Journal of Machine Learning Research 18 (2018) 1-43 Submitted 10/14; Revised 5/17; Published 5/18

Gradient Hard Thresholding Pursuit

Xiao-Tong Yuan XTYUAN@NUIST.EDU.CN
B-DAT Lab, Nanjing University of Information Science and Technology
Nanjing 210044, China

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

Tong Zhang TONGZHANG@TONGZHANG-ML.ORG
Tencent AI Lab
Shenzhen 518057, China

Editor: Yoram Singer

Abstract

Hard Thresholding Pursuit (HTP) is an iterative greedy selection procedure for finding sparse solutions of underdetermined linear systems. This method has been shown to have strong theoretical guarantee and impressive numerical performance. In this article, we generalize HTP from compressed sensing to a generic problem setup of sparsity-constrained convex optimization. The proposed algorithm iterates between a standard gradient descent step and a hard-thresholding step with or without debiasing. We analyze the parameter estimation and sparsity recovery performance of the proposed method. Extensive numerical results confirm our theoretical predictions and demonstrate the superiority of our method to the state-of-the-art greedy selection methods in sparse linear regression, sparse logistic regression and sparse precision matrix estimation problems.[1]

Keywords: Hard Thresholding Pursuit, Sparsity Recovery, Greedy Selection

1. Introduction

In the past decade, high-dimensional data analysis has received broad research interest in data mining and scientific discovery, with many significant results obtained in theory, algorithm and application. The major driving force is the rapid development of data collection technologies in many application domains such as social networks, natural language processing, bioinformatics and computer vision. In these applications it is not unusual that data samples are represented with millions or even billions of features using which an underlying statistical learning model must be fit. In many circumstances, however, the number of collected samples is substantially smaller than the dimensionality of features, implying that consistent estimators cannot be hoped for unless additional assumptions are imposed on the model. One of the most popular prior assumptions is that the data exhibit low-dimensional structure, which can often be captured by imposing sparsity constraint on model parameter space. It is thus crucial to develop robust and efficient computational procedures for high-dimensional estimation with sparsity constraint.

1. A conference version of this work appeared in ICML 2014 (Yuan et al., 2014).

In this article, we consider the following generic sparsity-constrained loss minimization problem:

$$\min_{x\in\mathbb{R}^p} f(x), \quad \text{s.t. } \|x\|_0 \leq k, \tag{1}$$

where $f : \mathbb{R}^p \mapsto \mathbb{R}$ is a smooth convex loss function and $\|x\|_0$ denotes the number of nonzero entries in parameter vector x. Among others, several popular examples falling into this framework include: (i) Sparsity-constrained linear regression model (Tropp & Gilbert, 2007) where the residual error is used to measure data reconstruction error; (ii) Sparsity-constrained logistic regression model (Bahmani et al., 2013) where the sigmoid loss is used to measure prediction error; (iii) Sparsity-constrained graphical model learning (Jalali et al., 2011) where the likelihood of samples drawn from an underlying probabilistic model is used to measure data fidelity.

Due to the presence of cardinality constraint $\|x\|_0 \leq k$, problem (1) is generally NP-hard even for the quadratic loss function (Natarajan, 1995). Thus, one must instead seek approximate solutions. For the special case of (1) with least squares error loss in compressed sensing (Donoho, 2006), a number of low-complexity greedy pursuit methods have been studied including matching pursuit (MP) (Mallat & Zhang, 1993), orthogonal matching pursuit (OMP) (Pati et al., 1993), iterative hard thresholding (IHT) (Blumensath & Davies, 2009), compressed sampling matching pursuit (CoSaMP) (Needell & Tropp, 2009) and hard thresholding pursuit (HTP) (Foucart, 2011) to name a few. These algorithms successively select the position of nonzero entries and estimate their values via exploring the residual error from the previous iteration. Comparing to those first-order convex optimization methods developed for ℓ_1-regularized sparse learning (Beck & Teboulle, 2009; Langford et al., 2009; Agarwal et al., 2012), these greedy pursuit algorithms often exhibit more attractive computational efficiency and scalability in practice.

The least squares error used in compressed sensing, however, is not an appropriate measure of discrepancy in a variety of applications beyond signal processing. For example, in statistical machine learning the log-likelihood function is commonly used in logistic regression (Bishop, 2006) and graphical model learning (Jalali et al., 2011; Ravikumar et al., 2011). Thus, it is desirable to investigate theory and algorithms applicable to a broader class of sparse learning problems as formulated by (1). To this end, several forward selection algorithms have been proposed to select the nonzero entries in a sequential fashion (Kim & Kim, 2004; Shalev-Shwartz et al., 2010; Yuan & Yan, 2013; Jaggi, 2011). This category of methods dates back to the Frank-Wolfe method (Frank & Wolfe, 1956). In the meanwhile, the forward greedy selection method has been generalized to convex loss minimization over the linear hull of a collection of atoms (Tewari et al., 2011; Yuan & Yan, 2013). To make the greedy selection procedure more adaptive, Zhang (2008) proposed a forward-backward algorithm which takes backward steps adaptively whenever beneficial. Jalali et al. (2011) have applied this forward-backward selection method to learn the sparse structure of graphical model. Bahmani et al. (2013) proposed a gradient support pursuit method that generalizes CoSaMP from compressed sensing to the generic sparse minimization problem (1). Jain et al. (2014) presented and analyzed several HTP/IHT-style algorithms for high-dimensional sparse estimation. In the paper of Blumensath (2013), a nonlinear-IHT algorithm was investigated in the generic setting of sparsity-constrained loss minimization. Recently, the extensions of HTP/IHT-style methods to structured and stochastic sparse estimation have

been extensively studied in machine learning community (Jain et al., 2016; Li et al., 2016; Shen & Li, 2016; Liu et al., 2017; Nguyen et al., 2017).

1.1 Overview of Our Contribution

In this article, inspired by the success of Hard Thresholding Pursuit (HTP) (Foucart, 2011, 2012) in compressed sensing, we propose and analyze the Gradient Hard Thresholding Pursuit (GraHTP) method to encompass the sparse estimation problems arising from applications with general nonlinear models. At each iteration, GraHTP performs standard gradient descent followed by a hard thresholding operation which first selects the top k (in magnitude) entries of the resultant vector and then (optionally) conducts debiasing on the selected entries. We show that in various settings with or without assuming RIP-type conditions, GraHTP has strong theoretical guarantees analogous to HTP in terms of parameter estimation accuracy.

Apart from the accuracy of objective value and parameter estimation, in many applications such as compressed sensing and graphical models learning, one property of central importance for sparse estimation is the recovery of sparsity pattern, which corresponds to the set of indices of nonzero components of the model parameters. Once the sparsity pattern is recovered, computing the actual nonzero coefficients just boils down to solving a convex minimization problem over the supporting indices. For perfect measurements, the results obtained by Foucart (2011) show that under proper conditions HTP can exactly recover the underlying true model parameters. For noisy models, however, the sparsity recovery analysis is a crucial challenge remains unsolved for HTP-style methods. As a core contribution of this work, we provide a systematic sparsity recovery analysis for GraHTP. Since the output of GraHTP is always k-sparse, the parameter estimation error bounds established in this article roughly imply a sufficient condition for sparsity recovery: as long as the smallest (in magnitude) nonzero entry of a k-sparse target model is larger than the estimation error bound, exact recovery of such a target model can be guaranteed. With more insightful analysis, we further derive some refined sparsity recovery results for GraHTP and for the k-sparse minimizer of problem (1) as well. Some preliminary results on sparsity recovery of GraHTP have been presented in a prior work of ours (Yuan et al., 2016), which we have improved largely in this article.

Comparing to the prior analysis for HTP-style methods, the merits of our main results can be distilled to the following two aspects:

- **Parameter estimation accuracy analysis with/without RIP-type conditions.** Our parameter estimation accuracy analysis for GraHTP simultaneously covers the setting where the target solution is an arbitrary k-sparse solution for which the RIP-type conditions are required, and the setting where the target solution is certain $\bar{k}$-sparse solutions with $\bar{k} \ll k$ for which the RIP-type conditions can be waived;

- **Systematic sparsity recovery analysis.** We extensively investigate the sparsity recovery performance of GraHTP which is of great importance and practical vale in many sparse learning applications including compressed sensing and graphical models learning.

Results	Target Solution	RIP Cond. Free	Sparsity Recovery
(Foucart, 2011)	True k-sparse signal x	$\times$	$\times$
(Blumensath, 2013)	$x^\star = \arg\min_{\|x\|_0 \le k} f(x)$	$\times$	$\times$
(Jain et al., 2014)	$\bar{x} = \arg\min_{\|x\|_0 \le \bar{k}} f(x)$ for proper $\bar{k} \ll k$	$\surd$	$\times$
This Work	$\bar{x}$ with $\|\bar{x}\|_0 \le k$	$\times$ (for $\|\bar{x}\|_0 = k$), $\surd$ (for $\|\bar{x}\|_0 \ll k$)	$\surd$

Table 1: Comparison between the results obtained in this work and several representative prior results for HTP-style algorithms.

Table 1 summarizes a high level comparison between our results and several representative state-of-the-art results for HTP-style algorithms, in terms of target solution, dependence on RIP-type conditions, and sparsity recovery analysis.

We have applied GraHTP to sparse linear regression, sparse logistic regression and sparse precision matrix estimation problems, with its algorithm and/or theory substantialized for these models. Empirically we demonstrate that GraHTP is competitive to the state-of-the-art greedy selection methods in these sparse learning problems.

1.2 Notation

In the following, x is a vector, A is a matrix, and F is an index set. The following notations will be used in this article.

- $[x]_i$: the ith entry of vector x.
- x_F: the restriction of x on F, i.e., $[x_F]_i = [x]_i$ if $i \in F$, and $[x_F]_i = 0$ otherwise.
- x_k: the restriction of x on its top k (in modulus) entries.
- $\|x\| = \sqrt{x^\top x}$: the Euclidean norm of x.
- $\|x\|_1 = \sum_i |[x]_i|$: the ℓ_1-norm of x.
- $\|x\|_\infty = \max_i |[x]_i|$: the ℓ_∞-norm of x.
- $\|x\|_0$: the number of nonzero entries of x.
- $\mathrm{supp}(x)$: the index set of nonzero entries of x.
- $\mathrm{supp}(x, k)$: the index set of the top k (in modulus) entries of x.
- $x_{\min} = \min_{i \in \mathrm{supp}(x)} |[x]_i|$: the smallest absolute value of nonzero element of x.
- $[A]_{ij}$: the element on the ith row and jth column of matrix A.

- $\|A\| = \sup_{\|x\|\leq 1} \|Ax\|$: the spectral norm of matrix A.
- $|A|_\infty = \max_{i,j} |[A]_{ij}|$: the element-wise ℓ_∞-norm of A.
- $\mathrm{Tr}(A)$: the trace (sum of diagonal elements) of a square matrix A.
- A_F: the restriction of A on index set F.
- A^-: the restriction of a square matrix A on its off-diagonal entries.
- $\mathrm{vect}(A)$: (column wise) vectorization of a matrix A.
- $\lambda_{\max}(A, k) = \max_{\|x\|=1, \|x\|_0 \leq k} x^\top Ax$: the largest k-sparse eigenvalue of a positive semi-definite matrix A.
- $\lambda_{\min}(A, k) = \min_{\|x\|=1, \|x\|_0 \leq k} x^\top Ax$: the smallest k-sparse eigenvalue of a positive semi-definite matrix A.

1.3 Organization

This article proceeds as follows: We present in Section 2 the GraHTP algorithm. The parameter estimation error and exact sparsity recovery guarantees of GraHTP are respectively analyzed in Section 3 and Section 4. The implications of GraHTP in linear regression, logistic regression and Gaussian graphical model learning are discussed in Section 5. Monte-Carlo simulations and real data experimental results are presented in Section 6. We conclude this article in Section 7.

2. Algorithm

GraHTP is an iterative greedy selection procedure for approximately optimizing the non-convex problem (1). A high level summary of GraHTP is described in the top panel of Algorithm 1. The procedure generates a sequence of intermediate k-sparse vectors $x^{(0)}, x^{(1)}, \ldots$ from an initial sparse approximation $x^{(0)}$ (typically $x^{(0)} = 0$). At the t-th iteration, the first step (**S1**), $\tilde{x}^{(t)} = x^{(t-1)} - \eta \nabla f(x^{(t-1)})$, computes the gradient descent at the point $x^{(t-1)}$ with step-size η. Then in the second step (**S2**), the k coordinates of the vector $\tilde{x}^{(t)}$ that have the largest magnitude are chosen as the support in which pursuing the minimization will be most effective. In the third step (**S3**), we find a vector with this support which minimizes the objective function, which becomes $x^{(t)}$. This last step, which is often referred to as *debiasing*, has been shown to improve the performance in other algorithms too (Yuan & Zhang, 2013; Bahmani et al., 2013). The iterations continue until the algorithm reaches certain terminating condition, e.g., the difference of objective value or model parameters between adjacent iterations converges. A more intuitive criterion is $F^{(t)} = F^{(t-1)}$ (see (**S2**) for the definition of $F^{(t)}$), since then $x^{(\tau)} = x^{(t)}$ for all $\tau \geq t$, although there is no guarantee that this should occur in general cases. It will be assumed throughout the article that the sparsity level k is known. In practice this integer parameter may be tuned via, for example, cross-validation in supervised learning tasks.

In the standard form of GraHTP, the debiasing step (**S3**) requires to minimize $f(x)$ over the supporting set $F^{(t)}$. If this step is judged too costly, we may consider instead a

fast variant of GraHTP, where the debiasing is replaced by a simple truncation operation $x^{(t)} = \tilde{x}_k^{(t)}$. This leads to the Fast GraHTP (FGraHTP) as described in the bottom panel of Algorithm 1, which can be understood as a projected gradient descent procedure for optimizing the nonconvex minimization problem (1). Up to the cost of truncation operation, its per-iteration computational overload is almost identical to that of the standard gradient descent procedure. The iteration procedure of FGraHTP is also known as the nonlinear-IHT algorithm (Blumensath, 2013). Comparing to that prior work, our analysis for FGraHTP is more comprehensive and the results are tighter especially in sparsity recovery analysis. While in this article we only study the FGraHTP outlined in Algorithm 1, we should mention that other fast variants of GraHTP can also be considered. For instance, to reduce the computational cost of the debiasing step (**S3**), we can take a restricted Newton step or a restricted gradient descent step to calculate $x^{(t)}$.

We close this section by pointing out that, in the special case where the squares error $f(x) = \frac{1}{2}\|y - Ax\|^2$ is the cost function, GraHTP reduces to HTP (Foucart, 2011). Specifically, the gradient descent step (**S1**) reduces to $\tilde{x}^{(t)} = x^{(t-1)} + \eta A^\top(y - Ax^{(t-1)})$ and the debiasing step (**S3**) reduces to the orthogonal projection $x^{(t)} = \arg\min\{\frac{1}{2}\|y - Ax\|^2, \mathrm{supp}(x) \subseteq F^{(t)}\}$. In the meanwhile, FGraHTP reduces to IHT (Blumensath & Davies, 2009), which is also known as Gradient Descent with Sparsification (Garg & Khandekar, 2009), of which the iteration is defined as $x^{(t)} = (x^{(t-1)} + \eta A^\top(y - Ax^{(t-1)}))_k$.

Algorithm 1: Gradient Hard Thresholding Pursuit (GraHTP).

Initialization: $x^{(0)}$ *with* $\|x^{(0)}\|_0 \le k$ *(typically* $x^{(0)} = 0$*),* $t = 1$.
Output: $x^{(t)}$.
repeat
  (**S1**) Compute $\tilde{x}^{(t)} = x^{(t-1)} - \eta \nabla f(x^{(t-1)})$;
  (**S2**) Let $F^{(t)} = \mathrm{supp}(\tilde{x}^{(t)}, k)$ be the indices of $\tilde{x}^{(t)}$ with the largest k absolute values;
  (**S3**) Compute $x^{(t)} = \arg\min\{f(x); \mathrm{supp}(x) \subseteq F^{(t)}\}$;
  $t = t + 1$;
until *halting condition holds*;

★ *Fast GraHTP* ★

repeat
  Compute $\tilde{x}^{(t)} = x^{(t-1)} - \eta \nabla f(x^{(t-1)})$;
  Compute $x^{(t)} = \tilde{x}_k^{(t)}$ as the truncation of $\tilde{x}^{(t)}$ with top k (in magnitude) entries preserved;
  $t = t + 1$;
until *halting condition holds*;

3. Parameter Estimation Analysis

In this section, we analyze the parameter estimation accuracy of GraHTP/FGraHTP. To simplify notation, we abbreviate $\nabla_F f = (\nabla f)_F$ and $\nabla_s f = (\nabla f)_s$. Our analysis relies on the conditions of Restricted Strong Convexity/Smoothness (RSC/RSS) which are conven-

tionally used in the analysis of greedy sparse optimization methods (Shalev-Shwartz et al., 2010; Bahmani et al., 2013; Jain et al., 2014).

Definition 1 (Restricted Strong Convexity/Smoothness) *For any integer $s > 0$, we say $f(x)$ is restricted m_s-strongly convex and M_s-smooth if there exist $m_s, M_s > 0$ such that*

$$\frac{m_s}{2}\|x-y\|^2 \leq f(x) - f(y) - \langle \nabla f(y), x-y\rangle \leq \frac{M_s}{2}\|x-y\|^2, \quad \forall \|x-y\|_0 \leq s. \tag{2}$$

The ratio number M_s/m_s, which measures the curvature of the loss function over sparse subspaces, will be referred to as *restricted strong condition number* in this article.

3.1 Main Results

The following theorem is our main result on the parameter estimation accuracy of GraHTP and FGraHTP with respect to arbitrary k-sparse target solutions. A proof of this theorem is provided in Appendix B.1.

Theorem 2 *Assume that f is M_{3k}-smooth and m_{3k}-strongly convex. Let $\bar{x}$ be an arbitrary k-sparse vector and $\rho = \sqrt{1 - 2\eta m_{3k} + \eta^2 M_{3k}^2}$.*

(a) Assume that $M_{3k}/m_{3k} < 2\sqrt{3}/3$ and the step-size η is chosen such that $\rho < 0.5$. Then GraHTP outputs $x^{(t)}$ satisfying

$$\|x^{(t)} - \bar{x}\| \leq \mu_1^t \|x^{(0)} - \bar{x}\| + \frac{2.83\eta\sqrt{k}}{1-2\rho}\|\nabla f(\bar{x})\|_\infty,$$

where $\mu_1 = \rho/(1-\rho) \in (0,1)$.

