Journal of Machine Learning Research 14 (2013) 807-841 Submitted 3/12; Revised 10/12; Published 3/13

Greedy Sparsity-Constrained Optimization

Sohail Bahmani SBAHMANI@CMU.EDU
Department of Electrical and Computer Engineering
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213, USA

Bhiksha Raj BHIKSHA@CS.CMU.EDU
Language Technologies Institute
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213, USA

Petros T. Boufounos PETROSB@MERL.COM
Mitsubishi Electric Research Laboratories
201 Broadway
Boston, MA 02139, USA

Editor: Francis Bach

Abstract

Sparsity-constrained optimization has wide applicability in machine learning, statistics, and signal processing problems such as feature selection and Compressed Sensing. A vast body of work has studied the sparsity-constrained optimization from theoretical, algorithmic, and application aspects in the context of sparse estimation in linear models where the fidelity of the estimate is measured by the squared error. In contrast, relatively less effort has been made in the study of sparsity-constrained optimization in cases where nonlinear models are involved or the cost function is not quadratic. In this paper we propose a greedy algorithm, Gradient Support Pursuit (GraSP), to approximate sparse minima of cost functions of arbitrary form. Should a cost function have a Stable Restricted Hessian (SRH) or a Stable Restricted Linearization (SRL), both of which are introduced in this paper, our algorithm is guaranteed to produce a sparse vector within a bounded distance from the true sparse optimum. Our approach generalizes known results for quadratic cost functions that arise in sparse linear regression and Compressed Sensing. We also evaluate the performance of GraSP through numerical simulations on synthetic and real data, where the algorithm is employed for sparse logistic regression with and without ℓ_2-regularization.

Keywords: sparsity, optimization, compressed sensing, greedy algorithm

1. Introduction

The demand for high-dimensional data analysis has grown significantly over the past decade by the emergence of applications such as social networking, bioinformatics, and mathematical finance. In these applications data samples often have thousands of features using which an underlying parameter must be inferred or predicted. In many circumstances the number of collected samples is significantly smaller than the dimensionality of the data, rendering any inference from the data ill-posed. However, it is widely acknowledged that the data sets that need to be processed usually

©2013 Sohail Bahmani, Bhiksha Raj and Petros T. Boufounos.

exhibit significant structure, which sparsity models are often able to capture. This structure can be exploited for robust regression and hypothesis testing, model reduction and variable selection, and more efficient signal acquisition in *underdetermined* regimes. Estimation of parameters with sparse structure is usually cast as an optimization problem, formulated according to specific application requirements. Developing techniques that are robust and computationally tractable to solve these optimization problems, even only approximately, is therefore critical.

In particular, theoretical and application aspects of sparse estimation in linear models have been studied extensively in areas such as signal processing, machine learning, and statistics. However, sparse estimation in problems where nonlinear models are involved have received comparatively little attention. Most of the work in this area extend the use of the ℓ_1-norm as a regularizer, effective to induce sparse solutions in linear regression, to problems with nonlinear models (see, e.g., Bunea, 2008; van de Geer, 2008; Kakade et al., 2010; Negahban et al., 2009). As a special case, logistic regression with ℓ_1 and elastic net regularization are studied by Bunea (2008). Furthermore, Kakade et al. (2010) have studied the accuracy of sparse estimation through ℓ_1-regularization for the exponential family distributions. A more general frame of study is proposed and analyzed by Negahban et al. (2009) where regularization with "decomposable" norms is considered in M-estimation problems. To provide the accuracy guarantees, these works generalize the Restricted Eigenvalue condition (Bickel et al., 2009) to ensure that the loss function is strongly convex over a restriction of its domain. We would like to emphasize that these sufficient conditions generally hold with proper constants and with high probability only if one assumes that the true parameter is bounded. This fact is more apparent in some of the mentioned work (e.g., Bunea, 2008; Kakade et al., 2010), while in some others (e.g., Negahban et al., 2009) the assumption is not explicitly stated. We will elaborate on this matter in Section 2. Tewari et al. (2011) also proposed a coordinate-descent type algorithm for minimization of a convex and smooth objective over the convex signal/parameter models introduced in Chandrasekaran et al. (2012). This formulation includes the ℓ_1-constrained minimization as a special case, and the algorithm is shown to converge to the minimum in objective value similar to the standard results in convex optimization.

