 2014)). Hence their result is suboptimal w.r.t. n, while ours allow n to scale with (b, k^*, d).

4. Main Proof

We first present two key technical lemmas that will be instrumental in developing the computational theory for our proposed algorithm and throughout the rest of the paper.

Lemma 4.1. Let $\boldsymbol{w}^* \in \mathbb{R}^d$ be a sparse vector such that $\|\boldsymbol{w}^*\|_0 \leq k^*$, and $\mathcal{H}_k(\cdot) : \mathbb{R}^d \to \mathbb{R}^d$ be the hard thresholding operator, which keeps the largest k entries (in magnitude) and sets the other entries equal to zero. Given $k > k^*$, for any vector $\boldsymbol{w} \in \mathbb{R}^d$, we have

$$\|\mathcal{H}_k(\boldsymbol{w}) - \boldsymbol{w}^*\|_2^2 \leq \left(1 + \frac{2\sqrt{k^*}}{\sqrt{k-k^*}}\right)\|\boldsymbol{w} - \boldsymbol{w}^*\|_2^2. \tag{4.1}$$

Proof Sketch. For notational convenience, define $\boldsymbol{w}' = \mathcal{H}_k(\boldsymbol{w})$. Let $\mathrm{supp}(\boldsymbol{w}^*) = \mathcal{I}^*$, $\mathrm{supp}(\boldsymbol{w}) = \mathcal{I}$, $\mathrm{supp}(\boldsymbol{w}') = \mathcal{I}'$, and $\boldsymbol{w}'' = \boldsymbol{w} - \boldsymbol{w}'$ with $\mathrm{supp}(\boldsymbol{w}'') = \mathcal{I}''$. Clearly we have $\mathcal{I}' \cup \mathcal{I}'' = \mathcal{I}$, $\mathcal{I}' \cap \mathcal{I}'' = \emptyset$, and $\|\boldsymbol{w}\|_2^2 = \|\boldsymbol{w}'\|_2^2 + \|\boldsymbol{w}''\|_2^2$. Then we have that

$$\|\boldsymbol{w}' - \boldsymbol{w}^*\|_2^2 - \|\boldsymbol{w} - \boldsymbol{w}^*\|_2^2 = 2\langle\boldsymbol{w}'', \boldsymbol{w}^*\rangle - \|\boldsymbol{w}''\|_2^2.$$

If $2\langle\boldsymbol{w}'', \boldsymbol{w}^*\rangle - \|\boldsymbol{w}''\|_2^2 \leq 0$, then (4.1) holds naturally. We will discuss when $2\langle\boldsymbol{w}'', \boldsymbol{w}^*\rangle - \|\boldsymbol{w}''\|_2^2 > 0$.

Let $\mathcal{I}^* \cap \mathcal{I}' = \mathcal{I}^{*1}$ and $\mathcal{I}^* \cap \mathcal{I}'' = \mathcal{I}^{*2}$, and denote $(\boldsymbol{w}^*)_{\mathcal{I}^{*1}} = \boldsymbol{w}^{*1}$, $(\boldsymbol{w}^*)_{\mathcal{I}^{*2}} = \boldsymbol{w}^{*2}$, $(\boldsymbol{w}')_{\mathcal{I}^{*1}} = \boldsymbol{w}^{1*}$, and $(\boldsymbol{w}'')_{\mathcal{I}^{*2}} = \boldsymbol{w}^{2*}$. Then we have that

$$2\langle\boldsymbol{w}'', \boldsymbol{w}^*\rangle - \|\boldsymbol{w}''\|_2^2 \leq 2\langle\boldsymbol{w}^{2*}, \boldsymbol{w}^{*2}\rangle - \|\boldsymbol{w}^{2*}\|_2^2$$
$$\leq 2\|\boldsymbol{w}^{2*}\|_2\|\boldsymbol{w}^{*2}\|_2 - \|\boldsymbol{w}^{2*}\|_2^2. \tag{4.2}$$

Let $|\mathrm{supp}(\boldsymbol{w}^{2*})| = |\mathcal{I}^{*2}| = k^{**}$ and $w_{2,\max} = \|\boldsymbol{w}^{2*}\|_\infty$, then consequently we have $\|\boldsymbol{w}^{2*}\|_2 = m\cdot w_{2,\max}$ for some $m \in [1, \sqrt{k^{**}}]$. Notice that we are interested in $1 \leq k^{**} \leq k^*$, because (4.1) holds naturally if $k^{**} = 0$. We consider three cases to maximize the RHS of (4.2):

Case 1: If $\|\boldsymbol{w}^{*2}\|_2 \leq w_{2,\max}$, then the RHS of (4.2) is maximized when $m = 1$, i.e. $\boldsymbol{w}^{2*}$ has only one nonzero element $w_{2,\max}$. By calculation, we have

$$\frac{\|\boldsymbol{w}' - \boldsymbol{w}^*\|_2^2 - \|\boldsymbol{w} - \boldsymbol{w}^*\|_2^2}{\|\boldsymbol{w} - \boldsymbol{w}^*\|_2^2} \leq \frac{1}{k-k^*}. \tag{4.3}$$

Case 2: If $w_{2,\max} < \|\boldsymbol{w}^{*2}\|_2 < \sqrt{k^{**}}w_{2,\max}$, then the RHS of (4.2) is maximized when $m = \frac{\|\boldsymbol{w}^{*2}\|_2}{w_{2,\max}}$. By calculation, we have

$$\frac{\|\boldsymbol{w}' - \boldsymbol{w}^*\|_2^2 - \|\boldsymbol{w} - \boldsymbol{w}^*\|_2^2}{\|\boldsymbol{w} - \boldsymbol{w}^*\|_2^2} \leq \frac{k^{**}}{k-k^*+k^{**}}. \tag{4.4}$$

Case 3: If $\|\boldsymbol{w}^{*2}\|_2 \geq \sqrt{k^{**}}w_{2,\max}$, then the RHS of (4.2) is maximized when $m = \sqrt{k^{**}}$. By calculation, we have

$$\frac{\|\boldsymbol{w}' - \boldsymbol{w}^*\|_2^2 - \|\boldsymbol{w} - \boldsymbol{w}^*\|_2^2}{\|\boldsymbol{w} - \boldsymbol{w}^*\|_2^2}$$
$$\leq \max\left\{\frac{k^{**}}{k-k^*+k^{**}}, \frac{2\sqrt{k^{**}}}{2\sqrt{k-k^*+\frac{5}{4}k^{**}}-\sqrt{k^{**}}}\right\}. \tag{4.5}$$

We have desired result from (4.3), (4.4) and (4.5). □

See detailed proof in Appendix C. Lemma 4.1 shows that the hard thresholding operator is nearly non-expansive when k is much larger than k^* such that $\frac{2\sqrt{k^*}}{\sqrt{k-k^*}}$ is very small (bounded away from 1).