(b) Assume that $M_{3k}/m_{3k} < 1.26$ and the step-size η is chosen such that $\rho < 0.62$. Then FGraHTP outputs $x^{(t)}$ satisfying

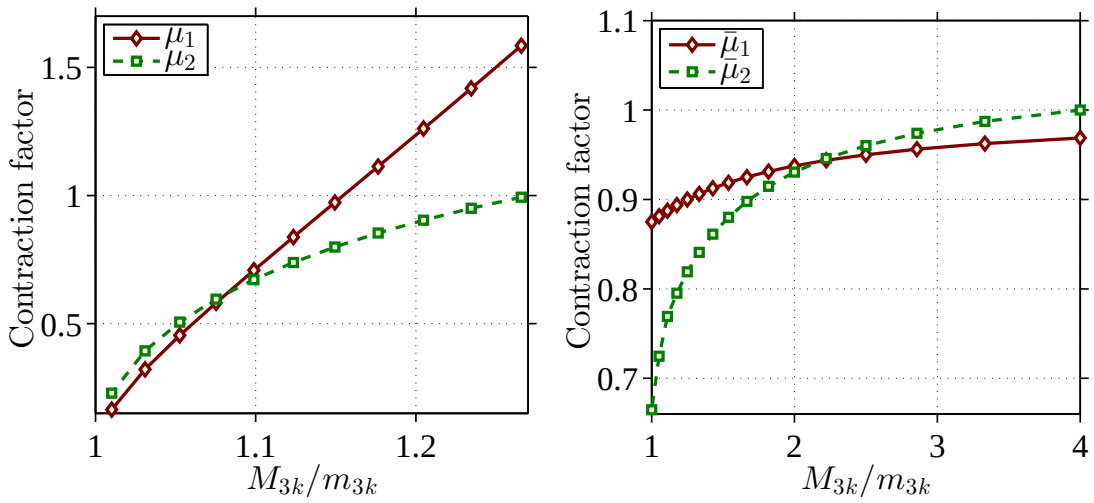
$$\|x^{(t)} - \bar{x}\| \leq \mu_2^t \|x^{(0)} - \bar{x}\| + \frac{2.81\eta\sqrt{k}}{1-1.62\rho}\|\nabla f(\bar{x})\|_\infty,$$

where $\mu_2 = 1.62\rho \in (0,1)$.

In the part (a) of Theorem 2, the contraction factor $\mu_1 < 1$ controls the convergence rate of GraHTP. The condition $\rho < 0.5$ requires the step-size to be selected according to

$$\frac{2m_{3k} - \sqrt{4m_{3k}^2 - 3M_{3k}^2}}{2M_{3k}^2} < \eta < \frac{2m_{3k} + \sqrt{4m_{3k}^2 - 3M_{3k}^2}}{2M_{3k}^2}, \tag{3}$$

from which we can see that $M_{3k}/m_{3k} < 2\sqrt{3}/3$ is a necessary condition to guarantee the existence of η such that $\rho < 0.5$ and $\mu_1 < 1$. The condition of $\rho < 0.5$ is analogous to the RIP condition for estimation from noisy measurements in compressed sensing (Candès et al., 2006; Needell & Tropp, 2009; Foucart, 2011). Indeed, in compressed sensing, GraHTP reduces to HTP which requires weaker RIP condition than prior compressed sensing algorithms. The condition in (3) also suggests that the value of η should be bounded from



(a) Theorem 2 (b) Theorem 5

Figure 1: Evolving curves of the contraction factors in Theorem 2 and Theorem 5.

above to guarantee convergence, and be bounded away from zero to avoid early stopping as well. Similarly, $M_{3k}/m_{3k} < 1.26$ in the part(b) is a necessary condition to guarantee the existence of η such that $\rho < 0.62$ and $\mu_2 < 1$. Figure 1(a) shows the evolving curves of contraction factors μ_1 and μ_2 as functions of M_{3k}/m_{3k} in the interval $[1, 1.26)$. It can be seen from this figure that $\mu_1 < \mu_2$ when $M_{3k}/m_{3k} \to 1$ and $\mu_1 > \mu_2$ for relatively larger M_{3k}/m_{3k}.

The non-vanishing terms in the error bounds of Theorem 2 indicate that the estimation errors of GraHTP and FGraHTP are controlled by the multiplier of $\sqrt{k}\|\nabla f(\bar{x})\|_\infty$. Particularly if the sparse vector $\bar{x}$ is sufficiently close to an unconstrained minimum of f, then the estimation error floor is negligible because $\|\nabla f(\bar{x})\|_\infty$ has small magnitude. The following corollary is a direct consequence of Theorem 2 which shows that exact support recovery is possible when $\bar{x}_{\min}$ is significantly larger than $\sqrt{k}\|\nabla f(\bar{x})\|_\infty$.

Corollary 3 *Assume the conditions in Theorem 2 hold.*

(a) Let $\bar{x}$ be an arbitrary k-sparse vector satisfying $\bar{x}_{\min} > \frac{5.66\eta\sqrt{k}}{1-2\rho}\|\nabla f(\bar{x})\|_\infty$. Then GraHTP will output $x^{(t)}$ satisfying $supp(x^{(t)}) = supp(\bar{x})$ after $t = \left\lceil \frac{1}{\mu_1} \ln\left(\frac{2\|x^{(0)}-\bar{x}\|}{\bar{x}_{\min}}\right)\right\rceil$ steps of iteration.

(b) Let $\bar{x}$ be an arbitrary k-sparse vector satisfying $\bar{x}_{\min} > \frac{5.62\eta\sqrt{k}}{1-1.62\rho}\|\nabla f(\bar{x})\|_\infty$. Then FGraHTP will output $x^{(t)}$ satisfying $supp(x^{(t)}) = supp(\bar{x})$ after $t = \left\lceil \frac{1}{\mu_2} \ln\left(\frac{2\|x^{(0)}-\bar{x}\|}{\bar{x}_{\min}}\right)\right\rceil$ steps of iteration.

Indeed, given the conditions in Corollary 3, for both GraHTP and FGraHTP we can show that $\|x^{(t)} - \bar{x}\| < \bar{x}_{\min}$ and thus $\text{supp}(x^{(t)}) = \text{supp}(\bar{x})$ must hold as $x^{(t)}$ and $\bar{x}$ are both k-sparse vectors.

Remark 4 *Corollary 3 shows that GraHTP/FGraHTP requires RIP-type conditions as in Theorem 2 to guarantee exact support recovery. As a comparison, the existing sparsity recovery results for ℓ_1-estimators (Wainwright, 2009; Li et al., 2015) are free of RIP-type conditions but instead relying on the irrepresentablility condition which is known to be stronger. For example, a case where the RIP-type condition holds while the irrepresentability condition does not was given by Van De Geer & Bühlmann (2009, Example 10.4).*

The RIP-type conditions assumed in Theorem 2 could still be restrictive in real-life high-dimensional statistical settings wherein pairs of variables can be arbitrarily correlated. In the following theorem, we further show that by properly relaxing sparsity levels, GraHTP and FGraHTP are able to accurately estimate parameters without assuming bounded restricted strong condition numbers. A proof of this theorem is deferred to Appendix B.2.

Theorem 5 *Let $\bar{x}$ be an arbitrary $\bar{k}$-sparse vector with $\bar{k} \leq k$. Assume that $s = 2k + \bar{k} < p$.*

(a) Assume that f is M_{2k}-smooth and m_{2k}-strongly convex. Assume the step-size $\eta < 1/M_{2k}$. If $k \geq \left(2 + \frac{4}{\eta^2 m_{2k}^2}\right)\bar{k}$, then GraHTP outputs $x^{(t)}$ satisfying

$$\|x^{(t)} - \bar{x}\| \leq \sqrt{\frac{2\bar{\mu}_1^t \bar{\triangle}^{(0)}}{m_{2k}}} + \frac{2.83\sqrt{k}\|\nabla f(\bar{x})\|_\infty}{m_{2k}},$$

where $\bar{\mu}_1 = 1 - \eta m_{2k}(1 - \eta M_{2k})/2$ and $\bar{\triangle}^{(0)} = \max\{f(x^{(0)}) - f(\bar{x}), 0\}$.

(b) Assume that f is M_s-smooth and m_s-strongly convex. Assume the step-size $\eta < 2m_s/M_s^2$ such that $\rho = \sqrt{1 - 2\eta m_s + \eta^2 M_s^2} < 1$. If $k > \rho\bar{k}/(1-\rho)^2$, then FGraHTP outputs $x^{(t)}$ satisfying

$$\|x^{(t)} - \bar{x}\| \leq \bar{\mu}_2^t \|x^{(0)} - \bar{x}\| + \frac{\gamma\eta\sqrt{s}}{1 - \bar{\mu}_2}\|\nabla f(\bar{x})\|_\infty,$$

where $\bar{\mu}_2 = \rho\gamma \in (0, 1)$ and $\gamma = \sqrt{1 + \left(\bar{k}/k + \sqrt{(4 + \bar{k}/k)\bar{k}/k}\right)/2}$.

Remark 6 *When using step-size $\eta = \frac{1}{2M_{2k}}$, the part(a) of Theorem 5 tells that GraHTP converges linearly towards an arbitrary $\bar{k}$-sparse vector $\bar{x}$ if the sparsity level is chosen as $k \geq \left(2 + \frac{16M_{2k}^2}{m_{2k}^2}\right)\bar{k}$. The estimation error is controlled by the multiplier of $\sqrt{k}\|\nabla f(\bar{x})\|_\infty$. Similarly, the part(b) of Theorem 5 establishes the convergence result of FGraHTP with proper relaxed $k \gg \bar{k}$. Note that the condition $k > \rho\bar{k}/(1-\rho)^2$ in part(b) actually enforces the contraction factor $\bar{\mu}_2 < 1$. Figure 1(b) shows the evolving curves of contraction factors $\bar{\mu}_1$ and $\bar{\mu}_2$ as functions of M_{3k}/m_{3k}, with the same target sparsity $\bar{k}$. We can see from this figure that $\bar{\mu}_2$ is superior to $\bar{\mu}_1$ when M_{3k}/m_{3k} is relatively small.*

The following corollary of Theorem 5 shows that GrHTP/FGraHTP with certain relaxed sparsity levels can guarantee $\mathrm{supp}(\bar{x}) \subseteq \mathrm{supp}(x^{(t)})$ without assuming RIP-type conditions.

Corollary 7 *Let $\bar{x}$ be an arbitrary $\bar{k}$-sparse vector with $\bar{k} \leq k$.*

(a) *Under the conditions in Theorem 5(a), if* $\bar{x}_{\min} > \frac{5.66\sqrt{k}}{m_{2k}}\|\nabla f(\bar{x})\|_\infty$, *then GraHTP will output* $x^{(t)}$ *satisfying* $supp(\bar{x}) \subseteq supp(x^{(t)})$ *after* $t = \left\lceil \frac{1}{\bar{\mu}_1} \ln\left(\frac{8\bar{\triangle}^{(0)}}{m_{2k}\bar{x}^2_{\min}}\right)\right\rceil$ *steps of iteration.*

(b) *Under the conditions in Theorem 5(b), if* $\bar{x}_{\min} > \frac{2\gamma\eta\sqrt{s}}{1-\bar{\mu}_2}\|\nabla f(\bar{x})\|_\infty$, *then FGraHTP will output* $x^{(t)}$ *satisfying* $supp(\bar{x}) \subseteq supp(x^{(t)})$ *after* $t = \left\lceil \frac{1}{\bar{\mu}_2} \ln\left(\frac{2\|x^{(0)}-\bar{x}\|}{\bar{x}_{\min}}\right)\right\rceil$ *steps of iteration.*

Indeed, the conditions in Corollary 7 imply $\|x^{(t)} - \bar{x}\| < \bar{x}_{\min}$ which leads to $\text{supp}(\bar{x}) \subseteq \text{supp}(x^{(t)})$. We note that the parameter estimation error bound derived by Jain et al. (2014, Theorem 3) implies a similar support recovery guarantee as in Corollary 7(a).

3.2 Comparison to Prior Results

Now we compare our method and parameter estimation error bounds to some prior relevant methods and results.

Our method versus nonlinear-IHT (Blumensath, 2013). As we remarked in Section 2 that FGraHTP is identical to the nonlinear-IHT method proposed by Blumensath (2013). The estimation error results of the two, however, are different: the error bound of nonlinear-IHT is relying on the objective value at the target solution; whereas ours in Theorem 2(b) is controlled by the infinity norm of gradient at the target solution.

Our method versus ℓ_1-norm ball constrained estimation (Agarwal et al., 2012). It is worthwhile to compare our ℓ_0-estimation results to those established by Agarwal et al. (2012, Theorem 1) for ℓ_1-norm ball constrained M-estimator (maximum likelihood type estimator). Let us consider $\bar{x}$ as the underlying $\bar{k}$-sparse nominal parameter in a statistical model. When using sparsity level $k = \bar{k}$, the $O(\sqrt{k}\|\nabla f(\bar{x})\|_\infty)$ estimation error bound in Theorem 2, which is at the same order of statistical error, is essentially identical to the error bound derived by Agarwal et al. (2012, Theorem 1). Our analysis, however, requires a bounding assumption on the restricted strong condition number which is not required in their result. This can be interpreted as the price of using nonconvex sparsity constraint rather than its convex relaxation. By using properly relaxed sparsity level $k = O(\bar{k})$, we obtain similar estimation error bounds in Theorem 5 but without assuming bounded restricted strong condition number. In this case, at a slight sacrifice in sparsity level, our methods gain better dependence on restricted strong condition number than those for convex models. Concerning the efficiency of projection steps, the ℓ_0-projection used in FGraHTP is more efficient than the ℓ_1-projection required by those first-order convex minimization methods. The projection operation of GraHTP is more expensive as it requires an additional debiasing step right after ℓ_0-projection.

Our method versus GraSP (Bahmani et al., 2013). A similar estimation error bound as in Theorem 2(a) has been established for the GraSP method (Bahmani et al., 2013). At time instance t, GraSP first conducts debiasing over the union of the top k entries of $x^{(t-1)}$ and the top $2k$ entries of $\nabla f(x^{(t-1)})$, and then preserves the top k entries of the resultant vector, which becomes $x^{(t)}$. Our GraHTP is connected to GraSP in the sense that the k largest absolute elements after the gradient descent step will come from some combination of the largest elements in $x^{(t-1)}$ and the largest elements in the gradient $\nabla f(x^{(t-1)})$.

Although having similar convergence behavior, the per-iteration cost of GraHTP is cheaper than GraSP: at each iteration, GraSP needs to minimize the objective over a support of size at least $2k$ while that size for GraHTP is k. FGraHTP is even cheaper for iteration as it does not need any debiasing operation. We will compare the actual numerical performance of these methods in the experiment section.

Our results versus the results obtained by Jain et al. (2014). The RIP-condition-free estimation error bound in Theorem 5(a) has also been proved by Jain et al. (2014, Theorem 3) with relaxed sparsity levels. As pointed out in Remark 6, the contraction factor $\bar{\mu}_1$ derived in Theorem 5(a) is inferior to the rate $\bar{\mu}_2$ in Theorem 5(b) when the restricted strong condition number is relatively small. Moreover, from Figure 1(b) we can see that $\bar{\mu}_1$ is valued in a quite restrictive interval $(0.87, 1)$ while $\bar{\mu}_2$ can be varied in a much wider range of $(0.65, 1)$. Figure 1(a) shows that the contraction factors μ_1 and μ_2 derived in Theorem 2 can be widely valued in $(0, 1)$. The more favorable contraction factors in Theorem 5(b) and Theorem 2 are resulted from a more careful analysis of GraHTP/FGraHTP and using a tight hard-thresholding bound derived by Shen & Li (2016).

4. Sparsity Recovery Analysis

In this section, we further analyze the sparsity recovery performance of GraHTP. In Corollary 3 and Corollary 7, we have already established some general sparsity recovery results for GraHTP. Here we will provide a refined analysis without assuming bounded restricted strong condition number. Moreover, we will analyze the sparsity recovery behavior of the sparse estimator $x^\star = \arg\min_{\|x\|_0 \le k} f(x)$ which to our knowledge has not been addressed elsewhere in literature. The main results obtained in this section are highlighted in below:

- For GraHTP algorithm, we derive in Theorem 8 an improved RIP-condition-free result for exactly recovering the support of a target $\bar{k}$-sparse vector with $\bar{k} < k$.
- For the global k-sparse minimizer $x^\star$, we provide in Theorem 10 a set of sufficient conditions under which $x^\star$ is able to recover the support of a target sparse vector.

4.1 Sparsity Recovery of GraHTP

In the following theorem, we show that for proper $k > \bar{k}$, GraHTP is able to recover the support of certain target $\bar{k}$-sparse vector without assuming bounded restricted strong condition numbers. A proof of this theorem is given in Appendix C.1.

Theorem 8 *Assume that f is M_{2k}-smooth and m_{2k}-strongly convex. Let $\bar{x}$ be an arbitrary $\bar{k}$-sparse vector satisfying $k \ge \left(1 + \frac{16M_{2k}^2}{m_{2k}^2}\right)\bar{k}$. Set the step-size to be $\eta = \frac{1}{2M_{2k}}$. If $\bar{x}_{\min} > 2.3\sqrt{\frac{f(\bar{x})-f(x^\star)}{m_{2k}}}$, then GraHTP will terminate and output $x^{(t)}$ satisfying $supp(x^{(t)}, \bar{k}) = supp(\bar{x})$ after at most*

$$t = \left\lceil \frac{2kM_{2k}}{m_{2k}} \ln \frac{\triangle^{(0)}}{\triangle^{-\star}} \right\rceil$$

steps of iteration, where $\triangle^{(0)} = f(x^{(0)}) - f(x^\star)$ and

$$\triangle^{-\star} = \min_{\|x\|_0 \le k, supp(x) \ne supp(x^\star), f(x) > f(x^\star)} f(x) - f(x^\star).$$

Results	Target Solution	RIP Condition	x-min Condition
Corollary 3(a)	Arbitrary k-sparse $\bar{x}$	Required	$\bar{x}_{\min} > \mathcal{O}\left(\frac{\sqrt{k}\|\nabla f(\bar{x})\|_\infty}{m_{2k}}\right)$
Corollary 7(a)	$\|\bar{x}\|_0 = \mathcal{O}\left((\frac{m_{2k}}{M_{2k}})^2 k\right)$	Free	$\bar{x}_{\min} > \mathcal{O}\left(\frac{\sqrt{k}\|\nabla f(\bar{x})\|_\infty}{m_{2k}}\right)$
Theorem 8	$\|\bar{x}\|_0 = \mathcal{O}\left((\frac{m_{2k}}{M_{2k}})^2 k\right)$	Free	$\bar{x}_{\min} > \mathcal{O}\left(\sqrt{\frac{f(\bar{x})-f(x^\star)}{m_{2k}}}\right)$

Table 2: Comparison of Theorem 8 against Corollary 3 and Corollary 7.