Furthermore, Shalev-Shwartz et al. (2010) proposed a number of greedy that sparsify a given estimate at the cost of relatively small increase of the objective function. However, their algorithms are not stand-alone. A generalization of Compressed Sensing is also proposed in Blumensath (2010), where the linear measurement operator is replaced by a nonlinear operator that applies to the sparse signal. Considering the norm of the residual error as the objective, Blumensath (2010) shows that if the objective satisfies certain sufficient conditions, the sparse signal can be accurately estimated by a generalization of the Iterative Hard Thresholding algorithm (Blumensath and Davies, 2009). The formulation of Blumensath (2010), however, has a limited scope because the metric of error is defined using a norm. For instance, the formulation does not apply to objectives such as the logistic loss. More recently, Jalali et al. (2011) studied a forward-backward algorithm using a variant of the sufficient conditions introduced in Negahban et al. (2009). Similar to our work, the main result in Jalali et al. (2011) imposes conditions on the function as restricted to sparse inputs whose non-zeros are fewer than a multiple of the target sparsity level. The multiplier used in their results has an *objective-dependent* value and is never less than 10. Furthermore, the multiplier is important in their analysis not only for determining the stopping condition of the algorithm, but also in the lower bound assumed for the minimal magnitude of the non-zero entries. In contrast, the multiplier in our results is fixed at 4, independent of the objective function itself, and we make no assumptions about the magnitudes of the non-zero entries.

This paper presents an extended version with improved guarantees of our prior work in Bahmani et al. (2011), where we proposed a greedy algorithm, the Gradient Support Pursuit (GraSP), for sparse estimation problems that arise in applications with general nonlinear models. We prove the accuracy of GraSP for a class of cost functions that have a *Stable Restricted Hessian* (SRH). The SRH, introduced in Bahmani et al. (2011), characterizes the functions whose restriction to sparse canonical subspaces have well-conditioned Hessian matrices. Similarly, we analyze the GraSP algorithm for non-smooth functions that have a *Stable Restricted Linearization* (SRL), a property introduced in this paper, analogous to SRH. The analysis and the guarantees for smooth and non-smooth cost functions are similar, except for less stringent conditions derived for smooth cost functions due to properties of symmetric Hessian matrices. We also prove that the SRH holds for the case of the ℓ_2-penalized logistic loss function.

1.1 Notation

In the remainder of this paper we use the notation listed in Table 1.

1.2 Paper Outline

In Section 2 we provide a background on sparse parameter estimation which serves as an overview of prior work. In Section 3 we state the general formulation of the problem and present our algorithm. Conditions that characterize the cost functions and the main accuracy guarantees of our algorithm are provided in Section 3 as well. The guarantees of the algorithm are proved in Appendices A and B. As an example where our algorithm can be applied, ℓ_2-regularized logistic regression is studied in Section 4. Some experimental results for logistic regression with sparsity constraints are presented in Section 5. Finally, Section 6 discusses the results and concludes.

2. Background

We first briefly review sparse estimation problems studied in the literature.