Remark 4.2. Though Lemma 4.1 looks similar to Lemma 1 in (Jain et al., 2014), they are essentially different. We provide a ratio of the distances between $\boldsymbol{w}$ and a k^*-sparse vector $\boldsymbol{w}^*$, before and after hard thresholding operation on $\boldsymbol{w}$, which is more intuitive than what is presented in (Jain et al., 2014) that gives a ratio of the distance between $\boldsymbol{w}$ and $\mathcal{H}_k(\boldsymbol{w})$, and the distance between $\boldsymbol{w}$ and $\boldsymbol{w}^*$. Besides, Lemma 4.1 is a also a key property that allow us to tolerate a large condition number κ_s, compared with a small bounded κ_s in (Nguyen et al., 2014), as long as the hard thresholding parameter k is large enough.

For notational simplicity, we denote the full gradient and the stochastic variance reduced gradient by

$$\widetilde{\boldsymbol{\mu}} = \nabla\mathcal{F}(\widetilde{\boldsymbol{w}}) = \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\widetilde{\boldsymbol{w}}) \quad \text{and}$$
$$\mathbf{g}^{(t)}(\boldsymbol{w}^{(t)}) = \nabla f_{i_t}(\boldsymbol{w}^{(t)}) - \nabla f_{i_t}(\widetilde{\boldsymbol{w}}) + \widetilde{\boldsymbol{\mu}}.$$

The next lemma shows that $\mathbf{g}^{(t)}(\boldsymbol{w}^{t-1})$ is an unbiased estimator of $\nabla\mathcal{F}(\boldsymbol{w}^{t-1})$ with a well controlled second order moment over a sparse support.

Lemma 4.3. Suppose that $\mathcal{F}(\boldsymbol{w})$ and all $f_i(\boldsymbol{w})$'s satisfy the RSC and RSS conditions with $s = 2k + k^*$. Let $\boldsymbol{w}^* \in \mathbb{R}^d$ be a sparse vector with $\|\boldsymbol{w}^*\|_0 \leq k^*$, $\mathcal{I}^* = \mathrm{supp}(\boldsymbol{w}^*)$, and $\boldsymbol{w}^{(t)}$ be a sparse vector with $\|\boldsymbol{w}^{(t)}\|_0 \leq k$, $\mathcal{I}^{(t)} = \mathrm{supp}(\boldsymbol{w}^{(t)})$. Then conditioning on $\boldsymbol{w}^{(t)}$, for any $\mathcal{I} \supseteq (\mathcal{I}^* \cup \mathcal{I}^{(t)})$, we have $\mathbb{E}[\mathbf{g}^{(t)}(\boldsymbol{w}^{(t)})] = \nabla\mathcal{F}(\boldsymbol{w}^{(t)})$ and

$$\mathbb{E}\|\mathbf{g}_{\mathcal{I}}^{(t)}(\boldsymbol{w}^{(t)})\|_2^2 \leq 12\rho_s^+\left[\mathcal{F}(\boldsymbol{w}^{(t)}) - \mathcal{F}(\boldsymbol{w}^*) + \mathcal{F}(\widetilde{\boldsymbol{w}}) - \mathcal{F}(\boldsymbol{w}^*)\right] + 3\|\nabla_{\mathcal{I}}\mathcal{F}(\boldsymbol{w}^*)\|_2^2. \tag{4.6}$$

Proof Sketch. It is straightforward that $\mathbf{g}^{(t)}(\boldsymbol{w}^{(t)})$ satisfies

$$\mathbb{E}\mathbf{g}^{(t)}(\boldsymbol{w}^{(t)}) = \mathbb{E}\nabla f_{i_t}(\boldsymbol{w}^{(t)}) - \mathbb{E}\nabla f_{i_t}(\widetilde{\boldsymbol{w}}) + \widetilde{\boldsymbol{\mu}} = \nabla\mathcal{F}(\boldsymbol{w}^{(t)}).$$

For any $i \in [n]$ and $\boldsymbol{w}$ with $\mathrm{supp}(\boldsymbol{w}) \subseteq \mathcal{I}$, consider

$$\phi_i(\boldsymbol{w}) = f_i(\boldsymbol{w}) - f_i(\boldsymbol{w}^*) - \langle\nabla f_i(\boldsymbol{w}^*), \boldsymbol{w} - \boldsymbol{w}^*\rangle.$$

Since $\nabla\phi_i(\boldsymbol{w}^*) = \nabla f_i(\boldsymbol{w}^*) - \nabla f_i(\boldsymbol{w}^*) = \mathbf{0}$, we have $\phi_i(\boldsymbol{w}^*) = \min_{\boldsymbol{w}}\phi_i(\boldsymbol{w})$, which implies

$$0 \leq \phi_i(\boldsymbol{w}) - \frac{1}{2\rho_s^+}\|\nabla\phi_i(\boldsymbol{w})\|_2^2,$$

This further results in

$$\|\nabla_{\mathcal{I}} f_i(\boldsymbol{w}) - \nabla_{\mathcal{I}} f_i(\boldsymbol{w}^*)\|_2^2 \leq \|\nabla f_i(\boldsymbol{w}) - \nabla f_i(\boldsymbol{w}^*)\|_2^2$$
$$\leq 2\rho_s^+\left[f_i(\boldsymbol{w}) - f_i(\boldsymbol{w}^*) - \langle\nabla_{\mathcal{I}} f_i(\boldsymbol{w}^*), \boldsymbol{w} - \boldsymbol{w}^*\rangle\right],$$