Remark 9 *The main message conveyed by Theorem 8 is: If $\bar{k} = \mathcal{O}\left(\frac{m_{2k}^2}{M_{2k}^2}k\right)$ and the nonzero elements in $\bar{x}$ are significantly larger than the value $\sqrt{(f(\bar{x}) - f(x^\star))/m_{2k}}$, then GraHTP will output $x^{(t)}$ whose top $\bar{k}$ entries are exactly the supporting set of $\bar{x}$. The implication of this result is that in order to recover certain $\bar{k}$-sparse signals, one may run GraHTP with a properly relaxed sparsity level k until convergence and then preserve the top $\bar{k}$ entries of the k-sparse output as the final estimation.*

In Table 2, we summarize the sparsity recovery results established in Theorem 8, Corollary 3 and Corollary 7. We claim that the x-min condition in Theorem 8 is no stronger than those in Corollary 3 and Corollary 7. Indeed, when $\bar{x} \neq x^\star$, from the restricted strong-convexity of f and the fact $x^\top y \leq \|x\|_\infty \|y\|_1$ we can derive the following inequality:

$$f(\bar{x}) - f(x^\star) \leq \frac{\|\nabla f(\bar{x})\|_\infty^2 \|\bar{x} - x^\star\|_1^2}{2m_{2k}\|\bar{x} - x^\star\|^2}.$$

It can be verified that the factor $\bar{l} = \|\bar{x} - x^\star\|_1^2/\|\bar{x} - x^\star\|^2$ is valued in the interval $[1, k + \bar{k}]$ if $\bar{x} \neq x^\star$. Since $k > \bar{k}$, we then always have $\sqrt{(f(\bar{x}) - f(x^\star))/m_{2k}} \leq \sqrt{k}\|\nabla f(\bar{x})\|_\infty/m_{2k}$. The closer $\bar{l}$ is to 1, the weaker lower bound condition can be imposed on $\bar{x}_{\min}$ in Theorem 8. In the extreme case when $\bar{l} = 1$, the $\bar{x}_{\min}$ condition becomes $\bar{x}_{\min} > \mathcal{O}\left(\|\nabla f(\bar{x})\|_\infty/m_{2k}\right)$ which is not dependent on factor $\sqrt{k}$ and thus is weaker than those in Corollary 3 and Corollary 7.

4.2 Sparsity Recovery of $x^\star$

Given a target solution $\bar{x}$, the following result gives some sufficient conditions under which the sparse estimator $x^\star$ is able to exactly recover the supporting set of $\bar{x}$. A proof of this result is provided in Appendix C.2.

Theorem 10 *Assume that f is M_{2k}-smooth and m_{2k}-strongly convex. Let $\bar{x}$ be an arbitrary $\bar{k}$-sparse vector with $\bar{k} \leq k$. Then $supp(\bar{x}) = supp(x^\star, \bar{k})$ if either of the following two conditions holds:*

(1) $\bar{x}_{\min} > \frac{4.59\sqrt{k}}{m_{2k}}\|\nabla f(\bar{x})\|_\infty$;

(2) $k \geq \left(1 + \frac{4M_{2k}^2}{m_{2k}^2}\right)\bar{k}$ *and* $\bar{x}_{\min} > 2.3\sqrt{\frac{f(\bar{x})-f(x^\star)}{m_{2k}}}$.

Remark 11 *Theorem 10 shows that when using sparsity level $k \geq \bar{k}$, the top $\bar{k}$ entries of the k-sparse global minimizer $x^\star$ is exactly the support of $\bar{x}$ if $\bar{x}_{\min}$ is significantly larger than $\sqrt{k}\|\nabla f(\bar{x})\|_\infty/m_{2k}$. By using a more relaxed sparsity level as in condition (2), the top $\bar{k}$ entries of $x^\star$ is exactly the support of $\bar{x}$ when $\bar{x}_{\min}$ is significantly larger than $\sqrt{(f(\bar{x}) - f(x^\star))/m_{2k}}$. Note that Theorem 10 is valid without imposing bounding assumptions on restricted strong condition number.*

We now compare the support recovery result in Theorem 10 for the ℓ_0-estimator (1) to those known for the following ℓ_1-regularized estimator:

$$\min_{x\in\mathbb{R}^p} f(x) + \lambda\|x\|_1, \tag{4}$$

where $f(x)$ is a convex loss function and λ is the regularization strength parameter. When the loss function is quadratic, a set of sufficient conditions were derived by Wainwright (2009) to guarantee exact sparsity recovery of Lasso-type estimators. For more general loss functions, a unified sparsity recovery analysis was presented in the paper of Li et al. (2015). We summarize in below a comparison between Theorem 10 and those sparsity recovery results for ℓ_1-regularized estimators (Li et al., 2015) with respect to several key conditions:

- **Local structured smoothness/convexity condition:** Theorem 10 only requires first-order local structured smoothness/convexity conditions (i.e., RSC/RSS) while the results obtained by Li et al. (2015, Theorem 5.1, Condition 1) rely on certain second-order and third-order local structured smoothness conditions.
- **Irrepresentablility condition:** Theorem 10 is free of the so called irrepresentablility condition which is typically required to guarantee the sparsistency of ℓ_1-regularized estimators (Li et al., 2015, Theorem 5.1, Condition 3).
- **x-min condition:** Comparing to the x-min condition derived by Li et al. (2015, Theorem 5.1, Condition 4) which is of order $\mathcal{O}(\sqrt{k}\|\nabla f(\bar{x})\|_\infty)$, the x-min condition (1) in Theorem 10 is comparable at the same order while the x-min condition (2) is sharper since $\sqrt{f(\bar{x}) - f(x^\star)/m_{2k}} \leq \sqrt{k}\|\nabla f(\bar{x})\|_\infty/m_{2k}$.

We comment that the above key differences also apply to the comparison between Theorem 8 for GraHTP and the sparsity recovery results for ℓ_1-regularized estimators. In Section 5.1, we will further specify our results to the setting of sparse linear regression and make a comparison against those sparsity recovery results for Lasso-type estimators (Wainwright, 2009).

5. Applications to Sparsity-Constrained M-estimation

We now specify GraHTP and its analysis to the M-estimation problem which is a popular formulation in statistical machine learning. Given a set of n independently drawn data samples $\{x^{(i)}\}_{i=1}^n$, the M-estimation problem is defined as to minimize the following empirical risk function averaged over the samples:

$$f(w) = \frac{1}{n}\sum_{i=1}^n \phi(x^{(i)} \mid w),$$

where ϕ is a loss function and w is a set of adjustable parameters. The sparsity-constrained M-estimation problem is then given by

$$\min_w f(w), \quad \text{subject to } \|w\|_0 \leq k. \tag{5}$$

In the subsections to follow, we will consider three instances of this model: linear regression, logistic regression and Gaussian precision matrix estimation.

5.1 Sparsity-constrained Linear Regression

Given a $\bar{k}$-sparse parameter vector $\bar{w}$, we assume the samples are generated according to the linear model $v^{(i)} = \bar{w}^\top u^{(i)} + \varepsilon^{(i)}$ where $\varepsilon^{(i)}$ are n i.i.d. sub-Gaussian random variables with parameter σ. The sparsity-constrained least squares regression model is then given by

$$\min_w f(w) = \frac{1}{2n}\sum_{i=1}^n \|v^{(i)} - w^\top u^{(i)}\|^2, \quad \text{subject to } \|w\|_0 \leq k. \tag{6}$$

In this case, GraHTP (and FGraHTP) reduces to the conventional HTP (and IHT) of which the parameter estimation performance has been extensively studied in compressed sensing (Foucart, 2011; Blumensath & Davies, 2009). Here we illustrate the sparsity recovery results we established in Section 4 and compare them against those for ℓ_1-estimators. Suppose $u^{(i)}$ are drawn from Gaussian distribution with covariance matrix $\Sigma \succ 0$. Then it holds with high probability that $f(w)$ has RSC constant $m_{2k} \geq \lambda_{\min}(\Sigma) - \mathcal{O}(\bar{k}\log p/n)$ and RSS constant $M_{2k} \leq \lambda_{\max}(\Sigma) + \mathcal{O}(\bar{k}\log p/n)$, and$\|\nabla f(\bar{w})\|_\infty = \mathcal{O}\left(\sigma\sqrt{\log p/n}\right)$. Assume that $k \geq \bar{k}$. We summarize in below the implications of our sparsity recovery results in sparse linear regression:

- Sparsity recovery of GraHTP. Corollary 3 shows that if $\bar{w}_{\min} > \mathcal{O}\left(\frac{\sigma\sqrt{k\log p/n}}{\lambda_{\min}(\Sigma)}\right)$ and $\frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}$ is well upper bounded, then after sufficient iteration GraHTP and FGraHTP with $k = \bar{k}$ will guarantee support recovery $\mathrm{supp}(x^{(t)}) = \mathrm{supp}(\bar{x})$ with high probability. Corollary 7 indicates that when using certain relaxed sparsity level $k = \mathcal{O}\left(\frac{\lambda^2_{\max}(\Sigma)}{\lambda^2_{\min}(\Sigma)}\bar{k}\right)$, GraHTP and FGraHTP are able to guarantee $\mathrm{supp}(x^{(t)}) \supseteq \mathrm{supp}(\bar{x})$ without assuming bounded condition number. Since $f(\bar{x}) - f(x^\star) \leq \frac{\bar{l}\|\nabla f(\bar{x})\|^2_\infty}{2m_{2k}}$ where $\bar{l} = \|\bar{x}-x^\star\|_1^2/\|\bar{x}-x^\star\|^2 \in [1, k+\bar{k}]$, Theorem 8 implies that if $\bar{w}_{\min} > \mathcal{O}\left(\frac{\sigma\sqrt{\bar{l}\log p/n}}{\lambda_{\min}(\Sigma)}\right)$ and $k = \mathcal{O}\left(\frac{\lambda^2_{\max}(\Sigma)}{\lambda^2_{\min}(\Sigma)}\bar{k}\right)$, then after finite iteration GraHTP will guarantee $\mathrm{supp}(x^{(t)}, \bar{k}) = \mathrm{supp}(\bar{x})$ with high probability.
- Sparsity recovery of the least squares estimator (6). Let $w^\star$ be the global k-sparse minimizer of (6). Theorem 10 shows that $\mathrm{supp}(w^\star, \bar{k}) = \mathrm{supp}(\bar{w})$ holds with high probability if $\bar{w}_{\min} > \mathcal{O}\left(\frac{\sigma\sqrt{k\log p/n}}{\lambda_{\min}(\Sigma)}\right)$. To compare our sparsity recovery results for ℓ_0-estimators against those established by Wainwright (2009, Theorem 1) for Lasso-type estimators, the signal-noise-ratio condition of $\bar{w}_{\min} > \mathcal{O}\left(\frac{\sigma\sqrt{k\log p/n}}{\lambda_{\min}(\Sigma)}\right)$ is shared

in that paper. The key difference is that our analysis is valid without imposing the irrepresentablility condition on design matrix which is required in the sparsity recovery analysis of Lasso-type estimators.

5.2 Sparsity-constrained Logistic Regression

Logistic regression is one of the most popular models in statistical machine learning (Bishop, 2006). In this model the relation between the random feature vector $u \in \mathbb{R}^p$ and its associated random binary label $v \in \{-1, +1\}$ is determined by the conditional probability

$$\mathbb{P}(v|u; \bar{w}) = \frac{\exp(2v\bar{w}^\top u)}{1 + \exp(2v\bar{w}^\top u)}, \tag{7}$$

where $\bar{w} \in \mathbb{R}^p$ denotes parameter vector. Given a set of n independently drawn data samples $\{(u^{(i)}, v^{(i)})\}_{i=1}^n$, logistic regression learns the parameters so as to minimize the following logistic loss function:

$$l(w) := -\frac{1}{n} \log \prod_i \mathbb{P}(u^{(i)} \mid v^{(i)}; w) = \frac{1}{n} \sum_{i=1}^n \log(1 + \exp(-2v^{(i)} w^\top u^{(i)})),$$

which is known to be convex. Unfortunately, in high-dimensional setting, i.e., $n < p$, the problem can be underdetermined and thus its minimum is not unique. A conventional way to handle this issue is to impose ℓ_2-regularization to the logistic loss to avoid singularity. The ℓ_2-penalty, however, does not promote sparse solutions which are often desirable in high-dimensional learning tasks. The sparsity-constrained ℓ_2-regularized logistic regression is then given by

$$\min_w f(w) = l(w) + \frac{\lambda}{2}\|w\|^2, \quad \text{subject to } \|w\|_0 \leq k, \tag{8}$$

where $\lambda > 0$ is the regularization strength parameter. Obviously $f(w)$ is λ-strongly convex. The cardinality constraint enforces the solution to be sparse.

Verifying restricted smoothness and strong convexity. Let $U = [u^{(1)}, ..., u^{(n)}] \in \mathbb{R}^{p\times n}$ be the design matrix and $\sigma(z) = 1/(1+\exp(-z))$ be the sigmoid function. In the case of ℓ_2-regularized logistic loss considered in this section we have $\nabla f(w) = Ua(w)/n + \lambda w$ in which the vector $a(w) \in \mathbb{R}^n$ is given by $[a(w)]_i = -2v^{(i)}(1-\sigma(2v^{(i)}w^\top u^{(i)}))$, and the Hessian $\nabla^2 f(w) = U\Lambda(w)U^\top/n + \lambda I$ where $\Lambda(w)$ is an $n \times n$ diagonal matrix whose diagonal entries $[\Lambda(w)]_{ii} = 4\sigma(2v_i w^\top u_i)(1 - \sigma(2v_i w^\top u_i))$. Given an integer s, recall that $\lambda_{\max}(A, s)$ denotes the largest s-sparse eigenvalue of a positive semi-definite matrix A and $\lambda_{\min}(A, s)$ denotes the smallest s-sparse eigenvalue of A. Assume that the algorithm is initialized with all-zero vector. Then it can be verified that $f(w)$ is $(\lambda_{\max}(UU^\top, s)+\lambda)$-smooth and $(\gamma_s+\lambda)$-strongly convex where $\gamma_s := \min_{f(w)\leq f(0)} \lambda_{\min}(U\Lambda(w)U^\top, s)$.

Bounding the value of $\|\nabla f(\bar{w})\|_\infty$. We now bound the infinity norm $\|\nabla f(\bar{w})\|_\infty$ which controls the estimation error and sparisty recovery bounds of GraHTP/FGraHTP. In the following derivation, we assume that the joint density of the random vector $(u, v) \in \mathbb{R}^{p+1}$ is given by the following exponential family distribution:

$$\mathbb{P}(u, v; \bar{w}) = \exp\left(v\bar{w}^\top u + B(u) - A(\bar{w})\right), \tag{9}$$

where

$$A(\bar{w}) := \log \sum_{v=\{-1,1\}} \int_{\mathbb{R}^p} \exp\left(v\bar{w}^\top u + B(u)\right) du$$

is the log-partition function. The term $B(u)$ characterizes the marginal behavior of u. Obviously, the conditional distribution of v given u, $\mathbb{P}(v \mid u; \bar{w})$, is given by the Bernoulli distribution in (7). By doing some elementary manipulations (see, e.g., Wainwright & Jordan, 2008) we can obtain the following standard result which shows that the first derivative of the logistic log-likelihood $l(w)$ yields the cumulants of the random variables $v[u]_j$:

$$\frac{\partial l}{\partial [w]_j} = \frac{1}{n}\sum_{i=1}^{n}\left\{-v^{(i)}[u^{(i)}]_j + \mathbb{E}_v[v[u^{(i)}]_j \mid u^{(i)}]\right\}. \tag{10}$$

Here the expectation $\mathbb{E}_v[\cdot \mid u]$ is taken over the conditional distribution (7). We introduce the following sub-Gaussian condition on the random variate $v[u]_j$.

Assumption 1 *For all j, we assume that there exists constant $\sigma > 0$ such that for all ζ,*

$$\mathbb{E}[\exp(\zeta v[u]_j)] \leq \exp\left(\sigma^2\zeta^2/2\right).$$

This assumption holds when $[u]_j$ are sub-Gaussian (e.g., Gaussian or bounded) random variables. The following result establishes the bound of $\|\nabla f(\bar{w})\|_\infty$.

Proposition 12 *If Assumption 1 holds, then with probability at least $1 - 4p^{-1}$,*

$$\|\nabla f(\bar{w})\|_\infty \leq 4\sigma\sqrt{\ln p/n} + \lambda\|\bar{w}\|_\infty.$$

A proof of this result is provided in Appendix D.1. If we choose $\lambda = O(\sqrt{\ln p/n})$, then with overwhelming probability $\|\nabla f(\bar{w})\|_\infty$ vanishes at the rate of $O(\sqrt{\ln p/n})$. This bound is superior to the bound obtained by Bahmani et al. (2013, Section 4.2) which is not vanishing as sample size increases. Based on the above discussion, we can similarly specify our parameter estimation and sparsity recovery results to sparse logistic regression. Here we omit the detailed specification of results for the sake of redundance reducing.

5.3 Sparsity-constrained Gaussian Precision Matrix Estimation

As an important class of sparse learning problems for exploring the interrelationship among a large number of random variables, the sparse Gaussian precision (inverse covariance) matrix estimation problem has received significant interest in a variety of scientific and engineering domains, including computational biology, natural language processing and document analysis.

Let x be a p-variate random vector with zero-mean Gaussian distribution $\mathcal{N}(0, \bar{\Sigma})$. Its density is parameterized by the precision matrix $\bar{\Omega} = \bar{\Sigma}^{-1} \succ 0$ as

$$\phi(x; \bar{\Omega}) = \frac{1}{\sqrt{(2\pi)^p (\det\bar{\Omega})^{-1}}} \exp\left(-\frac{1}{2}x^\top\bar{\Omega}x\right).$$

It is well known that the conditional independence between the variables $[x]_i$ and $[x]_j$ given $\{[x]_k, k \neq i, j\}$ is equivalent to $[\bar{\Omega}]_{ij} = 0$. The conditional independence relations between

components of x, on the other hand, can be represented by a graph $\mathcal{G} = (V, E)$ in which the vertex set V has p elements corresponding to $[x]_1, ..., [x]_p$, and the edge set E consists of edges between node pairs $\{[x]_i, [x]_j\}$. The edge between $[x]_i$ and $[x]_j$ is excluded from E if and only if $[x]_i$ and $[x]_j$ are conditionally independent given other variables. This graphical model is known as Gaussian Markov random field (GMRF) (Edwards, 2000). Thus for multivariate Gaussian distribution, estimating the support of the precision matrix $\bar{\Omega}$ is equivalent to learning the structure of GMRF $\mathcal{G}$.