2.1 Sparse Linear Regression and Compressed Sensing

The special case of sparse estimation in linear models has gained significant attention under the title of Compressed Sensing (CS) (Donoho, 2006). In standard CS problems the aim is to estimate a sparse vector $\mathbf{x}^\star$ from noisy linear measurements $\mathbf{y} = \mathbf{A}\mathbf{x}^\star + \mathbf{e}$, where $\mathbf{A}$ is a known $n \times p$ measurement matrix with $n \ll p$ and $\mathbf{e}$ is the additive measurement noise. To find the sparsest estimate in this *underdetermined* problem that is consistent with the measurements $\mathbf{y}$ one needs to solve the optimization problem

$$\widehat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\mathbf{x}\|_0 \quad \text{s.t.} \;\; \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2 \leq \varepsilon, \tag{1}$$

where ε is a given upper bound for $\|\mathbf{e}\|_2$ (Candès et al., 2006). In the absence of noise (i.e., when $\varepsilon = 0$), if $\mathbf{x}^\star$ is s-sparse (i.e., it has at most s nonzero entries) one merely needs every $2s$ columns of $\mathbf{A}$ to be linearly independent to guarantee exact recovery (Donoho and Elad, 2003). Unfortunately, the ideal solver (1) is computationally NP-hard in general (Natarajan, 1995) and one must seek approximate solvers instead.

It is shown in Candès et al. (2006) that under certain conditions, minimizing the ℓ_1-norm as a convex proxy for the ℓ_0-norm yields accurate estimates of $\mathbf{x}^\star$. The resulting approximate solver

Symbol	**Description**
$[n]$	the set $\{1,2,\ldots,n\}$ for any $n \in \mathbb{N}$
$\mathcal{I}$	calligraphic letters denote sets unless stated otherwise (e.g., $\mathcal{N}\left(\mu,\sigma^2\right)$ denotes a normal distribution)
$\mathcal{I}^c$	complement of set $\mathcal{I}$
$\mathbf{v}$	bold face small letters denote column vectors in $\mathbb{R}^b$ for some $b \in \mathbb{N}$
$\lVert\mathbf{v}\rVert_q$	the ℓ_q-norm of vector $\mathbf{v}$, that is $\left(\sum_{i=1}^{b} \lvert v_i\rvert^q\right)^{1/q}$, for a real number $q \geq 1$
$\lVert\mathbf{v}\rVert_0$	the "ℓ_0-norm" of vector $\mathbf{v}$ that merely counts its nonzero entries
$\mathbf{v}\vert_{\mathcal{I}}$	depending on the context 1. restriction of vector $\mathbf{v}$ to the rows indicated by indices in $\mathcal{I}$, or 2. a vector that equals $\mathbf{v}$ except for coordinates in $\mathcal{I}^c$ where it is zero
$\mathbf{v}_r$	the best r-term approximation of vector $\mathbf{v}$
$\operatorname{supp}(\mathbf{v})$	the support set (i.e., indices of the non-zero entries) of $\mathbf{v}$
$\mathbf{M}$	bold face capital letters denote matrices in $\mathbb{R}^{a\times b}$ for some $a,b \in \mathbb{N}$
$\mathbf{M}^{\mathrm{T}}$	transpose of matrix $\mathbf{M}$
$\mathbf{M}^{\dagger}$	pseudo-inverse of matrix $\mathbf{M}$
$\mathbf{M}_{\mathcal{I}}$	restriction of matrix $\mathbf{M}$ to the columns enumerated by $\mathcal{I}$
$\lVert\mathbf{M}\rVert$	the operator norm of matrix $\mathbf{M}$ which is equal to $\sqrt{\lambda_{\max}\left(\mathbf{M}^{\mathrm{T}}\mathbf{M}\right)}$
$\mathbf{I}$	the identity matrix
$\mathbf{P}_{\mathcal{I}}$	restriction of the identity matrix to the columns indicated by $\mathcal{I}$
$\mathbf{1}$	column vector of all ones
$\mathbb{E}\left[\cdot\right]$	expectation
$\mathbf{H}_f\left(\cdot\right)$	Hessian of the function f

Table 1: Notation used in this paper

basically returns the solution to the convex optimization problem

$$\widehat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\mathbf{x}\|_1 \quad \text{s.t. } \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2 \leq \varepsilon, \tag{2}$$

The required conditions for approximate equivalence of (1) and (2), however, generally hold only if measurements are collected at a higher rate. Ideally, one merely needs $n = O(s)$ measurements to estimate $\mathbf{x}^\star$, but $n = O(s \log \frac{p}{s})$ measurements are necessary for the accuracy of (2) to be guaranteed.