Since the sampling of i from $[n]$ is uniform sampling, this implies

$$\mathbb{E}\|\nabla_{\mathcal{I}} f_i(\boldsymbol{w}) - \nabla_{\mathcal{I}} f_i(\boldsymbol{w}^*)\|_2^2 = \tfrac{1}{n}\textstyle\sum_{i=1}^n\|\nabla_{\mathcal{I}} f_i(\boldsymbol{w}) - \nabla_{\mathcal{I}} f_i(\boldsymbol{w}^*)\|_2^2$$
$$\leq 2\rho_s^+\left[\mathcal{F}(\boldsymbol{w}) - \mathcal{F}(\boldsymbol{w}^*) - \langle\nabla_{\mathcal{I}}\mathcal{F}(\boldsymbol{w}^*), \boldsymbol{w} - \boldsymbol{w}^*\rangle\right]$$
$$\leq 4\rho_s^+\left[\mathcal{F}(\boldsymbol{w}) - \mathcal{F}(\boldsymbol{w}^*)\right], \tag{4.7}$$

By the definition of $\mathbf{g}_{\mathcal{I}}^{(t)}$, we can verify the second claim as

$$\mathbb{E}\|\mathbf{g}_{\mathcal{I}}^{(t)}(\boldsymbol{w}^{(t)})\|_2^2$$
$$\leq 3\mathbb{E}\|\left[\nabla_{\mathcal{I}} f_{i_t}(\widetilde{\boldsymbol{w}}) - \nabla_{\mathcal{I}} f_{i_t}(\boldsymbol{w}^*)\right] - \nabla_{\mathcal{I}}\mathcal{F}(\widetilde{\boldsymbol{w}}) + \nabla_{\mathcal{I}}\mathcal{F}(\boldsymbol{w}^*)\|_2^2$$
$$+3\mathbb{E}\|\nabla_{\mathcal{I}} f_{i_t}(\boldsymbol{w}^{(t)}) - \nabla_{\mathcal{I}} f_{i_t}(\boldsymbol{w}^*)\|_2^2 + 3\|\nabla_{\mathcal{I}}\mathcal{F}(\boldsymbol{w}^*)\|_2^2$$
$$\overset{(i)}{\leq} 3\mathbb{E}\|\nabla_{\mathcal{I}} f_{i_t}(\boldsymbol{w}^{(t)}) - \nabla_{\mathcal{I}} f_{i_t}(\boldsymbol{w}^*)\|_2^2$$
$$+3\mathbb{E}\|\nabla_{\mathcal{I}} f_{i_t}(\widetilde{\boldsymbol{w}}) - \nabla_{\mathcal{I}} f_{i_t}(\boldsymbol{w}^*)\|_2^2 + 3\|\nabla_{\mathcal{I}}\mathcal{F}(\boldsymbol{w}^*)\|_2^2.$$

Then (i) results in (4.6) from (4.7). □

See detailed proof in Appendix D. When $\boldsymbol{w}^*$ is a global minimizer, for convex problems, we have $\nabla\mathcal{F}(\boldsymbol{w}^*) = \mathbf{0}$ (or the differential of $\mathcal{F}(\boldsymbol{w}^*)$ contains $\mathbf{0}$ in composite minimization settings). However, we are working on a nonconvex optimization problem without such a convenience. This eventually results in this additional $\|\nabla_{\mathcal{I}}\mathcal{F}(\boldsymbol{w}^*)\|_2^2$ on the R.H.S of (4.6), which is different from existing variance reduction results in (Johnson & Zhang, 2013; Konečnỳ & Richtárik, 2013; Zhao et al., 2014b).

Remark 4.4. $\boldsymbol{w}^*$ can be arbitrary k^* sparse vector (1.1) in our analysis. While in establishing the statistical properties of the obtained estimator $\widetilde{\boldsymbol{w}}^{(r)}$, if $\boldsymbol{w}^*$ is the true model parameter, we have the estimate of expected $\mathcal{F}(\widetilde{\boldsymbol{w}}^{(r)})$ within the $\varepsilon + c\|\nabla_{\widetilde{\mathcal{I}}}\mathcal{F}(\boldsymbol{w}^*)\|_2^2$ distance to the expected $\mathcal{F}(\boldsymbol{w}^*)$, which results in that our estimator $\widetilde{\boldsymbol{w}}^{(r)}$ is within the optimal statistical accuracy to the true model parameter $\boldsymbol{w}^*$.

Now we are ready to provide the proof of Theorem 3.3. We present the proof as two parts.

Part 1: We first demonstrate (3.3) and (3.4). let $\boldsymbol{v} = \boldsymbol{w}^{(t)} - \eta\mathbf{g}_{\mathcal{I}}^{(t)}(\boldsymbol{w}^{(t)})$ and $\mathcal{I} = \mathcal{I}^* \cup \mathcal{I}^{(t)} \cup \mathcal{I}^{(t+1)}$, where $\mathcal{I}^* = \mathrm{supp}(\boldsymbol{w}^*)$, $\mathcal{I}^{(t)} = \mathrm{supp}(\boldsymbol{w}^{(t)})$ and $\mathcal{I}^{(t+1)} = \mathrm{supp}(\boldsymbol{w}^{(t+1)})$. Conditioning on $\boldsymbol{w}^{(t)}$, we have the following expectation