Given i.i.d. samples $\mathbb{X}_n = \{x^{(i)}\}_{i=1}^n$ drawn from $\mathcal{N}(0, \bar{\Sigma})$, the negative log-likelihood, up to a constant, can be written in terms of the precision matrix as

$$\mathcal{L}(\mathbb{X}_n; \bar{\Omega}) := -\log\det\bar{\Omega} + \langle\Sigma_n, \bar{\Omega}\rangle,$$

where Σ_n is the sample covariance matrix. We are interested in the problem of estimating a sparse precision $\bar{\Omega}$ with no more than a pre-specified number of off-diagonal nonzero entries. For this purpose, we consider the following cardinality constrained log-determinant program:

$$\min_{\Omega\succ 0} L(\Omega) := -\log\det\Omega + \langle\Sigma_n, \Omega\rangle, \quad \text{s.t. } \|\Omega^-\|_0 \leq 2k, \tag{11}$$

where Ω^- is the restriction of Ω on the off-diagonal entries, $\|\Omega^-\|_0 = |\text{supp}(\Omega^-)|$ is the cardinality of the supporting set of Ω^- and the integer $k > 0$ controls the number of edges, i.e., $|E|$, in the graph.

Verifying restricted smoothness and strong convexity. It can be verified that the Hessian matrix of $L(\Omega)$ is given by $\nabla^2 L(\Omega) = \Omega^{-1} \otimes \Omega^{-1}$, where $\otimes$ denotes the Kronecker product operator. Suppose that $\|\Omega^-\|_0 \leq s$ and $\alpha_s I \preceq \Omega \preceq \beta_s I$ for some $0 < \alpha_s \leq \beta_s$. Due to the fact that the eigenvalues of Kronecker products of symmetric matrices are the products of the eigenvalues of their factors, it holds that $\beta_s^{-2} I \preceq \Omega^{-1} \otimes \Omega^{-1} \preceq \alpha_s^{-2} I$. Therefore we have $\beta_s^{-2} \leq \|\nabla^2 L(\Omega)\| \leq \alpha_s^{-2}$ which implies that $L(\Omega)$ is β_s^{-2}-strongly convex and α_s^{-2}-smooth. Inspired by this property, we consider applying GraHTP to the following variant of problem (11):

$$\min_{\alpha I\preceq\Omega\preceq\beta I} L(\Omega), \quad \text{s.t. } \|\Omega^-\|_0 \leq 2k, \tag{12}$$

where $0 < \alpha \leq \beta$ are two constants which respectively lower and upper bound the eigenvalues of the desired solution. To roughly estimate α and β, we employ a rule proposed by Lu (2009, Proposition 3.1) for the ℓ_1-regularized log-determinant program. Specifically, we set

$$\alpha = (\|\Sigma_n\|_2 + n\xi)^{-1}, \quad \beta = \xi^{-1}(n - \alpha\text{Tr}(\Sigma_n)),$$

where ξ is a small enough positive number (e.g., $\xi = 10^{-2}$ as used in our implementation).

Bounding the value of $|\nabla L(\bar{\Omega})|_\infty$. It is standard to know that $|\nabla L(\bar{\Omega})|_\infty = |\Sigma_n - \bar{\Sigma}|_\infty = \mathcal{O}(\sqrt{\log p/n})$ with probability at least $1 - c_0 p^{-c_1}$ for some positive constants c_0 and c_1 and sufficiently large n (see, e.g., Ravikumar et al., 2011, Lemma 1). Therefore, with overwhelming probability we have $|\nabla L(\bar{\Omega})|_\infty = \mathcal{O}(\sqrt{\log p/n})$ when n is sufficiently large.

A Modified GraHTP. Note that GraHTP is not directly applicable to the problem (12) due to the presence of the constraint $\alpha I \preceq \Omega \preceq \beta I$ in addition to the sparsity

constraint. To address this issue, we need to accordingly modify the debiasing step (**S3**) of GraHTP to minimize $L(\Omega)$ over the constraints $\alpha I \preceq \Omega \preceq \beta I$ and $\mathrm{supp}(\Omega) \subseteq F^{(t)}$:

$$\min_{\alpha I \preceq \Omega \preceq \beta I} L(\Omega), \quad \text{s.t. } \mathrm{supp}(\Omega) \subseteq F^{(t)}. \tag{13}$$

Since this problem is convex, any off-the-shelf convex solver can be applied for optimization. In our implementation, we resort to the alternating direction method of multipliers (ADMM) (Boyd et al., 2010; Yuan, 2012) which has been observed to be efficient in our numerical practice. The implementation details of ADMM for solving the subproblem (13) are deferred to Appendix D.2. The modified GraHTP for sparse Gaussian precision matrix estimation is outlined in Algorithm 2.

Algorithm 2: A Modified GraHTP for Sparse Gaussian Precision Matrix Estimation.

Initialization: $\Omega^{(0)}$ *with* $\|(\Omega^{(0)})^-\|_0 \leq 2k$ *and* $\alpha I \preceq \Omega^{(0)} \preceq \beta I$ *(typically* $\Omega^{(0)} = \alpha I$*),* $t = 1$.
Output: $\Omega^{(t)}$.
repeat
  (**S1**) Compute $\tilde{\Omega}^{(t)} = \Omega^{(t-1)} - \eta \nabla L(\Omega^{(t-1)})$;
  (**S2**) Let $\tilde{F}^{(t)} = \mathrm{supp}((\tilde{\Omega}^{(t)})^-, 2k)$ be the indices of $(\tilde{\Omega}^{(t)})^-$ with the largest $2k$ absolute values and $F^{(t)} = \tilde{F}^{(t)} \cup \{(1,1), ..., (p,p)\}$;
  (**S3**) Compute $\Omega^{(t)} = \arg\min \left\{ L(\Omega); \alpha I \preceq \Omega \preceq \beta I, \mathrm{supp}(\Omega) \subseteq F^{(t)} \right\}$;
  $t = t + 1$;
until *halting condition holds*;

6. Experimental Results

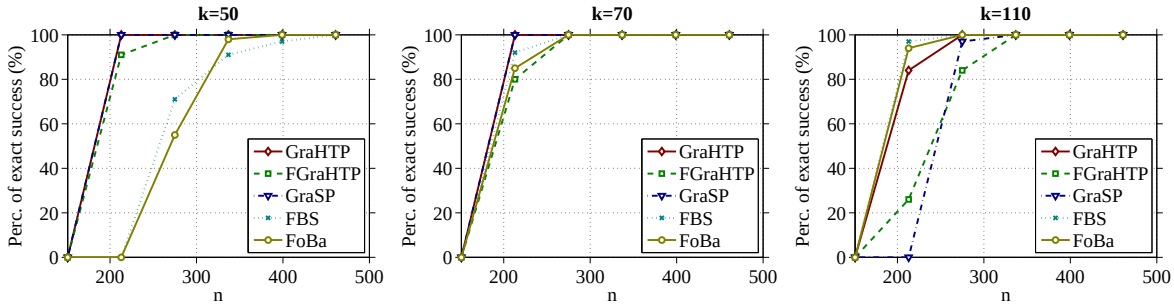
This section is devoted to illustrating the empirical performance of GraHTP/FGraHTP when applied to sparse learning tasks. Our algorithms are implemented in Matlab 7.12 running on a desktop with Intel Core i7 3.2G CPU and 16G RAM.

6.1 Sparsity-constrained Linear Regression

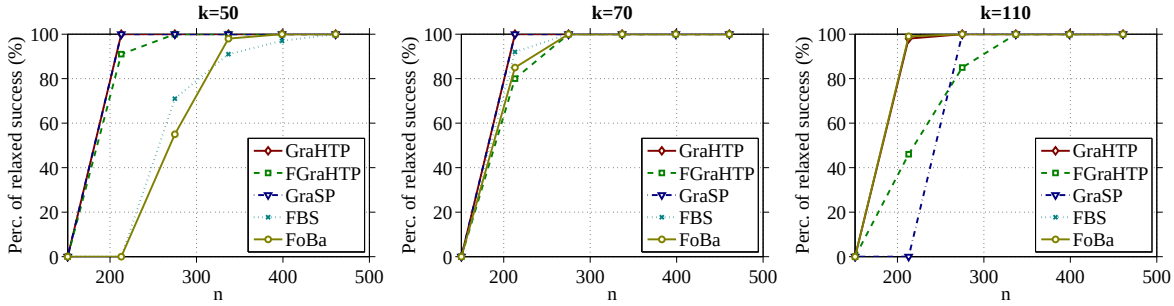
We conduct a group of Monte-Carlo simulation experiments on sparse linear regression model to verify the sparsity recovery results presented in Section 4.

Data generation. We consider a synthetic data model in which the sparse parameter $\bar{w}$ is a $p = 500$ dimensional vector that has $\bar{k} = 50$ nonzero entries drawn independently from a Gaussian distribution with significant mean. Each data sample u is a normally distributed dense vector. The responses are generated by $v = \bar{w}^\top u + \varepsilon$ where ε is a standard Gaussian noise. We allow the sample size n to be varying and for each n, we generate 100 random copies of data independently.

Baselines and evaluation metric. We test GraHTP and FGraHTP with varying sparsity level $k \geq \bar{k}$ and compare their performance with three state-of-the-art greedy selection methods: GraSP (Bahmani et al., 2013), FBS (Yuan & Yan, 2013) and FoBa (Zhang, 2008). As we have mentioned, GraSP is also a hard-thresholding-type method. This method



(a) Exact support recovery



(b) Relaxed support recovery

Figure 2: Sparse linear regression on simulated data: chance of success curves for support recovery under varying sample size and sparsity level.

simultaneously selects at each iteration k nonzero entries and update their values via exploring the top k entries in the previous iterate as well as the top $2k$ entries in the previous gradient. FBS is a forward-selection-type method which iteratively selects an atom from the dictionary and minimizes the objective function over the linear combinations of all the selected atoms. FoBa is an adaptive forward-backward greedy selection algorithm which allows elimination of selected variables when the objective value does not increase significantly. We use two metrics to measure the support recovery performance. We say a *relaxed support recovery* is successful if $\mathrm{supp}(\bar{w}) \subseteq \mathrm{supp}(w^{(t)})$ and an *exact support recovery* is successful if $\mathrm{supp}(\bar{w}) = \mathrm{supp}(w^{(t)}, \bar{k})$. We replicate the experiment over the 100 trials and record the percentage of relaxed success and percentage of exact success for each configuration of the pair (n, k).

Results. Figure 2 shows the percentage of exact (relaxed) success curves as functions of sample size n, under different sparsity levels $k \in \{50, 70, 110\}$. From these curves we can make the following observations:

- For each curve, the chance of success increases as sample size n increases. This is as expected because the larger sample size is, the easier the x-min conditions can be fulfilled so as to guarantee exact support recovery;
- GraHTP is superior to FGraHTP for sparsity recovery, especially when using sparsity level $k > \bar{k}$ and relatively small sample size. This indicates that the debiasing step

conducted in GraHPT can significantly improve the accuracy of sparsity recovery, especially in noisy settings.

- The left panel of Figure 2 shows that when $k = \bar{k}$, GraHTP/FGraHTP and GraSP are comparable and they all significantly outperform FBS and FoBa, especially when the sample size is relatively small. This observation suggests that hard-thresholding-type methods are more accurate than forward and/or backward selection methods for sparsity recovery with exact sparsity level. The middle panel shows that for slightly increased sparsity level $k = 70$, GraHTP and GraSP still exhibit superior performance, while the performance gap among all the considered algorithms decreases. From the right panel we can see that for relatively large $k > \bar{k}$, FBS, Foba and GraHTP have much better performance than FGraHTP and GraSP.

From the above observations we conclude that GraHTP is able to achieve better trade-off between accuracy and stability of sparsity recovery than the other considered methods.

6.2 Sparsity-constrained Logistic Regression

We present in this subsection the experimental results on several synthetic and real-data sparse logistic regression tasks.

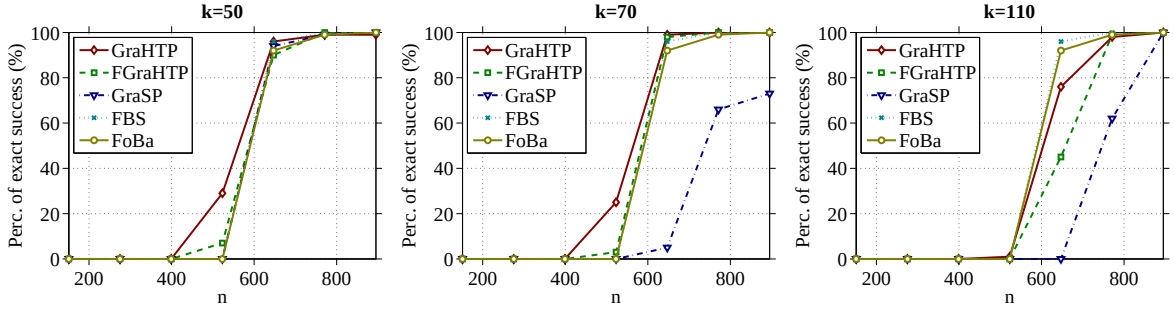
6.2.1 Monte-Carlo simulation

In this group of Monte-Carlo experiments, we use a simulated data to verify the sparsity recovery performance of GraHTP and FGraHTP on logistic regression model. The sparse parameter and design matrix are generated in an identical way to that of the linear regression model. The data labels, $v \in \{-1, 1\}$, are generated randomly according to the Bernoulli distribution $\mathbb{P}(v = 1|u; \bar{w}) = \exp(2\bar{w}^\top u)/(1 + \exp(2\bar{w}^\top u))$. The same experiment protocol as used in the previous linear regression setting applies here. Inspired by Theorem 8 and the discussion in Section 5.2, we set the step-size $\eta = \frac{1}{2M_{2k}}$ where $M_{2k} = \lambda_{\max}(UU^\top, 2k) + \lambda$. The sparse eigenvalue $\lambda_{\max}(UU^\top, 2k)$ can be computed using the truncated power method (Yuan & Zhang, 2013).

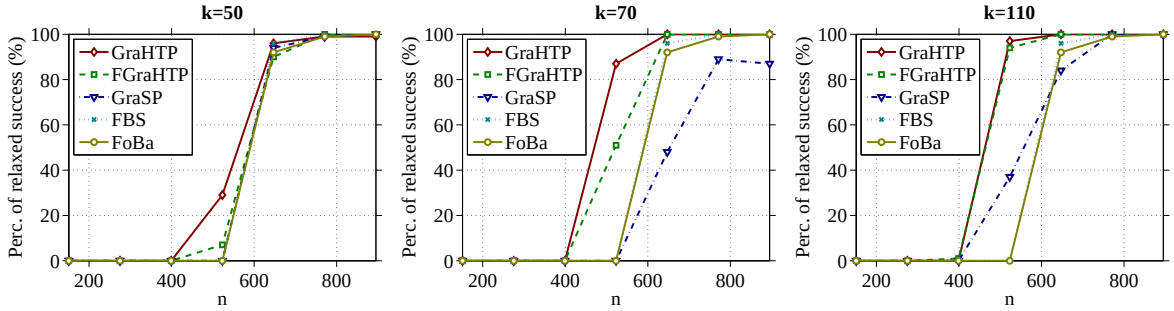
Results. For different sparsity levels $k \geq \bar{k}$, Figure 3 shows the chance of exact (relaxed) success curves as functions of sample size n. Again, from these curves we can observe that: 1) in a wide range of sparsity level, GraHP achieves better trade-off between accuracy and stability than the other considered sparsity recovery methods; and 2) GraHTP consistently outperforms FGraHTP in noisy settings when using $k > \bar{k}$.

6.2.2 Real data experiments

We further illustrate the performance of GraHTP/FGraHTP on real data for binary logistic regression. The data used for evaluation include two *dense* data sets gisette (Guyon et al., 2005) and breast cancer (Hess et al., 2006), and two *sparse* data sets rcv1.binary (Lewis et al., 2005) and news20.binary (Keerthi & DeCoste, 2005). Table 3 summaries the statistics of these data sets. For each data set, we test with sparsity parameters $k \in \{100, 200, ..., 1000\}$ and fix the regularization parameter $\lambda = 10^{-5}$. We initialize $w^{(0)} = 0$ and set the stopping criterion as $\|w^{(t)} - w^{(t-1)}\|/\|w^{(t-1)}\| \leq 10^{-4}$.



(a) Exact support recovery



(b) Relaxed support recovery

Figure 3: Sparse logistic regression on simulated data: chance of success curves for support recovery under varying sample size and sparsity level.

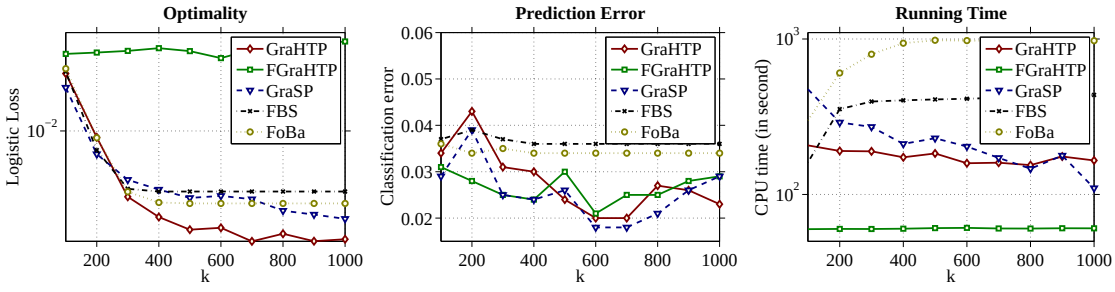
Datasets	Training Size	Testing Size	Dimensionality
gisette	6,000	1,000	5,000
breast cancer	54	79	22,283
rcv1.binary	20,242	20,000	47,236
news20.binary	10,000	9,996	1,355,191

Table 3: Statistics of data sets used in binary logistic regression experiment.