The convex program (2) can be solved in polynomial time using interior point methods. However, these methods do not scale well as the size of the problem grows. Therefore, several first-order convex optimization methods are developed and analyzed as more efficient alternatives (see, e.g., Beck and Teboulle, 2009; Agarwal et al., 2010). Another category of low-complexity algorithms in CS are the non-convex *greedy pursuits* including Orthogonal Matching Pursuit (OMP) (Pati et al., 1993; Tropp and Gilbert, 2007), Compressive Sampling Matching Pursuit (CoSaMP) (Needell and Tropp, 2009), Iterative Hard Thresholding (IHT) (Blumensath and Davies, 2009), and Subspace Pursuit (Dai and Milenkovic, 2009) to name a few. These greedy algorithms implicitly approximate the solution to the ℓ_0-constrained least squares problem

$$\widehat{\mathbf{x}} = \arg\min_{\mathbf{x}} \frac{1}{2} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 \quad \text{s.t. } \|\mathbf{x}\|_0 \leq s. \tag{3}$$

The main theme of these iterative algorithms is to use the residual error from the previous iteration to successively approximate the position of non-zero entries and estimate their values. These algorithms have shown to exhibit accuracy guarantees similar to those of convex optimization methods, though with more stringent requirements.

As mentioned above, to guarantee accuracy of the CS algorithms the measurement matrix should meet certain conditions such as *incoherence* (Donoho and Huo, 2001), Restricted Isometry Property (RIP) (Candès et al., 2006), Nullspace Property (Cohen et al., 2009), etc. Among these conditions RIP is the most commonly used and the best understood condition.

Matrix $\mathbf{A}$ is said to satisfy the RIP of order k—in its symmetric form—with constant δ_k, if $\delta_k < 1$ is the smallest number that

$$(1 - \delta_k) \|\mathbf{x}\|_2^2 \leq \|\mathbf{A}\mathbf{x}\|_2^2 \leq (1 + \delta_k) \|\mathbf{x}\|_2^2$$

holds for all k-sparse vectors $\mathbf{x}$. Several CS algorithms are shown to produce accurate solutions provided that the measurement matrix has a sufficiently small RIP constant of order ck with c being a small integer. For example, solving (2) is guaranteed to yield an accurate estimate of s-sparse $\mathbf{x}^\star$ if $\delta_{2s} < \sqrt{2} - 1$ (Candès, 2008). Interested readers can find the best known RIP-based accuracy guarantees for some of the CS algorithms in Foucart (2012).

2.2 Beyond Linear Models

The CS reconstruction algorithms attempt to provide a sparse vector that incurs only a small squared error which measures consistency of the solution versus the acquired data. While this measure of discrepancy is often desirable for signal processing applications, it is not the appropriate choice for a variety of other applications. For example, in statistics and machine learning the logistic loss function is also commonly used in regression and classification problems (see Liu et al., 2009, and references therein). Thus, it is desirable to develop theory and algorithms that apply to a broader class of optimization problems with sparsity constraints.

The existing studies on this subject are mostly in the context of statistical estimation. The majority of these studies consider the cost function to be convex everywhere and rely on the ℓ_1-regularization as the means to induce sparsity in the solution. For example, Kakade et al. (2010) have shown that for the exponential family of distributions maximum likelihood estimation with ℓ_1-regularization yields accurate estimates of the underlying sparse parameter. Furthermore, Negahban et al. have developed a unifying framework for analyzing statistical accuracy of *M-estimators* regularized by "decomposable" norms in (Negahban et al., 2009). In particular, in their work ℓ_1-regularization is applied to Generalized Linear Models (GLM) (Dobson and Barnett, 2008) and shown to guarantee a bounded distance between the estimate and the true statistical parameter. To establish this error bound they introduced the notion of *Restricted Strong Convexity* (RSC), which basically requires a lower bound on the curvature of the cost function around the true parameter in a restricted set of directions. The achieved error bound in this framework is inversely proportional to this curvature bound. Furthermore, Agarwal et al. (2010) have studied Projected Gradient Descent as a method to solve ℓ_1-constrained optimization problems and established accuracy guarantees using a slightly different notion of RSC and *Restricted Smoothness* (RSM).