$$\mathbb{E}\|\boldsymbol{v} - \boldsymbol{w}^*\|_2^2 = \mathbb{E}\|\boldsymbol{w}^{(t)} - \eta\mathbf{g}_{\mathcal{I}}^{(t)}(\boldsymbol{w}^{(t)}) - \boldsymbol{w}^*\|_2^2$$
$$= \mathbb{E}\|\boldsymbol{w}^{(t)} - \boldsymbol{w}^*\|_2^2 + \eta^2\mathbb{E}\|\mathbf{g}_{\mathcal{I}}^{(t)}(\boldsymbol{w}^{(t)})\|_2^2 - 2\eta\langle\boldsymbol{w}^{(t)} - \boldsymbol{w}^*, \mathbb{E}\mathbf{g}_{\mathcal{I}}^{(t)}(\boldsymbol{w}^{(t)})\rangle$$
$$= \mathbb{E}\|\boldsymbol{w}^{(t)} - \boldsymbol{w}^*\|_2^2 + \eta^2\mathbb{E}\|\mathbf{g}_{\mathcal{I}}^{(t)}(\boldsymbol{w}^{(t)})\|_2^2 - 2\eta\langle\boldsymbol{w}^{(t)} - \boldsymbol{w}^*, \nabla_{\mathcal{I}}\mathcal{F}(\boldsymbol{w}^{(t)})\rangle$$
$$\leq \mathbb{E}\|\boldsymbol{w}^{(t)} - \boldsymbol{w}^*\|_2^2 + \eta^2\mathbb{E}\|\mathbf{g}_{\mathcal{I}}^{(t)}(\boldsymbol{w}^{(t)})\|_2^2 - 2\eta\left[\mathcal{F}(\boldsymbol{w}^{(t)}) - \mathcal{F}(\boldsymbol{w}^*)\right]$$
$$\leq \mathbb{E}\|\boldsymbol{w}^{(t)} - \boldsymbol{w}^*\|_2^2 + 12\eta^2\rho_s^+\left[\mathcal{F}(\boldsymbol{w}^{(t)}) - \mathcal{F}(\boldsymbol{w}^*) + \mathcal{F}(\widetilde{\boldsymbol{w}}) - \mathcal{F}(\boldsymbol{w}^*)\right]$$
$$-2\eta\left[\mathcal{F}(\boldsymbol{w}^{(t)}) - \mathcal{F}(\boldsymbol{w}^*)\right] + 3\eta^2\|\nabla_{\mathcal{I}}\mathcal{F}(\boldsymbol{w}^*)\|_2^2$$
$$= \mathbb{E}\|\boldsymbol{w}^{(t)} - \boldsymbol{w}^*\|_2^2 - 2\eta(1 - 6\eta\rho_s^+)\left[\mathcal{F}(\boldsymbol{w}^{(t)}) - \mathcal{F}(\boldsymbol{w}^*)\right]$$
$$+12\eta^2\rho_s^+\left[\mathcal{F}(\widetilde{\boldsymbol{w}}) - \mathcal{F}(\boldsymbol{w}^*)\right] + 3\eta^2\|\nabla_{\mathcal{I}}\mathcal{F}(\boldsymbol{w}^*)\|_2^2, \tag{4.8}$$

where the first inequality follows from the RSC condition (3.2) and the second inequality follows from Lemma 4.3. Since $\boldsymbol{w}^{(t+1)} = \bar{\boldsymbol{w}}_k^{(t+1)} = \boldsymbol{v}_k$, i.e. $\boldsymbol{w}^{(t+1)}$ is the best k-sparse approximation of $\boldsymbol{v}$, then we have from Lemma 4.1,

$$\|\boldsymbol{w}^{(t+1)} - \boldsymbol{w}^*\|_2^2 \leq \left(1 + \frac{2\sqrt{k^*}}{\sqrt{k - k^*}}\right)\|\boldsymbol{v} - \boldsymbol{w}^*\|_2^2. \tag{4.9}$$

Let $\alpha = 1 + \frac{2\sqrt{k^*}}{\sqrt{k-k^*}}$. Combining (4.8) and (4.9), we have

$$\mathbb{E}\|\boldsymbol{w}^{(t+1)} - \boldsymbol{w}^*\|_2^2 \leq \alpha\mathbb{E}\|\boldsymbol{w}^{(t)} - \boldsymbol{w}^*\|_2^2 + 12\alpha\eta^2\rho_s^+ \left[\mathcal{F}(\widetilde{\boldsymbol{w}}) - \mathcal{F}(\boldsymbol{w}^*)\right] + 3\alpha\eta^2\|\nabla_{\mathcal{I}}\mathcal{F}(\boldsymbol{w}^*)\|_2^2 - 2\alpha\eta(1-6\eta\rho_s^+)\left[\mathcal{F}(\boldsymbol{w}^{(t)}) - \mathcal{F}(\boldsymbol{w}^*)\right]. \tag{4.10}$$

Notice that $\widetilde{\boldsymbol{w}} = \boldsymbol{w}^{(0)} = \widetilde{\boldsymbol{w}}^{(r-1)}$. By summing (4.10) over $t = 0, 1, \ldots, m-1$ and taking expectation with respect to all t's, we have

$$\begin{aligned}
&\mathbb{E}\|\boldsymbol{w}^{(m)} - \boldsymbol{w}^*\|_2^2 + \tfrac{2\eta(1-6\eta\rho_s^+)(\alpha^m-1)}{\alpha-1}\mathbb{E}\left[\mathcal{F}(\widetilde{\boldsymbol{w}}^{(r)}) - \mathcal{F}(\boldsymbol{w}^*)\right] \\
&\leq \alpha^m\mathbb{E}\|\widetilde{\boldsymbol{w}}^{(r-1)} - \boldsymbol{w}^*\|_2^2 + \tfrac{3\eta^2(\alpha^m-1)}{\alpha-1}\mathbb{E}\|\nabla_{\mathcal{I}}\mathcal{F}(\boldsymbol{w}^*)\|_2^2 \\
&\quad + \tfrac{12\eta^2\rho_s^+(\alpha^m-1)}{\alpha-1}\mathbb{E}\left[\mathcal{F}(\widetilde{\boldsymbol{w}}^{(r-1)}) - \mathcal{F}(\boldsymbol{w}^*)\right] \\
&\leq \tfrac{2\alpha^m}{\rho_s^-}\mathbb{E}\left[\mathcal{F}(\widetilde{\boldsymbol{w}}^{(r-1)}) - \mathcal{F}(\boldsymbol{w}^*)\right] + \tfrac{3\eta^2(\alpha^m-1)}{\alpha-1}\|\nabla_{\widetilde{\mathcal{I}}}\mathcal{F}(\boldsymbol{w}^*)\|_2^2 \\
&\quad + \tfrac{12\eta^2\rho_s^+(\alpha^m-1)}{\alpha-1}\mathbb{E}\left[\mathcal{F}(\widetilde{\boldsymbol{w}}^{(r-1)}) - \mathcal{F}(\boldsymbol{w}^*)\right],
\end{aligned} \tag{4.11}$$

where the last inequality follows from the RSC condition (3.1) and the definition of $\widetilde{\mathcal{I}}$. It further follows from (4.11),

$$\mathbb{E}\left[\mathcal{F}(\widetilde{\boldsymbol{w}}^{(r)}) - \mathcal{F}(\boldsymbol{w}^*)\right] \leq \beta\mathbb{E}\left[\mathcal{F}(\widetilde{\boldsymbol{w}}^{(r-1)}) - \mathcal{F}(\boldsymbol{w}^*)\right] + \tfrac{3\eta}{2(1-6\eta\rho_s^+)}\|\nabla_{\widetilde{\mathcal{I}}}\mathcal{F}(\boldsymbol{w}^*)\|_2^2. \tag{4.12}$$

where $\beta = \frac{\alpha^m(\alpha-1)}{\eta\rho_s^-(1-6\eta\rho_s^+)(\alpha^m-1)} + \frac{6\eta\rho_s^+}{1-6\eta\rho_s^+}$. Apply (4.12) recursively, we obtain (3.3) when $\beta \leq \frac{3}{4} < 1$.