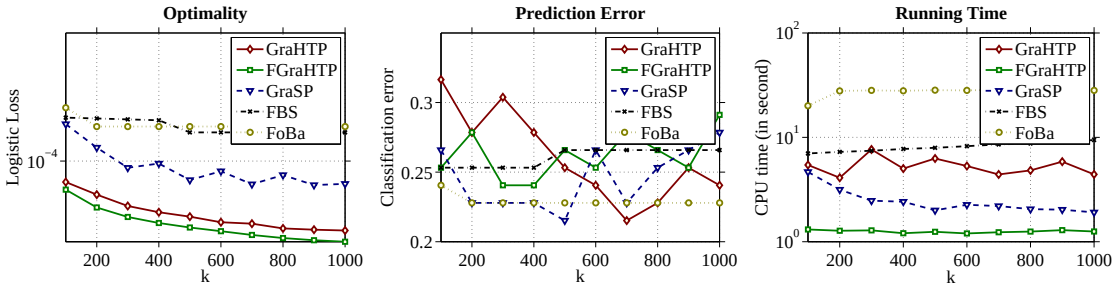
Results. The objective value, test classification error and CPU running time curves under varying sparsity level k are plot in Figure 4. From these curves we have the following observations:

- On optimality: GraHTP is superior to the other considered algorithms in most cases. FGraHTP is less optimal on gisette data, while it is comparable to the other algorithms on the other three data sets.

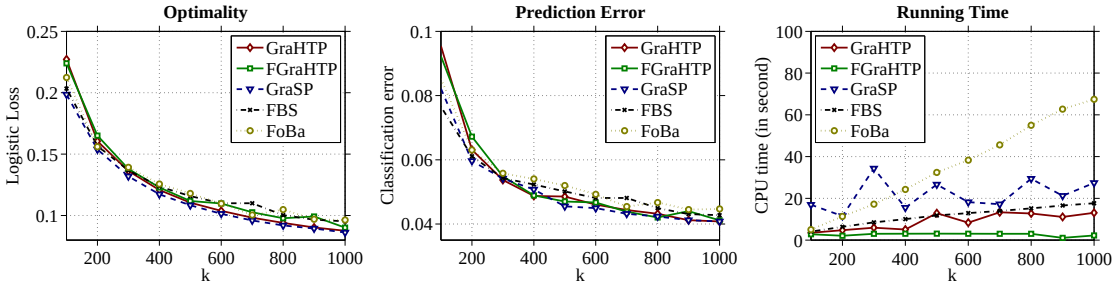
- On classification accuracy: GraHTP and GraSP are comparable to each other and they are slightly superior to the other algorithms in most cases; FGraHTP is average in classification accuracy in most cases.



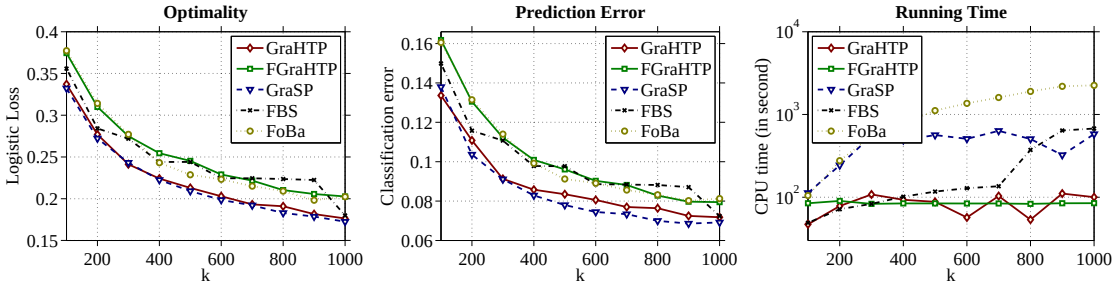
(a) gisette



(b) breast cancer



(c) rcv1.binary



(d) news20.binary

Figure 4: Sparse logistic regression on real data: objective value, classification error and CPU running time curves under varying sparsity level.

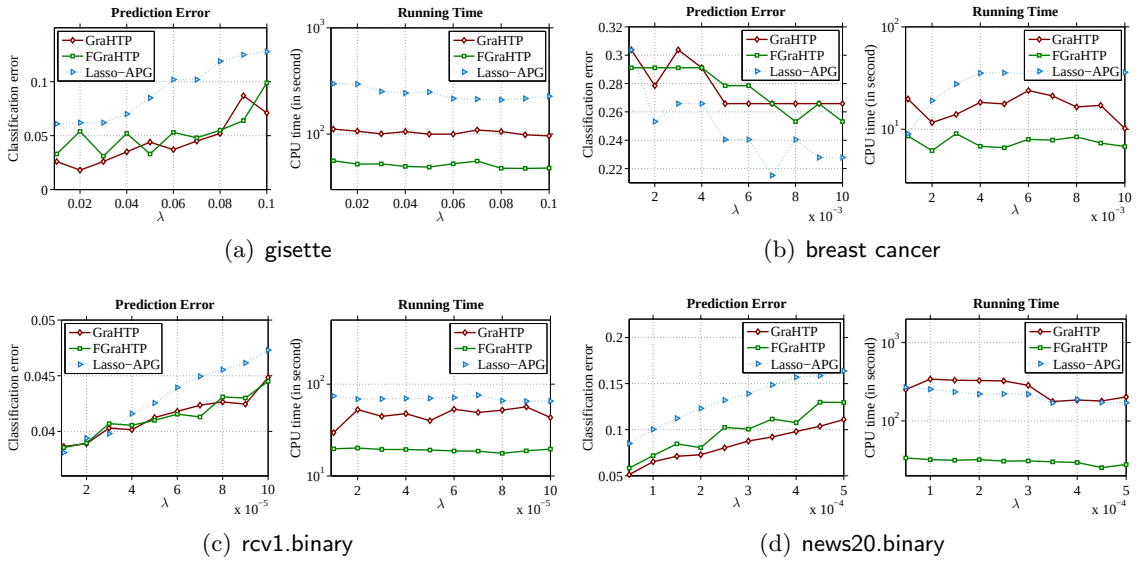


Figure 5: Sparse logistic regression on real data: comparison between GraHTP/FGraHTP and Lasso-type estimator in classification error and CPU running time.

- On execution time: FGraHTP is the most efficient one and GraHTP is the runner-up except on breast cancer. Particularly, as shown in Figure 4(d) that the computational advantage of FGraHTP/GraHTP over the other considered methods becomes significant on news20.binary which is relatively large in scale.

To summarize, GraHTP and FGraHTP are able to achieve desirable trade-off between accuracy and efficiency on the considered data sets.

Comparison against Lasso-type estimator. We have also conducted a set of experiments to compare GraHTP/FGraHTP against the Lasso-type estimator (4) for ℓ_1-regularized sparse learning. To make a fair comparison, we first solve the Lasso-type estimator (4) using an accelerated proximal gradient method (Beck & Teboulle, 2009), which we call Lasso-APG, and then run GraHTP with the sparsity level of the Lasso-APG solution. Figure 5 shows the test classification error and CPU running time curves under varying regularization parameter λ. We can observe from this group of results that: (1) GraHTP and FGraHTP outperform Lasso-APG in classification accuracy on three out of the four data sets in use; and (2) FGraHTP is the most efficient one on all the data sets and GraHTP is faster than Lasso-APG on three of the data sets. Based on these observations, we can conclude that GraHTP and FGraHTP tend to be more accurate and efficient than Lasso-type estimator when their output solutions are at the same sparsity level.

6.3 Sparsity-constrained Gaussian Precision Matrix Estimation

We further assess the performance of GraHTP/FGraHTP when applied to sparse precision matrix estimation.

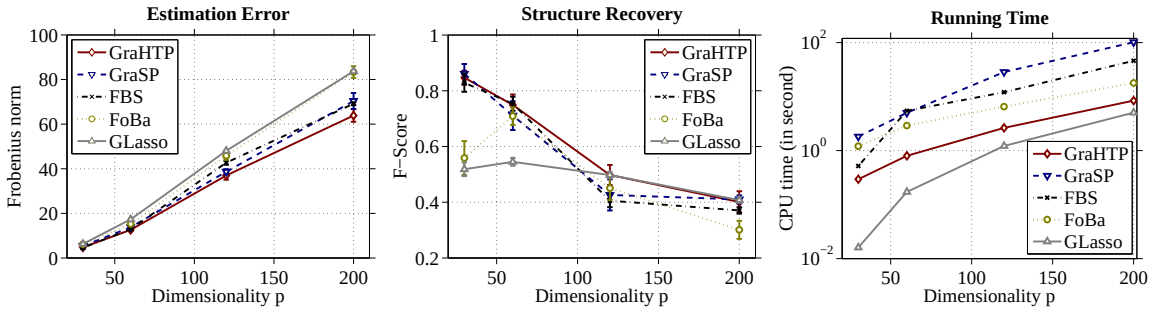


Figure 6: Sparse precision matrix estimation on simulated data: Matrix Frobenius norm loss, support recovery F-score and CPU running time curves under varying data dimensionality. The larger the F-score, the better the support recovery performance.

6.3.1 MONTE-CARLO SIMULATION

Our simulation study employs the sparse precision matrix model $\bar{\Omega} = \Theta + \sigma I$ where each off-diagonal entry in Θ is generated independently and equals 1 with probability $P = 0.1$ or 0 with probability $1 - P = 0.9$. Θ has zeros on the diagonal, and σ is chosen so that the condition number of $\bar{\Omega}$ is p. Let $\bar{\Sigma} = \bar{\Omega}^{-1}$ be the covariance matrix. We generate a training sample of size $n = 100$ from $\mathcal{N}(0, \bar{\Sigma})$, and an independent sample of size 100 from the same distribution for tuning the parameter k. The numerical performance is evaluated with different values of $p \in \{30, 60, 120, 200\}$, replicated 100 times each.

We compare the modified GraHTP (as outlined in Algorithm 2) with GraSP, FBS and FoBa. To adopt GraSP to sparse precision matrix estimation, we modify the algorithm with a similar two-stage strategy as used in the modified GraHTP such that it can handle the eigenvalue bounding constraint in addition to the sparsity constraint. FBS and FoBa have already been applied to sparse precision matrix estimation problems in literature (Yuan & Yan, 2013; Jalali et al., 2011). Also, we compare GraHTP with Graphical Lasso (GLasso) which is one of the representative Lasso-type convex estimators for ℓ_1-penalized log-determinant program (Friedman et al., 2008). The quality of precision matrix estimation is measured by its distance to the truth in Frobenius norm and the support recovery F-score. The larger the F-score, the better the support recovery performance.

Figure 6 compares the matrix error in Frobenius norm, support recovery F-score and CPU running time achieved by each of the considered algorithms for different p. The results show that GraHTP performs favorably in terms of estimation error and support recovery accuracy. We note from the error bars in the curves that the standard error (in 100 replication) of GraHTP is relatively larger than Glasso. This is because GraHTP approximately solves a nonconvex problem via greedy selection at each iteration; the procedure is less stable than those convex solvers such as GLasso. Similar phenomenon of instability has also been observed for the other considered ℓ_0-estimators. The right panel of Figure 6 shows the computational time of the considered algorithms. We can see that GLasso is more efficient than the four greedy selection methods. Although inferior to GLasso, GraHTP is still computationally more attractive than the other considered greedy selection solvers.

Methods	Specificity	Sensitivity	MCC	CPU Time (sec.)
GraHTP	0.77 (0.11)	0.77 (0.19)	**0.49** (0.19)	1.92
GraSP	0.73 (0.10)	**0.78** (0.18)	0.45 (0.17)	4.06
FBS	0.78 (0.11)	0.74 (0.18)	0.48 (0.19)	8.73
FoBa	0.72 (0.11)	**0.78** (0.18)	0.44 (0.18)	6.73
GLasso	**0.81** (0.11)	0.64 (0.21)	0.45 (0.19)	**1.19**

Table 4: Sparse precision matrix estimation on breast cancer data: comparison of average (std) classification accuracy and average CPU running time over 100 replications.

6.3.2 Real Data

We consider the task of LDA (linear discriminant analysis) classification of tumors using the breast cancer data set. This data consists of 133 subjects, each of which is associated with 22,283 gene expression levels. Among these subjects, 34 are with pathological complete response (pCR) and 99 are with residual disease (RD). The pCR subjects are considered to have a high chance of cancer free survival in the long term. Based on the estimated precision matrix of the gene expression levels, we apply LDA to predict whether a subject can achieve the pCR state or the RD state.

Experiment protocol. In this experiment, we follow the same protocol as what was used in the paper of Cai et al. (2011). The data are randomly divided into the training and test sets. In each random division, 5 pCR subjects and 16 RD subjects are randomly selected to constitute the test data, and the remaining subjects form the training set with size $n = 112$. By using two-sample t test, $p = 113$ most significant genes are selected as covariates. Following the LDA framework, we assume that the normalized gene expression data are normally distributed as $\mathcal{N}(\mu_l, \bar{\Sigma})$, where the two classes are assumed to have the same covariance matrix, $\bar{\Sigma}$, but different means, μ_l, $l = 1$ for pCR state and $l = 2$ for RD state. Given a test data sample x, we calculate its LDA scores, $\delta_l(x) = x^\top \hat{\Omega}\hat{\mu}_l - \frac{1}{2}\hat{\mu}_l^\top \hat{\Omega}\hat{\mu}_l + \log \hat{\pi}_l$, $l = 1, 2$, using the precision matrix $\hat{\Omega}$ estimated by the considered methods. Here $\hat{\mu}_l = (1/n_l)\sum_{i \in \text{class}_l} x_i$ is the within-class mean in the training set and $\hat{\pi}_l = n_l/n$ is the proportion of class l subjects in the training set. The classification rule is $\hat{l}(x) = \arg\max_{l=1,2} \delta_l(x)$. Clearly, the classification performance is directly affected by the estimation quality of $\hat{\Omega}$. Hence, we assess the precision matrix estimation performance on the test data and compare GraHTP with GraSP, FBS, FoBa and GLasso. We use a 6-fold cross-validation on the training data for tuning the sparsity level parameter in ℓ_0-estimators and the regularization strength parameter in GLasso. We replicate the experiment 100 times.

Evaluation metric and results. To evaluate classification performance, we use the following defined specificity, sensitivity (or recall), and Mathews correlation coefficient

(MCC) criteria as used by Cai et al. (2011):

$$\text{Specificity} = \frac{\text{TN}}{\text{TN}+\text{FP}}, \qquad \text{Sensitivity} = \frac{\text{TP}}{\text{TP}+\text{FN}},$$
$$\text{MCC} = \frac{\text{TP}\times\text{TN}-\text{FP}\times\text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}},$$

where TP and TN stand for true positives (pCR) and true negatives (RD), respectively, and FP and FN stand for false positives/negatives, respectively. The larger the criterion value, the better the classification performance. Since one can adjust decision threshold in any specific algorithm to trade-off specificity and sensitivity (increase one while reduce the other), the MCC is more meaningful as a single performance metric. Table 4 lists the averages and standard deviations, in the parentheses, of the three classification criteria over 100 replications. It can be observed that GraHTP is quite competitive to the leading methods in all the three metrics. The average CPU running time of each considered method is listed in the rightmost column of Table 4.

7. Conclusion

In this article, we proposed GraHTP as a generalization of HTP from compressed sensing to the generic problem of sparsity-constrained loss minimization. The main idea is to force the gradient descent iteration to be sparse via hard thresholding. Theoretically, we proved that under mild conditions, GraHTP converges geometrically and its estimation error is controlled by the restricted norm of gradient at the target sparse solution. Under properly strengthened conditions, we further established the sparsity recovery performance of GraHTP which to our knowledge has not been systematically analyzed elsewhere in literature. Also, we have proposed and analyzed the FGraHTP algorithm as a fast variant of GraHTP without applying the debiasing operation after truncation. Empirically, we showed that GraHTP and FGraHTP are superior or competitive to the state-of-the-art greedy pursuit methods when applied to sparse learning problems including linear regression, logistic regression and precision matrix estimation. To conclude, simply combining gradient descent with hard thresholding leads to an accurate and computationally tractable procedure for solving sparsity-constrained loss minimization problems.

Acknowledgments

The authors would like to thank the anonymous referees for their constructive comments which are extremely helpful for improving this work. Xiao-Tong Yuan and Ping Li were partially supported by NSF-Bigdata-1419210, NSF-III-1360971, ONR-N00014-13-1-0764, and AFOSR-FA9550-13-1-0137. Xiao-Tong Yuan is also partially supported by NSFC-61522308 and Tencent AI Lab Rhino-Bird Joint Research Program (No.JR201801). Tong Zhang was supported by NSF-IIS-1407939 and NSF-IIS-1250985.

Appendix A. Technical Lemmas

We present in this appendix section a few technical lemmas to be used in the proofs of main results.

Lemma 13 *Let x be a k-sparse vector and $y = x - \eta\nabla f(x)$. If f is M_{2k}-smooth, then the following inequality holds:*

$$f(y_k) \leq f(x) - \frac{1-\eta M_{2k}}{2\eta}\|y_k - x\|^2.$$

Proof Since f is M_{2k}-smooth, it follows that

$$\begin{aligned} f(y_k) - f(x) \leq & \langle \nabla f(x), y_k - x\rangle + \frac{M_{2k}}{2}\|y_k - x\|^2 \\ \overset{\xi_1}{\leq} & -\frac{1}{2\eta}\|y_k - x\|^2 + \frac{M_{2k}}{2}\|y_k - x\|^2 \\ = & -\frac{1-\eta M_{2k}}{2\eta}\|y_k - x\|^2, \end{aligned}$$

where "ξ_1" follows from the fact that y_k is the best k-support approximation to y such that

$$\|y_k - y\|^2 = \|y_k - x + \eta\nabla f(x)\|^2 \leq \|x - x + \eta\nabla f(x)\|^2 = \|\eta\nabla f(x)\|^2,$$

which implies $2\eta\langle\nabla f(x), y_k - x\rangle \leq -\|y_k - x\|^2$. ■

Lemma 14 *Assume that f is m_s-strongly convex. Then for any $\|x - x'\|_0 \leq s$ it holds that*

$$\|x - x'\| \leq \sqrt{\frac{2\max\{f(x) - f(x'), 0\}}{m_s}} + \frac{2\|\nabla_{F\cup F'} f(x')\|}{m_s},$$

where $F = supp(x)$ and $F' = supp(x')$.