Note that the guarantees provided for majority of the ℓ_1-regularization algorithms presume that the true parameter is bounded, albeit implicitly. For instance, the error bound for ℓ_1-regularized logistic regression is recognized by Bunea (2008) to be dependent on the true parameter (Bunea, 2008, Assumption A, Theorem 2.4, and the remark that succeeds them). Moreover, the result proposed by Kakade et al. (2010) implicitly requires the true parameter to have a sufficiently short length to allow the choice of the desirable regularization coefficient (Kakade et al., 2010, Theorems 4.2 and 4.5). Negahban et al. (2009) also assume that the true parameter is inside the unit ball to establish the required condition for their analysis of ℓ_1-regularized GLM, although this restriction is not explicitly stated (see the longer version of Negahban et al., 2009, p. 37). We can better understand why restricting the length of the true parameter may generally be inevitable by viewing these estimation problems from the perspective of empirical processes and their convergence. The empirical processes, including those considered in the studies mentioned above, are generally good approximations of their corresponding expected process (see Vapnik, 1998, chap. 5 and van de Geer, 2000). Therefore, if the expected process is not strongly convex over an unbounded, but perhaps otherwise restricted, set the corresponding empirical process cannot be strongly convex over the same set. This reasoning applies in many cases including the studies mentioned above, where it would be impossible to achieve the desired restricted strong convexity properties—with high probability—if the true parameter is allowed to be unbounded.

Furthermore, the methods that rely on the ℓ_1-norm are known to result in sparse solutions, but, as mentioned in Kakade et al. (2010), the sparsity of these solutions is not known to be optimal in general. One can intuit this fact from definitions of RSC and RSM. These two properties bound the curvature of the function from below and above in a restricted set of directions around the true optimum. For quadratic cost functions, such as squared error, these curvature bounds are absolute constants. As stated before, for more general cost functions such as the loss functions in GLMs, however, these constants will depend on the location of the true optimum. Consequently, depending on the location of the true optimum these error bounds could be extremely large, albeit finite. When error bounds are significantly large, the sparsity of the solution obtained by ℓ_1-regularization may not be satisfactory. This motivates investigation of algorithms that do not rely on ℓ_1-norm to induce sparsity.

3. Problem Formulation and the GraSP Algorithm

As seen in Section 2.1, in standard CS the squared error $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2$ is used to measure fidelity of the estimate. While this is appropriate for a large number of signal acquisition applications, it is not the right cost in other fields. Thus, the significant advances in CS cannot readily be applied in these fields when estimation or prediction of sparse parameters become necessary. In this paper we focus on a generalization of (3) where a generic cost function replaces the squared error. Specifically, for the cost function $f : \mathbb{R}^p \mapsto \mathbb{R}$, it is desirable to approximate

$$\arg\min_{\mathbf{x}} f(\mathbf{x}) \quad \text{s.t.}\ \|\mathbf{x}\|_0 \leq s. \tag{4}$$

We propose the Gradient Support Pursuit (GraSP) algorithm, which is inspired by and generalizes the CoSaMP algorithm, to approximate the solution to (4) for a broader class of cost functions.