We then demonstrate (3.4). It follows from RSC condition

$$\mathcal{F}(\boldsymbol{w}^*) \leq \mathcal{F}(\widetilde{\boldsymbol{w}}^{(r)}) + \langle\nabla\mathcal{F}(\boldsymbol{w}^*), \boldsymbol{w}^* - \widetilde{\boldsymbol{w}}^{(r)}\rangle - \tfrac{\rho_s^-}{2}\|\widetilde{\boldsymbol{w}}^{(r)} - \boldsymbol{w}^*\|_2^2. \tag{4.13}$$

Let $\zeta = \left(\frac{3}{4}\right)^r\left[\mathcal{F}(\widetilde{\boldsymbol{w}}^{(0)}) - \mathcal{F}(\boldsymbol{w}^*)\right] + \frac{6\eta}{\left(1-6\eta\rho_s^+\right)}\|\nabla_{\widetilde{\mathcal{I}}}\mathcal{F}(\boldsymbol{w}^*)\|_2^2$. Combining (3.3) and (4.13), we have

$$\begin{aligned}
&\mathbb{E}\left[\mathcal{F}(\widetilde{\boldsymbol{w}}^{(r)}) - \zeta\right] \leq \mathcal{F}(\boldsymbol{w}^*) \\
&\leq \mathbb{E}\left[\mathcal{F}(\widetilde{\boldsymbol{w}}^{(r)}) + \langle\nabla\mathcal{F}(\boldsymbol{w}^*), \boldsymbol{w}^* - \widetilde{\boldsymbol{w}}^{(r)}\rangle - \tfrac{\rho_s^-}{2}\|\widetilde{\boldsymbol{w}}^{(r)} - \boldsymbol{w}^*\|_2^2\right].
\end{aligned} \tag{4.14}$$

Using the duality of norms, we have

$$\begin{aligned}
&\mathbb{E}\langle\nabla\mathcal{F}(\boldsymbol{w}^*), \boldsymbol{w}^* - \widetilde{\boldsymbol{w}}^{(r)}\rangle \leq \|\nabla\mathcal{F}(\boldsymbol{w}^*)\|_\infty\mathbb{E}\|\widetilde{\boldsymbol{w}}^{(r)} - \boldsymbol{w}^*\|_1 \\
&\leq \sqrt{s}\|\nabla\mathcal{F}(\boldsymbol{w}^*)\|_\infty\mathbb{E}\|\widetilde{\boldsymbol{w}}^{(r)} - \boldsymbol{w}^*\|_2.
\end{aligned} \tag{4.15}$$

Combining (4.14), (4.15) and $(\mathbb{E}[x])^2 \leq \mathbb{E}[x^2]$, we have

$$\tfrac{\rho_s^-}{2}(\mathbb{E}\|\widetilde{\boldsymbol{w}}^{(r)} - \boldsymbol{w}^*\|_2)^2 \leq \sqrt{s}\|\nabla\mathcal{F}(\boldsymbol{w}^*)\|_\infty\mathbb{E}\|\widetilde{\boldsymbol{w}}^{(r)} - \boldsymbol{w}^*\|_2 + \zeta. \tag{4.16}$$

Let $a = \mathbb{E}\|\widetilde{\boldsymbol{w}}^{(r)} - \boldsymbol{w}^*\|_2$. From (4.16), we solve the following quadratic function of a,

$$\frac{\rho_s^-}{2}a^2 - \sqrt{s}\|\nabla\mathcal{F}(\boldsymbol{w}^*)\|_\infty a - \zeta \leq 0,$$

which yields the bound (3.4).

Now we show that with k, η and m specified in the theorem, we have the guarantee that $\beta \leq 1$ provided appropriate choices of constants. More specifically, let $\eta \leq \frac{C_3}{\rho_s^+} \leq \frac{1}{18\rho_s^+}$, then we have $\frac{6\eta\rho_s^+}{1-6\eta\rho_s^+} \leq \frac{6C_3}{1-6C_3} \leq \frac{1}{2}$. If $k \geq C_1\kappa_s^2 k^*$ and $\eta \geq \frac{C_2}{\rho_s^+}$ with $C_2 \leq C_3$, then we have that $\alpha \leq 1 + \frac{2}{\sqrt{C_1-1}\cdot\kappa_s}$ and

$$\begin{aligned}
\frac{\alpha^m(\alpha-1)}{\eta\rho_s^-(1-6\eta\rho_s^+)(\alpha^m-1)} &\leq \frac{\frac{2}{\sqrt{C_1-1}\cdot\kappa_s}}{\frac{2C_2}{3\kappa_s}\left(1-(1+\frac{2}{\sqrt{C_1-1}\cdot\kappa_s})^{-m}\right)} \\
&= \frac{3}{C_2\sqrt{C_1-1}\left(1-(1+\frac{2}{\sqrt{C_1-1}\cdot\kappa_s})^{-m}\right)}.
\end{aligned} \tag{4.17}$$

Then (4.17) is guaranteed to be strictly smaller than $\frac{1}{2}$, i.e. $\beta < 1$, if we have

$$m \geq \log_{1+\frac{2}{\sqrt{C_1-1}\cdot\kappa_s}}\frac{C_2\sqrt{C_1-1}}{C_2\sqrt{C_1-1}-6}. \tag{4.18}$$

Using the the fact that $\ln(1+x) > x/2$ for $x \in (0,1)$, it follows that

$$\begin{aligned}
\log_{1+\frac{2}{\sqrt{C_1-1}\cdot\kappa_s}}\frac{C_2\sqrt{C_1-1}}{C_2\sqrt{C_1-1}-6} &= \frac{\log\frac{C_2\sqrt{C_1-1}}{C_2\sqrt{C_1-1}-6}}{\log 1+\frac{2}{\sqrt{C_1-1}\cdot\kappa_s}} \\
&\leq \log\frac{C_2\sqrt{C_1-1}}{C_2\sqrt{C_1-1}-6}\cdot\sqrt{C_1-1}\cdot\kappa_s.
\end{aligned}$$

Then (4.18) holds if m satisfies

$$m \geq \log\frac{C_2\sqrt{C_1-1}}{C_2\sqrt{C_1-1}-6}\cdot\sqrt{C_1-1}\cdot\kappa_s$$

If we choose $C_1 = 161^2$, $C_2 = \frac{1}{20}$, $C_3 = \frac{1}{18}$ and $C_4 = 222$, then we have $\beta \leq \frac{3}{4}$.