Proof Since f is m_s-strongly convex, we have

$$\begin{aligned} f(x) \geq & f(x') + \langle \nabla f(x'), x - x'\rangle + \frac{m_s}{2}\|x - x'\|^2 \\ \geq & f(x') - \|\nabla_{F\cup F'} f(x')\|\|x - x'\| + \frac{m_s}{2}\|x - x'\|^2, \end{aligned}$$

where the second inequality follows from Cauchy-Schwarz inequality. From this above inequality we can see that if $f(x) \leq f(x')$, then

$$\|x - x'\| \leq \frac{2\|\nabla_{F\cup F'} f(x')\|}{m_s}.$$

If otherwise $f(x) > f(x')$, then we have

$$\begin{aligned} \|x - x'\| \leq & \frac{\|\nabla_{F\cup F'} f(x')\| + \sqrt{\|\nabla_{F\cup F'} f(x')\|^2 + 2m_s(f(x) - f(x'))}}{m_s} \\ \leq & \frac{2\|\nabla_{F\cup F'} f(x')\| + \sqrt{2m_s(f(x) - f(x'))}}{m_s}. \end{aligned}$$

By combining the above two cases we get the desired bound. ■

Lemma 15 *Assume that f is m_s-strongly convex and M_s-smooth. For any index set F with cardinality $|F| \leq s$ and any x, y with $supp(x) \cup supp(y) \subseteq F$, if $\eta \in (0, 2m_s/M_s^2)$, then*

$$\|x - y - \eta\nabla_F f(x) + \eta\nabla_F f(y)\| \leq \sqrt{1 - 2\eta m_s + \eta^2 M_s^2}\|x - y\|,$$

and $\sqrt{1 - 2\eta m_s + \eta^2 M_s^2} < 1$.

Proof By adding two copies of the inequality (2) with x and y interchanged and applying Theorem 2.1.5 in the textbook (Nesterov, 2004) on the supporting set F, we can show that

$$(x - y)^\top(\nabla f(x) - \nabla f(y)) \geq m_s\|x - y\|^2, \quad \|\nabla_F f(x) - \nabla_F f(y)\| \leq M_s\|x - y\|.$$

Then for any $\eta > 0$ we have

$$\|x - y - \eta\nabla_F f(x) + \eta\nabla_F f(y)\|^2 \leq (1 - 2\eta m_s + \eta^2 M_s^2)\|x - y\|^2.$$

It is clear that $1 - 2\eta m_s + \eta^2 M_s^2 \geq 1 - m_s^2/M_s^2 \geq 0$. The condition $\eta < 2m_s/M_s^2$ implies $\sqrt{1 - 2\eta m_s + \eta^2 M_s^2} < 1$. This proves the lemma. ■

Lemma 16 *Assume that f is M_s-smooth and m_s-strongly convex. Let F and F' be two index sets with cardinality $|F \cup F'| = s$. Let $x = \arg\min_{supp(y) \subseteq F} f(y)$ and $supp(x') \subseteq F'$. Then for any $\eta \in (0, 2m_s/M_s^2)$, the following two inequalities hold:*

$$\|(x - x')_F\| \leq \frac{\rho\|x'_{F' \setminus F}\|}{1 - \rho} + \frac{\eta\|\nabla_{F \cup F'} f(x')\|}{1 - \rho}, \tag{14}$$

$$\|x - x'\| \leq \frac{\|x'_{F' \setminus F}\|}{1 - \rho} + \frac{\eta\|\nabla_{F \cup F'} f(x')\|}{1 - \rho}, \tag{15}$$

where $\rho = \sqrt{1 - 2\eta m_s + \eta^2 M_s^2} < 1$.

Proof Since x is the minimum of $f(y)$ restricted over the supporting set F, we have $\langle\nabla f(x), z\rangle = 0$ whenever $\mathrm{supp}(z) \subseteq F$. Then

$$\begin{aligned}
&\|(x - x')_F\|^2 \\
=&\langle x - x', (x - x')_F\rangle \\
=&\langle x - x' - \eta\nabla_{F \cup F'} f(x) + \eta\nabla_{F \cup F'} f(x'), (x - x')_F\rangle - \eta\langle\nabla_{F \cup F'} f(x'), (x - x')_F\rangle \\
\overset{\xi_1}{\leq}&\sqrt{1 - 2\eta m_s + \eta^2 M_s^2}\|x - x'\|\|(x - x')_F\| + \eta\|\nabla_{F \cup F'} f(x')\|\|(x - x')_F\|,
\end{aligned}$$

where "ξ_1" follows from Lemma 15. Let us abbreviate $\rho = \sqrt{1 - 2\eta m_s + \eta^2 M_s^2}$. After simplification, we have

$$\|(x - x')_F\| \leq \rho\|x - x'\| + \eta\|\nabla_{F \cup F'} f(x')\|. \tag{16}$$

It follows that

$$\begin{aligned}
\|x - x'\| \leq&\|(x - x')_F\| + \|(x - x')_{F' \setminus F}\| \\
\leq&\rho\|x - x'\| + \eta\|\nabla_{F \cup F'} f(x')\| + \|(x - x')_{F' \setminus F}\|.
\end{aligned}$$

After rearrangement we obtain

$$\begin{aligned}\|x - x'\| \leq & \frac{\|(x - x')_{F' \setminus F}\|}{1 - \rho} + \frac{\eta \|\nabla_{F \cup F'} f(x')\|}{1 - \rho} \\ = & \frac{\|x'_{F' \setminus F}\|}{1 - \rho} + \frac{\eta \|\nabla_{F \cup F'} f(x')\|}{1 - \rho}.\end{aligned} \tag{17}$$

By combining (16) and (17) we get

$$\|(x - x')_F\| \leq \frac{\rho \|x'_{F' \setminus F}\|}{1 - \rho} + \frac{\eta \|\nabla_{F \cup F'} f(x')\|}{1 - \rho}.$$

This proves the desired bounds in this lemma. ■

The following lemma is established by Shen & Li (2016, Theorem 1) for bounding the estimation error of hard-thresholding operation. This result will be extensively used in our analysis.

Lemma 17 *Let $b \in \mathbb{R}^p$ be an arbitrary p-dimensional vector and $a \in \mathbb{R}^p$ be any k-sparse vector. Denote $\bar{k} = \|a\|_0 \leq k$. Then, we have the following universal bound:*

$$\|b_k - a\|^2 \leq \nu \|b - a\|^2, \quad \nu = 1 + \frac{\beta + \sqrt{(4 + \beta)\beta}}{2}, \quad \beta = \frac{\min\{\bar{k}, p - k\}}{k - \bar{k} + \min\{\bar{k}, p - k\}}.$$

Appendix B. Proofs of Main Theorems in Section 3

The technical proofs of main results in Section 3 are collected in this appendix section.

B.1 Proof of Theorem 2

Before proving Theorem 2, we first present two lemmas which are respectively key to the proof of part(a) and part(b) of Theorem 2.

Lemma 18 *Assume that f is M_{3k}-smooth and m_{3k}-strongly convex. Let $\bar{x}$ be an arbitrary k-sparse vector. Then at time instance t, for any $\eta \in (0, 2m_{3k}/M_{3k}^2)$, GraHTP will output $x^{(t)}$ satisfying*

$$\|x^{(t)} - \bar{x}\| \leq \frac{\rho}{1 - \rho} \|x^{(t-1)} - \bar{x}\| + \frac{2\eta \|\nabla_{2k} f(\bar{x})\|}{1 - \rho},$$

where $\rho = \sqrt{1 - 2\eta m_{3k} + \eta^2 M_{3k}^2} < 1$.

Proof Denote $\bar{F} = \mathrm{supp}(\bar{x})$. Since $x^{(t)}$ is the minimum of $f(x)$ restricted over the supporting set $F^{(t)}$, it is directly known from the inequality (15) in Lemma 16 that

$$\|x^{(t)} - \bar{x}\| \leq \frac{\|(x^{(t)} - \bar{x})_{\bar{F} \setminus F^{(t)}}\|}{1 - \rho} + \frac{\eta \|\nabla_{F^{(t)}} f(\bar{x})\|}{1 - \rho}. \tag{18}$$

According to the definition of $F^{(t)}$,

$$\|(x^{(t-1)} - \eta\nabla f(x^{(t-1)}))_{\bar{F}}\| \leq \|(x^{(t-1)} - \eta\nabla f(x^{(t-1)}))_{F^{(t)}}\|.$$

By eliminating the contribution on $\bar{F} \cap F^{(t)}$ we get

$$\|(x^{(t-1)} - \eta\nabla f(x^{(t-1)}))_{\bar{F}\setminus F^{(t)}}\| \leq \|(x^{(t-1)} - \eta\nabla f(x^{(t-1)}))_{F^{(t)}\setminus \bar{F}}\|. \tag{19}$$

For the right-hand side, we can derive

$$\begin{aligned} &\|(x^{(t-1)} - \eta\nabla f(x^{(t-1)}))_{F^{(t)}\setminus \bar{F}}\| \\ \leq &\|(x^{(t-1)} - \bar{x} - \eta\nabla f(x^{(t-1)}) + \eta\nabla f(\bar{x}))_{F^{(t)}\setminus \bar{F}}\| + \eta\|\nabla_{F^{(t)}\setminus \bar{F}} f(\bar{x})\|. \end{aligned} \tag{20}$$

As for the left-hand side, we can see that

$$\begin{aligned} &\|(x^{(t-1)} - \eta\nabla f(x^{(t-1)}))_{\bar{F}\setminus F^{(t)}}\| \\ \geq &\|(x^{(t-1)} - \bar{x} - \eta\nabla f(x^{(t-1)}) + \eta\nabla f(\bar{x}))_{\bar{F}\setminus F^{(t)}} - (x^{(t)} - \bar{x})_{\bar{F}\setminus F^{(t)}} - \eta\nabla_{\bar{F}\setminus F^{(t)}} f(\bar{x})\| \\ \geq &\|(x^{(t)} - \bar{x})_{\bar{F}\setminus F^{(t)}}\| - \|(x^{(t-1)} - \bar{x} - \eta\nabla f(x^{(t-1)}) + \eta\nabla f(\bar{x}))_{\bar{F}\setminus F^{(t)}}\| \\ &- \eta\|\nabla_{\bar{F}\setminus F^{(t)}} f(\bar{x})\|. \end{aligned} \tag{21}$$

Denote $\bar{F}\Delta F^{(t)}$ the symmetric difference of $\bar{F}$ and $F^{(t)}$ and let $F = \bar{F} \cup F^{(t)} \cup F^{(t-1)}$. It can be shown from (19), (20) and (21) that

$$\begin{aligned} &\|(x^{(t)} - \bar{x})_{\bar{F}\setminus F^{(t)}}\| \\ \leq &\|(x^{(t-1)} - \bar{x} - \eta\nabla f(x^{(t-1)}) + \eta\nabla f(\bar{x}))_{\bar{F}\Delta F^{(t)}}\| + \eta\|\nabla_{\bar{F}\Delta F^{(t)}} f(\bar{x})\| \\ \leq &\|x^{(t-1)} - \bar{x} - \eta\nabla_F f(x^{(t-1)}) + \eta\nabla_F f(\bar{x})\| + \eta\|\nabla_{F^{(t)}\Delta \bar{F}} f(\bar{x})\| \\ \overset{\xi_1}{\leq} &\rho\|x^{(t-1)} - \bar{x}\| + \eta\|\nabla_{F^{(t)}\Delta \bar{F}} f(\bar{x})\|, \end{aligned} \tag{22}$$

where "ξ_1" follows from Lemma 15. As a final step, combining (18) and (22) gives us

$$\begin{aligned} \|x^{(t)} - \bar{x}\| &\leq \frac{\rho}{1-\rho}\|x^{(t-1)} - \bar{x}\| + \frac{2\eta\|\nabla_{F^{(t)}\cup \bar{F}} f(\bar{x})\|}{1-\rho} \\ &\leq \frac{\rho}{1-\rho}\|x^{(t-1)} - \bar{x}\| + \frac{2\eta\|\nabla_{2k} f(\bar{x})\|}{1-\rho}. \end{aligned}$$

This completes the proof. ■

Lemma 19 *Let $\bar{x}$ be an arbitrary $\bar{k}$-sparse vector. Assume that $s = 2k + \bar{k} \leq p$ and f is M_s-smooth and m_s-strongly convex. Then at time instance t, for any $\eta \in (0, 2m_s/M_s^2)$, FGraHTP will output $x^{(t)}$ satisfying*

$$\|x^{(t)} - \bar{x}\| \leq \gamma\rho\|x^{(t-1)} - \bar{x}\| + \gamma\eta\|\nabla_s f(\bar{x})\|,$$

where $\rho = \sqrt{1 - 2\eta m_s + \eta^2 M_s^2} < 1$ and $\gamma = \sqrt{1 + \left(\bar{k}/k + \sqrt{(4 + \bar{k}/k)\bar{k}/k}\right)/2}$.

Proof Recall that $F^{(t)} = \text{supp}(x^{(t)})$ and $F = F^{(t-1)} \cup F^{(t)} \cup \text{supp}(\bar{x})$. Consider the following vector

$$y = x^{(t-1)} - \eta \nabla_F f(x^{(t-1)}).$$

By using triangular inequality,

$$\begin{aligned} \|y - \bar{x}\| &= \|x^{(t-1)} - \eta \nabla_F f(x^{(t-1)}) - \bar{x}\| \\ &\leq \|x^{(t-1)} - \bar{x} - \eta \nabla_F f(x^{(t-1)}) + \eta \nabla_F f(\bar{x})\| + \eta \|\nabla_F f(\bar{x})\| \\ &\leq \rho \|x^{(t-1)} - \bar{x}\| + \eta \|\nabla_s f(\bar{x})\|, \end{aligned}$$

where the last inequality follows from Lemma 15 and $\|\nabla_F f(\bar{x})\| \leq \|\nabla_s f(\bar{x})\|$. We note that $x^{(t)} = y_k$ in FGraHTP. Then, by invoking Lemma 17 we get

$$\|x^{(t)} - \bar{x}\| \leq \gamma \|y - \bar{x}\|,$$

where $\gamma = \sqrt{1 + \left(\bar{k}/k + \sqrt{(4 + \bar{k}/k)\bar{k}/k}\right)/2}$. It follows that

$$\|x^{(t)} - \bar{x}\| \leq \gamma\rho \|x^{(t-1)} - \bar{x}\| + \gamma\eta \|\nabla_s f(\bar{x})\|.$$

This proves the desired bound. ■

Equipped with Lemma 18 and Lemma 19, we can now prove Theorem 2 in a straightforward way.

Proof [of Theorem 2]

Part(a): Since $M_{3k}/m_{3k} < 2\sqrt{3}/3$, there exists $\eta \in (0, 2m_{3k}/M_{3k}^2)$ such that $\rho = \sqrt{1 - 2\eta m_{3k} + \eta^2 M_{3k}^2} < 0.5$ and thus $\rho/(1-\rho) < 1$. By recursively applying Lemma 18 and noting the fact $\|\nabla_s f(x)\| \leq \sqrt{s}\|\nabla f(x)\|_\infty$ we obtain the desired bound in this part.

Part(b): Note that $\gamma = 1.62$ when $\bar{k} = k$ in Lemma 19. Since $M_{3k}/m_{3k} < 1.26$, there exists $\eta \in (0, 2m_{3k}/M_{3k}^2)$ such that $\rho < 0.62$ and thus $1.62\rho < 1$. Then by recursively applying Lemma 19 with $\bar{k} = k$ we obtain the desired bound in this part. ■

B.2 Proof of Theorem 5

We need the following lemma to prove Theorem 5.

Lemma 20 *Assume that f is M_{2k}-smooth and m_{2k}-strongly convex. Assume the step-size $\eta < 1/M_{2k}$. Let $\bar{x}$ be an arbitrary $\bar{k}$-sparse vector with $k \geq \left(2 + \frac{4}{\eta^2 m_{2k}^2}\right)\bar{k}$. Then GraHTP outputs $x^{(t)}$ satisfying*

$$f(x^{(t)}) \leq f(\bar{x}) + (1 - \bar{\nu})^t \bar{\triangle}^{(0)},$$

where $\bar{\nu} = \eta m_{2k}(1 - \eta M_{2k})/2 \in (0, 0.125 m_{2k}/M_{2k})$ and $\bar{\triangle}^{(0)} = f(x^{(0)}) - f(\bar{x})$.