Of course, even for a simple quadratic objective, (4) can have combinatorial complexity and become NP-hard. However, similar to the results of CS, knowing that the cost function obeys certain properties allows us to obtain accurate estimates through tractable algorithms. To guarantee that GraSP yields accurate solutions and is a tractable algorithm, we also require the cost function to have certain properties that will be described in Section 3.2. These properties are analogous to and generalize the RIP in the standard CS framework. For smooth cost functions we introduce the notion of a Stable Restricted Hessian (SRH) and for non-smooth cost functions we introduce the Stable Restricted Linearization (SRL). Both of these properties basically bound the Bregman divergence of the cost function restricted to sparse canonical subspaces. However, the analysis based on the SRH is facilitated by matrix algebra that results in somewhat less restrictive requirements for the cost function.

3.1 Algorithm Description

Algorithm 1: The GraSP algorithm

```
input  : $f(\cdot)$ and $s$
output: $\hat{\mathbf{x}}$
initialize: $\widehat{\mathbf{x}} = 0$
repeat
    compute local gradient: $\mathbf{z} = \nabla f(\widehat{\mathbf{x}})$
    identify directions: $\mathcal{Z} = \operatorname{supp}(\mathbf{z}_{2s})$
    merge supports: $\mathcal{T} = \mathcal{Z} \cup \operatorname{supp}(\widehat{\mathbf{x}})$
    minimize over support: $\mathbf{b} = \arg\min f(\mathbf{x})$ s.t. $\mathbf{x}|_{\mathcal{T}^c} = \mathbf{0}$
    prune estimate: $\widehat{\mathbf{x}} = \mathbf{b}_s$
until halting condition holds
```

GraSP is an iterative algorithm, summarized in Algorithm 1, that maintains and updates an estimate $\widehat{\mathbf{x}}$ of the sparse optimum at every iteration. The first step in each iteration, $\mathbf{z} = \nabla f(\widehat{\mathbf{x}})$, evaluates the gradient of the cost function at the current estimate. For nonsmooth functions, instead of the gradient we use a *restricted subgradient* $\mathbf{z} = \nabla_f(\widehat{\mathbf{x}})$ defined in Section 3.2. Then $2s$ coordinates of the vector $\mathbf{z}$ that have the largest magnitude are chosen as the directions in which pursuing the minimization will be most effective. Their indices, denoted by $\mathcal{Z} = \operatorname{supp}(\mathbf{z}_{2s})$, are then merged with the support of the current estimate to obtain $\mathcal{T} = \mathcal{Z} \cup \operatorname{supp}(\widehat{\mathbf{x}})$. The combined support is a set of at most $3s$ indices over which the function f is minimized to produce an intermediate estimate $\mathbf{b} = \arg\min f(\mathbf{x})$ s.t. $\mathbf{x}|_{\mathcal{T}^c} = 0$. The estimate $\widehat{\mathbf{x}}$ is then updated as the best s-term approximation of the intermediate estimate $\mathbf{b}$. The iterations terminate once certain condition, for instance, on the change of the cost function or the change of the estimated minimum from the previous iteration, holds.

In the special case where the squared error $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2$ is the cost function, GraSP reduces to CoSaMP. Specifically, the gradient step reduces to the proxy step $\mathbf{z} = \mathbf{A}^{\mathrm{T}}(\mathbf{y} - \mathbf{A}\hat{\mathbf{x}})$ and minimization over the restricted support reduces to the constrained pseudoinverse step $\mathbf{b}|_{\mathcal{T}} = \mathbf{A}_{\mathcal{T}}^{\dagger}\mathbf{y}$, $\mathbf{b}|_{\mathcal{T}^c} = \mathbf{0}$ in CoSaMP.

Variants Although in this paper we only analyze the standard form of GraSP outlined in Algorithm 1, other variants of the algorithm can also be studied. Below we list some of these variants.