Part 2: Next, we demonstrate (3.5) and (3.6). Let ξ_1, ξ_2, $\xi_3, \ldots$ be a non-negative sequence of random variables, which is defined as

$$\xi_r = \max\left\{\mathcal{F}(\widetilde{\boldsymbol{w}}^{(r)}) - \mathcal{F}(\boldsymbol{w}^*) - \tfrac{3\eta}{2\left(1-6\eta\rho_s^+\right)}\|\nabla_{\widetilde{\mathcal{I}}}\mathcal{F}(\boldsymbol{w}^*)\|_2^2,\ 0\right\}.$$

For a fixed $\varepsilon > 0$, it follows from (3.3) and Markov inequality,

$$\mathbb{P}\left(\xi_r \geq \varepsilon\right) \leq \frac{\mathbb{E}\xi_r}{\varepsilon} \leq \frac{\left(\frac{3}{4}\right)^r\left[\mathcal{F}(\widetilde{\boldsymbol{w}}^{(0)}) - \mathcal{F}(\boldsymbol{w}^*)\right]}{\varepsilon}. \tag{4.19}$$

For a given $\delta \in (0,1)$, let the RHS of (4.19) be no greater than δ, which requires

$$r \geq \log_{\left(\frac{3}{4}\right)^{-1}}\frac{\mathcal{F}(\widetilde{\boldsymbol{w}}^{(0)}) - \mathcal{F}(\boldsymbol{w}^*)}{\varepsilon\delta}.$$

Therefore, we have that if $r = \left\lceil 4\log\left(\frac{\mathcal{F}(\widetilde{\boldsymbol{w}}^{(0)}) - \mathcal{F}(\boldsymbol{w}^*)}{\varepsilon\delta}\right)\right\rceil$, then (3.5) holds with probability at least $1-\delta$. Finally, (3.6) holds consequently via combining (3.4) and (3.5).

5. Experiments

We compare the empirical performance of the SVR-GHT algorithm with two other candidate algorithms: GHT proposed in (Jain et al., 2014) and SGHT proposed in (Nguyen et al., 2014) on both synthetic data and real data.

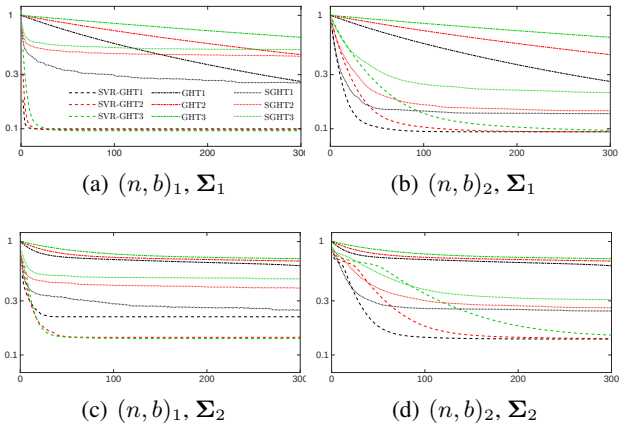
5.1. Simulated Data

We consider a sparse linear regression problem. We generate each row of the design matrix $\mathbf{A}_{i*} \in \mathbb{R}^d$ independently from $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$. The response vector is generated from the linear model $\mathbf{y} = \mathbf{A}\boldsymbol{w}^* + \boldsymbol{\epsilon}$, where $\boldsymbol{w}^* \in \mathbb{R}^d$ is the regression coefficient vector, and $\boldsymbol{\epsilon} \in \mathbb{R}^n$ is generated from $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ with $\sigma = 1$. We set $nb = 10000$, $d = 25000$, $k^* = 200$ and $k = 500$. For $\mathbf{\Sigma}$, we set $\mathbf{\Sigma}_{ii} = 1$ and $\mathbf{\Sigma}_{ij} = c$ for some constant $c \in (0, 1)$ for all $i \neq j$. The nonzero entries in $\boldsymbol{w}^*$ are sampled independently from a uniform distribution over the interval $(-2, +2)$. We divide 10000 samples into n mini batches evenly, and each mini batch contains $b = 10000/n$ samples.

We compare the performance of GHT, SGHT, and SVR-GHT for four different settings: (1) $(n, b) = (10000, 1)$, $\mathbf{\Sigma}_{ij} = 0.1$; (2) $(n, b) = (10000, 1)$, $\mathbf{\Sigma}_{ij} = 0.5$; (3) $(n, b) = (200, 50)$, $\mathbf{\Sigma}_{ij} = 0.1$; (4) $(n, b) = (200, 50)$, $\mathbf{\Sigma}_{ij} = 0.5$. For simplicity, we choose $m = n$ throughout our experiments[1]. Figure 1 illustrates the result of objective value vs. iterations under settings when $(n, b) = (10000, 1)$, and results for $(n, b) = (200, 50)$ are analogous. Since the SGHT and SVR-GHT algorithms are stochastic, we plot the objective values averaged over 50 different runs. We illustrate step sizes $\eta = 1/256, 1/512$ and $1/1024$. The horizontal axis corresponds to the number of passes over the entire dataset; computing a full gradient is counted as 1 pass, while computing a stochastic gradient is counted as $1/n$-th of a pass. The vertical axis corresponds to the ratio of current objective value over the objective value using $\widetilde{\boldsymbol{w}}^{(0)} = \mathbf{0}$. We further provide the optimal relative estimation error $\|\widetilde{\boldsymbol{w}}^{(10^6)} - \boldsymbol{w}^*\|_2 / \|\boldsymbol{w}^*\|_2$ after 10^6 effective passes of the entire dataset for each setting of three algorithms in Table 1. The optimal estimation error is obtained by averaging over 50 different runs, each of which is chosen from a sequence of step sizes $\eta \in \{1/2^5, 1/2^6, \ldots, 1/2^{14}\}$.