Proof From the definition of $\tilde{x}^{(t)}$ we know that the following inequality holds:

$$\|\tilde{x}^{(t)}_{F^{(t)}} - x^{(t-1)}\| \geq \eta \|\nabla_{F^{(t)} \setminus F^{(t-1)}} f(x^{(t-1)})\|. \tag{23}$$

From Lemma 13 we get

$$f(x^{(t)}) - f(x^{(t-1)}) \leq f(\tilde{x}^{(t)}_{F^{(t)}}) - f(x^{(t-1)}) \leq -\frac{1-\eta M_{2k}}{2\eta}\|\tilde{x}^{(t)}_{F^{(t)}} - x^{(t-1)}\|^2. \tag{24}$$

Combining the above two inequalities (23) and (24) gives us

$$f(x^{(t)}) - f(x^{(t-1)}) \leq -\frac{(1-\eta M_{2k})\eta}{2}\|\nabla_{F^{(t)}\setminus F^{(t-1)}} f(x^{(t-1)})\|^2. \tag{25}$$

Let $\bar{F} = \text{supp}(\bar{x})$. Under the conditions in the theorem, we claim

$$\|\nabla_{F^{(t)}\setminus F^{(t-1)}} f(x^{(t-1)})\|^2 \geq m_{2k}\left[f(x^{(t-1)}) - f(\bar{x})\right]. \tag{26}$$

To prove this , let us distinguish the following two mutually complementary cases:

- Case I: $|F^{(t)} \setminus F^{(t-1)}| \geq \bar{k}$. In this case, we have $|F^{(t)} \setminus F^{(t-1)}| \geq |\bar{F} \setminus F^{(t-1)}|$. From the m_{2k}-strong convexity of f we have

$$\begin{aligned}
&\frac{m_{2k}}{2}\|\bar{x} - x^{(t-1)}\|^2 \\
\leq& f(\bar{x}) - f(x^{(t-1)}) - (\bar{x} - x^{(t-1)})^\top \nabla f(x^{(t-1)}) \\
\overset{\xi_1}{\leq}& f(\bar{x}) - f(x^{(t-1)}) + \frac{m_{2k}}{2}\|\bar{x} - x^{(t-1)}\|^2 + \frac{1}{2m_{2k}}\|\nabla_{\bar{F}\setminus F^{(t-1)}} f(x^{(t-1)})\|^2,
\end{aligned}$$

  where "ξ_1" follows from Cauchy-Schwartz inequality, a basic inequality $ma^2/2 + b^2/(2m) \geq ab$ for any $m > 0$, and $\nabla_{F^{(t-1)}} f(x^{(t-1)}) = 0$. This implies

$$\|\nabla_{\bar{F}\setminus F^{(t-1)}} f(x^{(t-1)})\|^2 \geq 2m_{2k}\left[f(x^{(t-1)}) - f(\bar{x})\right]. \tag{27}$$

  Since $F^{(t)} \setminus F^{(t-1)}$ contains the top $|F^{(t)} \setminus F^{(t-1)}|$ (in magnitude) entries in $\nabla f(x^{(t-1)})$ and $|F^{(t)} \setminus F^{(t-1)}| \geq |\bar{F} \setminus F^{(t-1)}|$, it follows that

$$\|\nabla_{F^{(t)}\setminus F^{(t-1)}} f(x^{(t-1)})\|^2 \geq \|\nabla_{\bar{F}\setminus F^{(t-1)}} f(x^{(t-1)})\|^2 \geq 2m_{2k}\left[f(x^{(t-1)}) - f(\bar{x})\right].$$

- Case II: $|F^{(t)} \setminus F^{(t-1)}| < \bar{k}$. In this case, from the step (**S2**) we know that each element of $\tilde{x}^{(t)}$ over $\bar{F} \setminus (F^{(t)} \cup F^{(t-1)})$ has smaller magnitude than that over $F^{(t)} \cap F^{(t-1)}$. This implies

$$\frac{\|\tilde{x}^{(t)}_{\bar{F}\setminus(F^{(t)}\cup F^{(t-1)})}\|^2}{|\bar{F} \setminus (F^{(t)} \cup F^{(t-1)})|} \leq \frac{\|\tilde{x}^{(t)}_{(F^{(t)}\cap F^{(t-1)})\setminus\bar{F}}\|^2}{|(F^{(t)} \cap F^{(t-1)}) \setminus \bar{F}|}.$$

  Since $\tilde{x}^{(t)}_{\bar{F}\setminus(F^{(t)}\cup F^{(t-1)})} = -\eta\nabla_{\bar{F}\setminus(F^{(t)}\cup F^{(t-1)})} f(x^{(t-1)})$, $\tilde{x}^{(t)}_{(F^{(t)}\cap F^{(t-1)})\setminus\bar{F}} = x^{(t-1)}_{(F^{(t)}\cap F^{(t-1)})\setminus\bar{F}}$, $|\bar{F} \setminus (F^{(t)} \cup F^{(t-1)})| \leq \bar{k}$ and $|(F^{(t)} \cap F^{(t-1)}) \setminus \bar{F}| \geq k - 2\bar{k}$, we have

$$\eta^2\|\nabla_{\bar{F}\setminus(F^{(t)}\cup F^{(t-1)})} f(x^{(t-1)})\|^2 \leq \frac{\bar{k}}{k-2\bar{k}}\|x^{(t-1)}_{(F^{(t)}\cap F^{(t-1)})\setminus\bar{F}}\|^2.$$

From the m_{2k}-strong convexity of f we have

$$\begin{aligned}
&\frac{m_{2k}}{2}\|\bar{x}-x^{(t-1)}\|^2\\
\leq& f(\bar{x})-f(x^{(t-1)})-(\bar{x}-x^{(t-1)})^\top\nabla f(x^{(t-1)})\\
\overset{\xi_1}{\leq}& f(\bar{x})-f(x^{(t-1)})+\frac{m_{2k}}{4}\|\bar{x}-x^{(t-1)}\|^2+\frac{1}{m_{2k}}\|\nabla_{\bar{F}\setminus F^{(t-1)}}f(x^{(t-1)})\|^2\\
\leq& f(\bar{x})-f(x^{(t-1)})+\frac{m_{2k}}{4}\|\bar{x}-x^{(t-1)}\|^2+\frac{1}{m_{2k}}\|\nabla_{\bar{F}\setminus(F^{(t)}\cup F^{(t-1)})}f(x^{(t-1)})\|^2\\
&+\frac{1}{m_{2k}}\|\nabla_{(F^{(t)}\setminus F^{(t-1)})\cap\bar{F}}f(x^{(t-1)})\|^2\\
\overset{\xi_2}{\leq}& f(\bar{x})-f(x^{(t-1)})+\frac{m_{2k}}{4}\|\bar{x}-x^{(t-1)}\|^2+\frac{\bar{k}}{\eta^2(k-2\bar{k})m_{2k}}\|x^{(t-1)}_{(F^{(t)}\cap F^{(t-1)})\setminus\bar{F}}\|^2\\
&+\frac{1}{m_{2k}}\|\nabla_{F^{(t)}\setminus F^{(t-1)}}f(x^{(t-1)})\|^2\\
\overset{\xi_3}{\leq}& f(\bar{x})-f(x^{(t-1)})+\frac{m_{2k}}{4}\|\bar{x}-x^{(t-1)}\|^2+\frac{\bar{k}}{\eta^2(k-2\bar{k})m_{2k}}\|\bar{x}-x^{(t-1)}\|^2\\
&+\frac{1}{m_{2k}}\|\nabla_{F^{(t)}\setminus F^{(t-1)}}f(x^{(t-1)})\|^2,
\end{aligned}$$

where "ξ_1" follows from Cauchy-Schwartz inequality, $ma^2/4+b^2/(m)\geq ab$ for any $m>0$, and $\nabla_{F^{(t-1)}}f(x^{(t-1)})=0$, "$\xi_2$" follows from the preceding inequality, and "ξ_3" is due to $\|x^{(t-1)}_{(F^{(t)}\cap F^{(t-1)})\setminus\bar{F}}\|\leq\|\bar{x}-x^{(t-1)}\|$. Since $k\geq\left(2+\frac{4}{\eta^2m_{2k}^2}\right)\bar{k}$, the above inequality leads to

$$\|\nabla_{\bar{F}\setminus F^{(t-1)}}f(x^{(t-1)})\|^2\geq m_{2k}\left[f(x^{(t-1)})-f(\bar{x})\right].$$

Since $\eta<1/M_{2k}$, from (25) and (26) we get that

$$\begin{aligned}
f(x^{(t)})&\leq f(x^{(t-1)})-\frac{\eta m_{2k}(1-\eta M_{2k})}{2}\left[f(x^{(t-1)})-f(\bar{x})\right]\\
&=f(x^{(t-1)})-\bar{\nu}\left[f(x^{(t-1)})-f(\bar{x})\right].
\end{aligned}$$

Therefore, we get

$$f(x^{(t)})-f(\bar{x})\leq(1-\bar{\nu})(f(x^{(t-1)})-f(\bar{x})).$$

Since $m_{2k}\leq M_{2k}$ and $\eta\in(0,1/M_{2k})$, it can be verified that $\bar{\nu}\in(0,0.125m_{2k}/M_{2k})$. By recursively applying the above inequality we obtain the desired result. ■

We are now in the position to prove Theorem 5.

Proof [of Theorem 5] **Part(a):** Since $\eta<1/M_{2k}$, it is directly known from Lemma 13 that $\{f(x^{(t)})\}$ is monotonically decreasing. From Lemma 14 we know that the result holds when

$f(x^{(t)}) \leq f(\bar{x})$. Therefore, we only need to consider the case when $f(x^{(t)}) > f(\bar{x})$. In this case, from Lemma 14 and Lemma 20 we get

$$\begin{aligned}\|x^{(t)} - \bar{x}\| \leq & \sqrt{\frac{2(f(x^{(t)}) - f(\bar{x}))}{m_{2k}}} + \frac{2\|\nabla_{2k} f(\bar{x})\|}{m_{2k}} \\ \leq & \sqrt{\frac{2(1-\bar{\nu})^t \bar{\triangle}^{(0)}}{m_{2k}}} + \frac{2\sqrt{2k}\|\nabla f(\bar{x})\|_\infty}{m_{2k}}.\end{aligned}$$

This proves the result in part(a).

Part(b): From the condition $k \geq \rho\bar{k}/(1-\rho)^2$ we can verify that $\bar{\mu}_2 = \rho\gamma < 1$. Thus, the result can be directly proved by recursively applying Lemma 19. ■

Appendix C. Proofs of Main Theorems in Section 4

The technical proofs of main results in Section 4 are collected in this appendix section.

C.1 Proof of Theorem 8

Before commencing with the actual proof, we first present an overview of the proof procedure which consists of the following three key ingredients:

(a) We first prove that under the given conditions, GraHTP will not terminate (i.e., $F^{(t)} \neq F^{(t-1)}$) whenever $\mathrm{supp}(\bar{x}) \nsubseteq \mathrm{supp}(x^{(t-1)})$.

(b) We then show that $\mathrm{supp}(\bar{x}) = \mathrm{supp}(x^{(t)}, \bar{k})$ when GraHTP terminates at $x^{(t)}$.

(c) Finally we show that the conditions in the theorem guarantee finite termination of GraHTP and analyze its iteration complexity before termination.

Proof [of Theorem 8]

We first show that $F^{(t)} \neq F^{(t-1)}$ whenever $\mathrm{supp}(\bar{x}) \nsubseteq \mathrm{supp}(x^{(t-1)})$. To this end, let us assume $\mathrm{supp}(\bar{x}) \nsubseteq \mathrm{supp}(x^{(t-1)})$. Recall $x^\star = \arg\min_{\|x\|_0 \leq k} f(x)$. Then

$$\begin{aligned}\bar{x}_{\min} + \|x^{(t-1)}_{F^{(t-1)} \setminus \bar{F}}\| \leq & \|\bar{x} - x^{(t-1)}\| \\ \overset{\xi_1}{\leq} & \sqrt{\frac{2\max\left\{f(\bar{x}) - f(x^{(t-1)}), 0\right\}}{m_{2k}}} + \frac{2\|\nabla_{\bar{F} \setminus F^{(t-1)}} f(x^{(t-1)})\|}{m_{2k}} \\ \overset{\xi_2}{\leq} & \sqrt{\frac{2(f(\bar{x}) - f(x^\star))}{m_{2k}}} + \frac{2\|\nabla_{\bar{F} \setminus F^{(t-1)}} f(x^{(t-1)})\|}{m_{2k}},\end{aligned}$$

where "ξ_1" follows from Lemma 14 and "ξ_2" is due to the fact of $f(x^{(t-1)}) \geq f(x^\star)$. Since it is assumed $\bar{x}_{\min} > 1.62\sqrt{\frac{2(f(\bar{x}) - f(x^\star))}{m_{2k}}}$, the above inequality implies

$$\|x^{(t-1)}_{F^{(t-1)} \setminus \bar{F}}\| < \frac{2\|\nabla_{\bar{F} \setminus F^{(t-1)}} f(x^{(t-1)})\|}{m_{2k}},$$

which then gives us

$$\sqrt{k-\bar{k}}x_{\min}^{(t-1)} < \frac{2\sqrt{\bar{k}}}{m_{2k}}\|\nabla f(x^{(t-1)})\|_\infty.$$

Since $\eta = \frac{1}{2M_{2k}}$ and $k \geq \left(1+\frac{16M_{2k}^2}{m_{2k}^2}\right)\bar{k}$, we then have

$$\eta\|\nabla f(x^{(t-1)})\|_\infty > x_{\min}^{(t-1)}.$$

This means that at least the smallest nonzero entry of $x^{(t-1)}$ and the largest entry of $\nabla f(x^{(t-1)})$ can be swapped in step (**S2**) of Algorithm 1, and thus $F^{(t)} \neq F^{(t-1)}$. Therefore, when the algorithm terminates at time instance t, i.e., $F^{(t)} = F^{(t-1)}$, we must have $\mathrm{supp}(\bar{x}) \subseteq \mathrm{supp}(x^{(t)})$.

Next we show that $\mathrm{supp}(\bar{x}) = \mathrm{supp}(x^{(t)}, \bar{k})$ when GraHTP terminates at time instance t with $\mathrm{supp}(\bar{x}) \subseteq \mathrm{supp}(x^{(t)})$. Assume otherwise $\mathrm{supp}(\bar{x}) \neq \mathrm{supp}(x^{(t)}, \bar{k})$. Then

$$\begin{aligned}
\bar{x}_{\min} &\leq \|\bar{x} - x_{\bar{k}}^{(t)}\| \\
&\overset{\xi_1}{\leq} 1.62\|\bar{x} - x^{(t)}\| \\
&\overset{\xi_2}{\leq} 1.62\left(\sqrt{\frac{2(f(\bar{x}) - f(x^\star))}{m_{2k}}} + \frac{2\|\nabla_{\bar{F}\setminus F^{(t)}} f(x^{(t)})\|}{m_{2k}}\right) \\
&\overset{\xi_3}{=} 1.62\sqrt{\frac{2(f(\bar{x}) - f(x^\star))}{m_{2k}}},
\end{aligned}$$

where "ξ_1" is based on the truncation error bound given by Shen & Li (2016, Theorem 1), "ξ_2" follows from Lemma 14 and the fact of $f(x^{(t)}) \geq f(x^\star)$, and "$\xi_3$" is the consequence of $\bar{F} \subseteq F^{(t)}$. This above inequality contradicts the assumption on $\bar{x}_{\min}$. Therefore, it must hold that $\mathrm{supp}(\bar{x}) = \mathrm{supp}(x^{(t)}, \bar{k})$.

Now we claim that GraHTP is finite under the assumed conditions. Indeed, based on Lemma 13 it is easy to verify that when $\eta = \frac{1}{2M_{2k}}$, the sequence $\{f(x^{(t)})\}$ generated by Algorithm 1 is monotonically decreasing. Since the number of k-support index sets is finite, the sequence $\{f(x^{(t)})\}$ will be eventually periodic, and thus must be eventually a constant. Therefore we deduce that $\tilde{x}_k^{(t)} = x^{(t-1)}$, i.e., $F^{(t)} = F^{(t-1)}$, when t is sufficiently enough.

Finally, we estimate the iteration complexity bound before algorithm termination. Suppose that $F^{(t)} \neq F^{(t-1)}$ (otherwise GraHTP terminates at time instance t). From the step (**S3**) we know that $\nabla_{F^{(t-1)}} f(x^{(t-1)}) = 0$. By definition of $F^{(t)}$ we may decompose $F^{(t)} = G_1 \cup (F^{(t-1)} \setminus G_2)$ with $G_1 \subseteq \mathrm{supp}(\nabla f(x^{(t-1)}))$, $G_2 \subseteq F^{(t-1)}$ and $|G_1| = |G_2| = k' \leq k$. Here, G_1 contains the top k' (in magnitude) entries in $\nabla f(x^{(t-1)})$ while G_2 contains the bottom k' nonzero entries in $x^{(t-1)}$. Since $F^{(t)} \neq F^{(t-1)}$, we have $k' \geq 1$. From the step (**S2**) we know that

$$\|x_{G_2}^{(t-1)}\| < \eta\|\nabla_{G_1} f(x^{(t-1)})\|. \tag{28}$$

Let $F = F^{(t-1)} \cup \mathrm{supp}(x^\star)$. From the conditions in the theorem we have

$$\begin{aligned}
\frac{m_{2k}}{2}\|x^\star - x^{(t-1)}\|^2 &\leq f(x^\star) - f(x^{(t-1)}) - (x^\star - x^{(t-1)})^\top \nabla f(x^{(t-1)}) \\
&\leq f(x^\star) - f(x^{(t-1)}) + \frac{m_{2k}}{2}\|x^\star - x^{(t-1)}\|^2 + \frac{1}{2m_{2k}}\|\nabla_F f(x^{(t-1)})\|^2,
\end{aligned}$$

where the second inequality follows from Cauchy-Schwartz inequality and a basic inequality $ma^2/2 + b^2/(2m) \geq ab$ for any $m > 0$. This implies

$$\|\nabla_F f(x^{(t-1)})\|^2 \geq 2m_{2k}\left[f(x^{(t-1)}) - f(x^\star)\right].$$

Let $F^\star = \text{supp}(x^\star)$ and $k'' = |F^\star \setminus F^{(t-1)}|$. Obviously, we have $k'' \leq k$. Based on the above arguments, it can be verified that

$$\begin{aligned} k\|\nabla_{G_1} f(x^{(t-1)})\|^2 &\geq (k''/k')\|\nabla_{G_1} f(x^{(t-1)})\|^2 \\ &\geq \|\nabla_F f(x^{(t-1)})\|^2 \\ &\geq 2m_{2k}\left[f(x^{(t-1)}) - f(x^\star)\right]. \end{aligned} \tag{29}$$

Now let $y^{(t)} := x^{(t-1)} + \delta^{(t-1)}$ in which

$$\delta^{(t-1)} = -\eta\nabla_{G_1} f(x^{(t-1)}) - x^{(t-1)}_{G_2}.$$

From the steps (**S1**) and (**S3**) in Algorithm 1 we get

$$\begin{aligned} f(x^{(t)}) &\leq f(y^{(t)}) \\ &\leq f(x^{(t-1)}) + \langle \nabla f(x^{(t-1)}), \Delta^{(t-1)}\rangle + \frac{M_{2k}}{2}\|\Delta^{(t-1)}\|^2 \\ &\leq f(x^{(t-1)}) + \frac{M_{2k}}{2}\|x^{(t-1)}_{G_2}\|^2 - \frac{2\eta - \eta^2 M_{2k}}{2}\|\nabla_{G_1} f(x^{(t-1)})\|^2 \\ &\overset{\xi_1}{\leq} f(x^{(t-1)}) - (\eta - \eta^2 M_{2k})\|\nabla_{G_1} f(x^{(t-1)})\|^2 \\ &\overset{\xi_2}{\leq} f(x^{(t-1)}) - \frac{m_{2k}}{2kM_{2k}}(f(x^{(t-1)}) - f(x^\star)), \end{aligned}$$

where "ξ_1" follows from (28) and "ξ_2" uses (29) and $\eta = \frac{1}{M_{2k}}$ as well. Therefore, we get

$$f(x^{(t)}) - f(x^\star) \leq \left(1 - \frac{m_{2k}}{2kM_{2k}}\right)(f(x^{(t-1)}) - f(x^\star)).$$

Note that $f(x^{(t)}) \geq f(x^\star)$ for all $t \geq 0$. By recursively using the above inequality we get

$$f(x^{(t)}) - f(x^\star) \leq \left(1 - \frac{m_{2k}}{2kM_{2k}}\right)^t (f(x^{(0)}) - f(x^\star)).$$

Let us define the following quantity

$$\triangle^{-\star} = \min_{\|x\|_0 \leq k, \text{supp}(x) \neq \text{supp}(x^\star), f(x) > f(x^\star)} f(x) - f(x^\star).$$

Then $f(x^{(t)}) - f(x^\star) \leq \triangle^{-\star}$ when $t \geq \frac{2kM_{2k}}{m_{2k}} \ln \frac{\triangle^{(0)}}{\triangle^{-\star}}$ (note that $\triangle^{-\star} > 0$ by definition). After that, we have $f(x^{(t)}) = f(x^\star)$, i.e., $x^{(t)}$ is also a k-sparse minimizer. Then according to Lemma 21 we have $x^{(t)}_{\min} \geq \frac{\|\nabla f(x^{(t)})\|_\infty}{M_{2k}} > \eta\|\nabla f(x^{(t)})\|_\infty$, and thus the algorithm terminates at $x^{(t)}$. Based on the above arguments, we can conclude that GraHTP terminates after at most $t = \left\lceil \frac{2kM_{2k}}{m_{2k}} \ln \frac{\triangle^{(0)}}{\triangle^{-\star}} \right\rceil$ steps of iteration. This completes the proof. ■

C.2 Proof of Theorem 10

The following lemma gives a necessary condition on the k-sparse minimizer $x^\star$. A similar result was proved by Beck & Eldar (2013).