We see that SVR-GHT uniformly outperforms the other two candidate algorithms in terms of the convergence rate under all settings. While GHT also enjoys linear converge guarantees, its computational complexity within each iteration is n times larger than SVR-GHT; consequently, it performances much worse than SVR-GHT. Besides, we also see that SGHT converges worse than SVR-GHT in all settings. This is perhaps because the largest eigenvalue of any 500 by 500 submatrix of the covariance matrix is large (larger than 50 or 250) such that the underlying design matrix violates the Restricted Isometry Property (RIP). This might explain the poor performance of SGHT. On the other hand, the optimal estimation error of SVR-GHT is comparable to GHT, both of which outperform SGHT uniformly, especially in noisy settings. It is important to note that with the optimal step size, the estimation of GHT usually becomes stable after $> 10^5$ passes, while the estimation of SVR-GHT usually becomes stable within a few dozen to a few hundred passes, which validates the significant improvement of computational cost of SVR-GHT over GHT.

[1]Larger m results in increasing number of effective passes of the entire dataset required to achieve the same decrease of objective values, which is also observed in a closed related proximal gradient method with variance reduction (Xiao & Zhang, 2014)



(a) $(n, b)_1$, $\mathbf{\Sigma}_1$ (b) $(n, b)_2$, $\mathbf{\Sigma}_1$

(c) $(n, b)_1$, $\mathbf{\Sigma}_2$ (d) $(n, b)_2$, $\mathbf{\Sigma}_2$

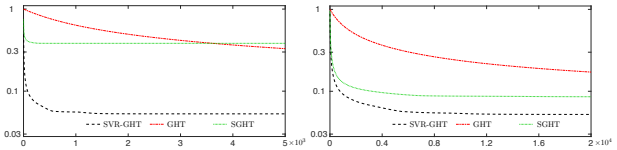
Figure 1. Comparison among three candidate algorithms under four different settings. We denote $\mathbf{\Sigma}_1 : \mathbf{\Sigma}_{ij} = 0.1$, $\mathbf{\Sigma}_2 : \mathbf{\Sigma}_{ij} = 0.5$, $(n, b)_1 = (10000, 1)$ and $(n, b)_2 = (200, 50)$. The horizontal axis is the number of passes over the entire dataset. The vertical axis is the ratio of current objective value over the objective value using $\widetilde{\boldsymbol{w}}^{(0)} = \mathbf{0}$. For each algorithm, option 1, 2 and 3 correspond to $\eta = 1/256, 1/512$ and $1/1024$ respectively. It is evident from the plots that SVR-GHT uniformly outperforms the other candidate algorithms in terms of the iteration complexity (over effective passes of data) in all settings.

Table 1. Comparison of optimal relative estimation errors between the three candidate algorithms under four different settings on simulated data sets. We denote $\mathbf{\Sigma}_1 : \mathbf{\Sigma}_{ij} = 0.1$, $\mathbf{\Sigma}_2 : \mathbf{\Sigma}_{ij} = 0.5$, $(n, b)_1 = (10000, 1)$ and $(n, b)_2 = (200, 50)$. SVR-GHT achieves comparable result with GHT, both of which uniformly outperforms SGHT in each of the eight settings.

Method		GHT	SGHT	SVR-GHT
$\mathbf{\Sigma}_1$	$(n, b)_1$	0.00851	0.02490	0.00968
	$(n, b)_2$		0.06412	0.00970
$\mathbf{\Sigma}_2$	$(n, b)_1$	0.02940	0.21676	0.02614
	$(n, b)_2$		0.18764	0.02823

5.2. Real Data

We adopt a subset of RCV1 dataset with 9625 documents and 29992 distinct words, including the classes of "C15", "ECAT", "GCAT", and "MCAT" (Cai & He, 2012). We apply logistic regression to perform binary classification for



(a) $(n, b) = (5000, 1)$ (b) $(n, b) = (100, 50)$

Figure 2. Comparison among the three candidate algorithms under two different settings on the RCV1 data set for class C15, ECAT, GCAT and MCAT. The horizontal axis corresponds to the number of passes over the entire dataset. The vertical axis corresponds to misclassification rate on the test data. It is evident from the plots that SVR-GHT uniformly outperforms the other candidate algorithms in both settings.

all classes, each of which uses 5000 documents for training, i.e., $nb = 5000$ and $d = 29992$, with the same proportion of documents from each class, and the rest for testing. We illustrates the computational performance of the GHT, SGHT, and SVR-GHT algorithms with two different settings for each class: Setting (1) has $(n, b) = (5000, 1)$; Setting (2) has $(n, b) = (100, 50)$. We choose $k = 200$ and $m = n$ for both settings of all classes. For all three algorithms, we plot their objective values and provide the optimal classification errors averaged over 10 different runs using different data separations. Figure 2 demonstrates the result for "C15", and the other classes have analogous performance. Similar to the synthetic numerical evaluations, SVR-GHT uniformly outperforms the other two candidate algorithms in terms of the convergence rate under both settings. We provide the optimal misclassification rates of all classes for three algorithms in Table 2, where the optimal step size η for each algorithm is chosen from a sequence of values $\{1/2^5, 1/2^6, \ldots, 1/2^{14}\}$.. Similar to the simulated data sets again, the optimal misclassification rate of SVR-GHT is comparable to GHT, both of which outperform SGHT uniformly. The estimation of GHT generally requires $> 10^6$ passes to become stable, while the estimation of SVR-GHT generally requires a few hundred to a few thousand passes to be stable, which validates the significant improvement of computational cost of SVR-GHT over GHT for this real dataset.