Lemma 21 *If f is M_{2k}-smooth, then the following inequality holds for the global k-sparse minimizer $x^\star = \arg\min_{\|x\|_0 \leq k} f(x)$:*

$$x^\star_{\min} \geq \frac{\|\nabla f(x^\star)\|_\infty}{M_{2k}}.$$

Proof Assume otherwise that $\vartheta^\star := \frac{M_{2k} x^\star_{\min}}{\|\nabla f(x^\star)\|_\infty} < 1$. Let us consider $\tilde{x}^\star = x^\star - \eta \nabla f(x^\star)$ with any $\eta \in (\vartheta^\star / M_{2k}, 1/M_{2k})$. From Lemma 13 we get that

$$f(\tilde{x}^\star_k)) \leq f(x^\star) - \frac{1 - \eta M_{2k}}{2\eta} \|\tilde{x}^\star_k - x^\star\|^2.$$

Since $\eta < \frac{1}{M_{2k}}$ and $x^\star_{\min} = \frac{\vartheta^\star \|\nabla f(x^\star)\|_\infty}{M_{2k}} < \eta \|\nabla f(x^\star)\|_\infty$, we have $\tilde{x}^\star_k \neq x^\star$ and thus it follows from the above inequality that $f(\tilde{x}^\star_k) < f(x^\star)$ which contradicts the optimality of $x^\star$. ■

Now we can prove the main result in Theorem 10.

Proof [of Theorem 10] We first show that $\mathrm{supp}(\bar{x}) = \mathrm{supp}(x^\star, \bar{k})$ if the condition (1) is satisfied. Assume otherwise $\mathrm{supp}(\bar{x}) \neq \mathrm{supp}(x^\star, \bar{k})$. From the optimality of $x^\star$ and $k \geq \bar{k}$ we have $f(x^\star) \leq f(\bar{x})$. By invoking Lemma 14 and the truncation error bound by Shen & Li (2016, Theorem 1) we get

$$\bar{x}_{\min} \leq \|x^\star_{\bar{k}} - \bar{x}\| \leq 1.62 \|x^\star - \bar{x}\| \leq \frac{3.24\sqrt{2k}\|\nabla f(\bar{x})\|_\infty}{m_{2k}} < \frac{4.59\sqrt{k}\|\nabla f(\bar{x})\|_\infty}{m_{2k}},$$

which contradicts the condition.

Next we prove that $\mathrm{supp}(\bar{x}) = \mathrm{supp}(x^\star, \bar{k})$ if the condition (2) is satisfied. Let $\bar{F} = \mathrm{supp}(\bar{x})$ and $F^\star = \mathrm{supp}(x^\star)$. We first claim that $\bar{F} \subseteq F^\star$. Indeed, if otherwise $\bar{F} \nsubseteq F^\star$, then

$$\begin{aligned}
\bar{x}_{\min} + \|x^\star_{F^\star \setminus \bar{F}}\| &\leq \|\bar{x} - x^\star\| \\
&\overset{\xi_1}{\leq} \sqrt{\frac{2\max\{f(\bar{x}) - f(x^\star), 0\}}{m_{2k}}} + \frac{2\|\nabla_{\bar{F}\setminus F^\star} f(x^\star)\|}{m_{2k}} \\
&= \sqrt{\frac{2(f(\bar{x}) - f(x^\star))}{m_{2k}}} + \frac{2\|\nabla_{\bar{F}\setminus F^\star} f(x^\star)\|}{m_{2k}},
\end{aligned}$$

where "ξ_1" follows from Lemma 14. Since $\bar{x}_{\min} > 1.62\sqrt{\frac{2(f(\bar{x}) - f(x^\star))}{m_{2k}}}$, the above inequality leads to

$$\sqrt{k - \bar{k}} x^\star_{\min} \leq \|x^\star_{F^\star \setminus \bar{F}}\| < \frac{2\|\nabla_{\bar{F}\setminus F^\star} f(x^\star)\|}{m_{2k}} \leq \frac{2\sqrt{\bar{k}}\|\nabla f(x^\star)\|_\infty}{m_{2k}}.$$

Since $k \geq \left(1+\frac{4M_{2k}^2}{m_{2k}^2}\right)\bar{k}$, we thus have $x^\star_{\min} < \frac{\|\nabla f(x^\star)\|_\infty}{M_{2k}}$. This contradicts Lemma 21. Therefore we must have $\bar{F} \subseteq F^\star$. Now let us assume $\mathrm{supp}(\bar{x}) \neq \mathrm{supp}(x^\star, \bar{k})$. Then

$$\begin{aligned}
\bar{x}_{\min} &\leq \|\bar{x} - x^\star_{\bar{k}}\| \\
&\overset{\xi_1}{\leq} 1.62\|\bar{x} - x^\star\| \\
&\overset{\xi_2}{\leq} 1.62\left(\sqrt{\frac{2(f(\bar{x}) - f(x^\star))}{m_{2k}}} + \frac{2\|\nabla_{\bar{F}\setminus F^\star} f(x^\star)\|}{m_{2k}}\right) \\
&\overset{\xi_3}{=} 1.62\sqrt{\frac{2(f(\bar{x}) - f(x^\star))}{m_{2k}}} < 2.3\sqrt{\frac{f(\bar{x}) - f(x^\star)}{m_{2k}}},
\end{aligned}$$

where "ξ_1" is based on the truncation error bound by Shen & Li (2016, Theorem 1), "ξ_2" follows from Lemma 14 and the fact of $f(\bar{x}) \geq f(x^\star)$, and "$\xi_3$" is the consequence of $\bar{F} \subseteq F^\star$. This above inequality contradicts the assumption on $\bar{x}_{\min}$. Therefore, it must hold that $\mathrm{supp}(\bar{x}) = \mathrm{supp}(x^\star, \bar{k})$. ■

Appendix D. Some Technical Details in Section 5

In this appendix section, we give the proof of Proposition 12 and present some implementation details of the proposed ADMM method for solving the subproblem (13).

D.1 Proof of Proposition 12

Proof It is straightforward to show that

$$\|\nabla f(\bar{w})\|_\infty \leq \|\nabla l(\bar{w})\|_\infty + \lambda\|\bar{w}\|_\infty. \tag{30}$$

We next bound the term $\|\nabla l(\bar{w})\|_\infty$. From (10) we have

$$\begin{aligned}
\left|\frac{\partial l}{\partial [\bar{w}]_j}\right| &= \left|\frac{1}{n}\sum_{i=1}^n -v^{(i)}[u^{(i)}]_j + \mathbb{E}_v[v[u^{(i)}]_j \mid u^{(i)}]\right| \\
&\leq \left|\frac{1}{n}\sum_{i=1}^n v^{(i)}[u^{(i)}]_j - \mathbb{E}[v[u]_j]\right| + \left|\frac{1}{n}\sum_{i=1}^n \mathbb{E}_v[v[u^{(i)}]_j \mid u^{(i)}] - \mathbb{E}[v[u]_j]\right|,
\end{aligned}$$

where $\mathbb{E}[\cdot]$ is taken over the distribution (9). Therefore, for any $\varepsilon > 0$,

$$\begin{aligned}
\mathbb{P}\left(\left|\frac{\partial l}{\partial [\bar{w}]_j}\right| > \varepsilon\right) \leq& \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n v^{(i)}[u^{(i)}]_j - \mathbb{E}[v[u]_j]\right| > \frac{\varepsilon}{2}\right) \\
&+ \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n \mathbb{E}_v[v[u^{(i)}]_j \mid u^{(i)}] - \mathbb{E}[v[u]_j]\right| > \frac{\varepsilon}{2}\right) \\
\overset{\xi_1}{\leq}& 4\exp\left\{-\frac{n\varepsilon^2}{8\sigma^2}\right\},
\end{aligned}$$

where "ξ_1" follows from the large deviation inequality of sub-Gaussian random variables which is standard (see, e.g., Vershynin, 2011). By the union bound we have

$$\mathbb{P}(\|\nabla l(\bar{w})\|_\infty > \varepsilon) \quad \leq \quad 4p \exp\left\{-\frac{n\varepsilon^2}{8\sigma^2}\right\}.$$

By choosing $\varepsilon = 4\sigma\sqrt{\ln p/n}$ in the above inequality we obtain that with probability at least $1 - 4p^{-1}$,

$$\|\nabla l(\bar{w})\|_\infty \leq 4\sigma\sqrt{\ln p/n}.$$

Combining the above bound with (30) yields the desired result. ■

D.2 The ADMM Method for Solving the Subproblem (13)

Now we present the algorithmic procedure of ADMM for solving the subproblem (13). By introducing an auxiliary variable $\Theta \in \mathbb{R}^{p\times p}$, this subproblem can be equivalently formulated as

$$\min_{\alpha I \preceq \Omega \preceq \beta I} L(\Omega), \quad \text{s.t. } \Omega = \Theta, \ \text{supp}(\Theta) \subseteq F. \tag{31}$$

Then, the augmented Lagrangian function of (31) is

$$J(\Omega, \Theta, \Gamma) := L(\Omega) - \langle \Gamma, \Omega - \Theta\rangle + \frac{\rho}{2}\|\Omega - \Theta\|_{Frob}^2,$$

where $\Gamma \in \mathbb{R}^{p\times p}$ is the multiplier of the linear constraint $\Omega = \Theta$ and $\rho > 0$ is the penalty strength parameter for the violation of the linear constraint. The ADMM method alternately solves the following problems to generate the new iterate:

$$\begin{aligned}
\Omega^{(\tau)} &= \mathop{\arg\min}_{\alpha I \preceq \Omega \preceq \beta I} J(\Omega, \Theta^{(\tau-1)}, \Gamma^{(\tau-1)}), &(32)\\
\Theta^{(\tau)} &= \mathop{\arg\min}_{\text{supp}(\Theta)\subseteq F} J(\Omega^{(\tau)}, \Theta, \Gamma^{(\tau-1)}), &(33)\\
\Gamma^{(\tau)} &= \Gamma^{(\tau-1)} - \rho(\Omega^{(\tau)} - \Theta^{(\tau)}).
\end{aligned}$$

Let us first consider the minimization problem (32) for updating $\Omega^{(\tau)}$. It is equivalent to the following minimization problem:

$$\Omega^{(\tau)} = \mathop{\arg\min}_{\alpha I \preceq \Omega \preceq \beta I} \frac{1}{2}\|\Omega - M\|_{Frob}^2 - \frac{1}{\rho}\log\det\Omega,$$

where

$$M = \Theta^{(\tau-1)} - \frac{1}{\rho}(\Sigma_n - \Gamma^{(\tau-1)}).$$

Let the eigenvalue decomposition of M be

$$M = V\Lambda V^\top, \quad \text{with } \Lambda = \text{diag}(\lambda_1, ..., \lambda_n).$$

It is easy to verify that the solution of problem (32) is given by

$$\Omega^{(\tau)} = V\tilde{\Lambda}V^\top, \quad \text{with } \tilde{\Lambda} = \text{diag}(\tilde{\lambda}_1, ..., \tilde{\lambda}_n),$$

where

$$\tilde{\lambda}_j = \min\left\{\beta, \max\left\{\alpha, \frac{\lambda_j + \sqrt{\lambda_j^2 + 4/\rho}}{2}\right\}\right\}.$$

Next, we consider the minimization problem (33) for updating $\Theta^{(\tau)}$. It is straightforward to see that the solution of problem (33) is given by

$$\Theta^{(\tau)} = \left[\Omega^{(\tau)} - \frac{1}{\rho}\Gamma^{(\tau-1)}\right]_F.$$

References

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

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

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

Beck, A. and Teboulle, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

Bishop, C. M. *Pattern Recognition and Machine Learning*. Springer-Verlag New York, 2006.

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

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

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2010.

Cai, T., Liu, W., and Luo, X. A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607, 2011.

Candès, E. J., Romberg, J. K., and Tao, T. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59(8): 1207–1223, 2006.

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

Edwards, D. M. *Introduction to Graphical Modelling*. Springer Science & Business Media New York, 2000.

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

Foucart, S. Sparse recovery algorithms: sufficient conditions in terms of restricted isometry constants. In *Approximation Theory XIII: San Antonio 2010*, pages 65–77, 2012.

Frank, M. and Wolfe, P. An algorithm for quadratic programming. *Naval Research Logistics (NRL)*, 3(1-2):95–110, 1956.

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

Garg, R. and Khandekar, R. Gradient descent with sparsification: an iterative algorithm for sparse recovery with restricted isometry property. In *International Conference on Machine Learning (ICML)*, pages 337–344, 2009.

Guyon, I., Gunn, S., Hur A. B., and Dror, G. Result analysis of the NIPS 2003 feature selection challenge. In *Advances in Neural Information Processing Systems (NIPS)*, pages 545–552, 2005.

Hess, K. R., Anderson, K., Symmans, W. F., and *et al.* Pharmacogenomic predictor of snesitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. *Journal of Clinical Oncology*, 24(26):4236–4244, 2006.

Jaggi, M. Sparse convex optimization methods for machine learning. Technical report, PhD thesis in Theoretical Computer Science, ETH Zurich, 2011.

Jain, P., Rao, N., and Dhillon, I. Structured sparse regression via greedy hard-thresholding. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1516–1524, 2016.

Jain, P., Tewari, A., and Kar, P. On iterative hard thresholding methods for high-dimensional m-estimation. In *Advances in Neural Information Processing Systems (NIPS)*, pages 685–693, 2014.

Jalali, A., Johnson, C. C., and Ravikumar, P. K. On learning discrete graphical models using greedy methods. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1935–1943, 2011.

Keerthi, S. S. and DeCoste, D. A modified finite newton method for fast solution of large scale linear svms. *Journal of Machine Learning Research*, 6:341–361, 2005.

Kim, Y. and Kim, J. Gradient lasso for feature selection. In *International Conference on Machine Learning (ICML)*, pages 60–67, 2004.

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

Lewis, D., Yang, Y., Rose, T., and Li, F. Rcv1: a new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397, 2004.

Li, X., Zhao, T., Arora, R., Liu, H., and Haupt, J. Stochastic variance reduced optimization for nonconvex sparse learning. In *International Conference on Machine Learning (ICML)*, pages 917–925, 2016.

Li, Y.-H., Scarlett, J., Ravikumar, P., and Cevher, V. Sparsistency of ℓ_1-regularized m-estimators. In *International Conference on Artifficial Intelligence and Statistics (AISTATS)*, pages 644–652, 2015.

Liu, B., Yuan, X.-T., Wang, L., Liu, Q., and Metaxas, D. N. Dual Iterative Hard Thresholding: From Non-convex Sparse Minimization to Non-smooth Concave Maximization. In *International Conference on Machine Learning (ICML)*, pages 2179–2187, 2017.

Lu, Z. Smooth optimization approach for sparse covariance selection. *SIAM Journal on Optimization*, 19(4):1807–1827, 2009.

Mallat, S. G. and Zhang, Z. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.

Natarajan, B. K. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24(2):227–234, 1995.

Needell, D. and Tropp, J. A. Cosamp: iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis*, 26(3):301–321, 2009.

Nesterov, Y. *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer Academic Publishers, 2004.

Nguyen, N., Needell, D., and Woolf, T. Linear convergence of stochastic iterative greedy algorithms with sparse constraints. *IEEE Transactions on Information Theory*, 63(11): 6869–6895, 2017.

Pati, Y. C., Rezaiifar, R., and Krishnaprasad, P. S. Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition. In *Annual Asilomar Conference on Signals, Systems, and Computers*, pages 40–44, 1993.

Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

Shalev-Shwartz, S., Srebro, N., and Zhang, T. Trading accuracy for sparsity in optimization problems with sparsity constraints. *SIAM Journal on Optimization*, 20(6):2807–2832, 2010.

Shen, J. and Li, P. A Tight Bound of Hard Thresholding. *arXiv preprint arXiv:1605.01656*, 2016. URL `http://arxiv.org/pdf/1605.01656.pdf`.

Tewari, A., Ravikumar, P., and Dhillon, I. S. Greedy algorithms for structurally constrained high-dimensional problems. In *Advances in Neural Information Processing Systems (NIPS)*, pages 882–890, 2011.

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

Van De Geer, S. and Bühlmann, P. On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

Vershynin, R. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2011. URL `http://arxiv.org/pdf/1011.3027.pdf`.

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

Wainwright, M. J. and Jordan, M.I. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.

Yuan, X. M. Alternating direction method for covariance selection models. *Journal of Scientific Computing*, 51(2):261–273, 2012.

Yuan, X.-T., Li, P., and Zhang, T. Gradient hard thresholding pursuit for sparsity-constrained optimization. In *International Conference on Machine Learning (ICML)*, pages 127–135, 2014.

Yuan, X.-T., Li, P., and Zhang, T. Exact recovery of hard thresholding pursuit. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3558–3566, 2016.

Yuan, X.-T. and Yan, S. Forward basis selection for pursuing sparse representations over a dictionary. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12): 3025–3036, 2013.

Yuan, X.-T. and Zhang, T. Truncated power method for sparse eigenvalue problems. *Journal of Machine Learning Research*, 14:899–925, 2013.

Zhang, T. Adaptive forward-backward greedy algorithm for sparse learning with linear models. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1921–1928, 2008.