6. Discussion

The SVR-GHT algorithm presented in this paper is closely related to some recent work on stochastic optimization algorithms, including Prox-SVRG algorithm (Xiao & Zhang, 2014), stochastic averaging gradient (SAG) algorithm (Roux et al., 2012) and stochastic dual coordinate ascent algorithm (SDCA, (Shalev-Shwartz & Zhang, 2013)). However, the focus in these previous works has been on establishing global linear convergence for optimization problems involving strongly convex objective with a convex constraint, whereas SVR-GHT guarantees linear convergence for optimization problems involving a nonstrongly convex objective with nonconvex cardinality constraint.

Table 2. Comparison of optimal classification errors among the three candidate algorithms for both settings of all four classes. We denote $(n, b)_1 = (5000, 1)$ and $(n, b)_2 = (100, 50)$. SVR-GHT achieves comparable result with GHT, both of which uniformly outperforms SGHT in all settings.

Method		GHT	SGHT	SVR-GHT
C15	$(n, b)_1$	0.02844	0.03259	0.02826
	$(n, b)_2$		0.03361	0.02867
ECAT	$(n, b)_1$	0.05581	0.06851	0.05628
	$(n, b)_2$		0.07179	0.05631
GCAT	$(n, b)_1$	0.03028	0.06263	0.03354
	$(n, b)_2$		0.09142	0.03444
MCAT	$(n, b)_1$	0.05703	0.07638	0.05877
	$(n, b)_2$		0.08228	0.05927

Other related work includes nonconvex regularized M-estimators proposed by (Loh & Wainwright, 2013). In particular, they consider the nonconvex optimization problem:

$$\min_{\boldsymbol{w}} \mathcal{F}(\boldsymbol{w}) + \mathcal{P}_{\lambda,\gamma}(\boldsymbol{w}) \quad \text{s.t. } \|\boldsymbol{w}\|_1 \leq R, \tag{6.1}$$

where $\mathcal{P}_{\lambda,\gamma}(\boldsymbol{w})$ is a nonconvex regularization function with tuning parameters λ and γ; Popular choices for $\mathcal{P}_{\lambda,\gamma}(\boldsymbol{w})$ are the SCAD (Fan & Li, 2001) and MCP (Zhang, 2010) regularization functions. (Loh & Wainwright, 2013) show that under restricted strong convexity and restricted strong smoothness conditions, similar to those studied here, the proximal gradient algorithm attains linear convergence to approximate global optima with optimal estimation accuracy. Accordingly, one could adopt the Prox-SVRG algorithm to solve (6.1) in a stochastic fashion, and trim the analyses in (Xiao & Zhang, 2014) and (Loh & Wainwright, 2013) to establish similar convergence guarantees. We remark, however, that Problem (6.1) involves three tuning parameters, λ, γ, and R which, in practice, requires enormous tuning effort to attain good empirical performance. In contrast, Problem (1.1) involves a single tuning parameter, k, which makes tuning more efficient.

Acknowledgements

This research is supported by NSF CCF-1217751; NSF AST-1247885; DARPA Young Faculty Award N66001-14-1-4047; NSF DMS-1454377-CAREER; NSF IIS-1546482-BIGDATA; NIH R01MH102339; NSF IIS-1408910; NSF IIS-1332109; NIH R01GM083084.

References

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

Banerjee, Onureena, El Ghaoui, Laurent, and d'Aspremont, Alexandre. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *The Journal of Machine Learning Research*, 9:485–516, 2008.

Blumensath, Thomas and Davies, Mike E. Iterative hard thresholding for compressed sensing. *Appl. Comp. Harm. Anal.*, 27 (3):594–607, 2009.

Cai, Deng and He, Xiaofei. Manifold adaptive experimental design for text categorization. *Knowledge and Data Engineering, IEEE Transactions on*, 24(4):707–719, 2012.

Fan, Jianqing and Li, Runze. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

Foucart, Simon. Hard thresholding pursuit: An algorithm for compressive sensing. *SIAM J. Numer. Anal.*, 49(6):2543–2563, 2011.

Friedman, Jerome, Hastie, Trevor, and Tibshirani, Robert. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

Jain, Prateek, Tewari, Ambuj, and Kar, Purushottam. On iterative hard thresholding methods for high-dimensional m-estimation. In *NIPS*, pp. 685–693. 2014.

Johnson, Rie and Zhang, Tong. Accelerating stochastic gradient descent using predictive variance reduction. In *NIPS*, pp. 315–323. 2013.

Konečnỳ, Jakub and Richtárik, Peter. Semi-stochastic gradient descent methods. *arXiv preprint arXiv:1312.1666*, 2013.

Loh, Po-Ling and Wainwright, Martin J. Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. In *Advances in Neural Information Processing Systems*, pp. 476–484, 2013.

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

Nesterov, Yu. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013.

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

Raskutti, Garvesh, Wainwright, Martin J, and Yu, Bin. Restricted eigenvalue properties for correlated gaussian designs. *The Journal of Machine Learning Research*, 11:2241–2259, 2010.

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

Roux, Nicolas L., Schmidt, Mark, and Bach, Francis R. A stochastic gradient method with an exponential convergence rate for finite training sets. In *NIPS*, pp. 2663–2671. 2012.

Shalev-Shwartz, Shai and Zhang, Tong. Stochastic dual coordinate ascent methods for regularized loss. *JMLR*, 14(1):567–599, 2013.

Shen, Xiaotong, Pan, Wei, and Zhu, Yunzhang. Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association*, 107(497):223–232, 2012.

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

Van de Geer, Sara A. High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, pp. 614–645, 2008.

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

Yuan, Ming and Lin, Yi. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

Yuan, Xiao-Tong, Li, Ping, and Zhang, Tong. Gradient hard thresholding pursuit for sparsity-constrained optimization. In *ICML*, pp. 71–79. 2013.

Zhang, Cun-Hui. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, pp. 894–942, 2010.

Zhao, Tuo and Liu, Han. Accelerated path-following iterative shrinkage thresholding algorithm with application to semiparametric graph estimation. *Journal of Computational and Graphical Statistics*, 2016. To appear.

Zhao, Tuo, Liu, Han, and Zhang, Tong. A general theory of pathwise coordinate optimization. *arXiv preprint arXiv:1412.7477*, 2014a.

Zhao, Tuo, Yu, Mo, Wang, Yiming, Arora, Raman, and Liu, Han. Accelerated mini-batch randomized block coordinate descent method. In *Advances in neural information processing systems*, pp. 3329–3337, 2014b.