1. *Debiasing*: In this variant, instead of performing a hard thresholding on the vector $\mathbf{b}$, the objective is minimized restricted to the support set of $\mathbf{b}_s$ to obtain the new iterate:

$$\widehat{\mathbf{x}} = \arg\min_{\mathbf{x}} \; f(\mathbf{x}) \quad \text{s.t. } \operatorname{supp}(\mathbf{x}) \subseteq \operatorname{supp}(\mathbf{b}_s).$$

2. *Restricted Newton Step*: To reduce the computations in each iteration, the minimization that yields $\mathbf{b}$, we can set $\mathbf{b}|_{\mathcal{T}^c} = \mathbf{0}$ and take a restricted Newton step as

$$\mathbf{b}|_{\mathcal{T}} = \widehat{\mathbf{x}}|_{\mathcal{T}} - \kappa\left(\mathbf{P}_{\mathcal{T}}^{\mathrm{T}}\mathbf{H}_f(\widehat{\mathbf{x}})\mathbf{P}_{\mathcal{T}}\right)^{-1}\widehat{\mathbf{x}}|_{\mathcal{T}},$$

where $\kappa > 0$ is a step-size. Of course, here we are assuming that the restricted Hessian, $\mathbf{P}_{\mathcal{T}}^{\mathrm{T}}\mathbf{H}_f(\widehat{\mathbf{x}})\mathbf{P}_{\mathcal{T}}$, is invertible.

3. *Restricted Gradient Descent*: The minimization step can be relaxed even further by applying a restricted gradient descent. In this approach, we again set $\mathbf{b}|_{\mathcal{T}^c} = \mathbf{0}$ and

$$\mathbf{b}|_{\mathcal{T}} = \widehat{\mathbf{x}}|_{\mathcal{T}} - \kappa\,\nabla f(\widehat{\mathbf{x}})|_{\mathcal{T}}.$$

Since $\mathcal{T}$ contains both the support set of $\widehat{\mathbf{x}}$ and the $2s$-largest entries of $\nabla f(\widehat{\mathbf{x}})$, it is easy to show that each iteration of this alternative method is equivalent to a standard gradient descent followed by a hard thresholding. In particular, if the squared error is the cost function as in standard CS, this variant reduces to the IHT algorithm.

3.2 Sparse Reconstruction Conditions

In what follows we characterize the functions for which accuracy of GraSP can be guaranteed. For twice continuously differentiable functions we rely on Stable Restricted Hessian (SRH), while for non-smooth cost functions we introduce the Stable Restricted Linearization (SRL). These properties that are analogous to the RIP in the standard CS framework, basically require that the curvature of the cost function over the sparse subspaces can be bounded locally from above and below such that the corresponding bounds have the same order. Below we provide precise definitions of these two properties.

Definition 1 (Stable Restricted Hessian)**.** Suppose that f is a twice continuously differentiable function whose Hessian is denoted by $\mathbf{H}_f(\cdot)$. Furthermore, let

$$A_k(\mathbf{x}) = \sup\left\{\Delta^{\mathrm{T}}\mathbf{H}_f(\mathbf{x})\Delta \,\middle|\, |\mathrm{supp}(\mathbf{x}) \cup \mathrm{supp}(\Delta)| \leq k, \|\Delta\|_2 = 1\right\} \tag{5}$$

and

$$B_k(\mathbf{x}) = \inf\left\{\Delta^{\mathrm{T}}\mathbf{H}_f(\mathbf{x})\Delta \,\middle|\, |\mathrm{supp}(\mathbf{x}) \cup \mathrm{supp}(\Delta)| \leq k, \|\Delta\|_2 = 1\right\}, \tag{6}$$

for all k-sparse vectors $\mathbf{x}$. Then f is said to have a Stable Restricted Hessian (SRH) with constant μ_k, or in short μ_k-SRH, if $1 \leq \frac{A_k(\mathbf{x})}{B_k(\mathbf{x})} \leq \mu_k$.

Remark 1. Since the Hessian of f is symmetric, an equivalent for Definition 1 is that a twice continuously differentiable function f has μ_k-SRH if the condition number of $\mathbf{P}_{\mathcal{K}}\mathbf{H}_f(\mathbf{x})\mathbf{P}_{\mathcal{K}}^{\mathrm{T}}$ is not greater than μ_k for all k-sparse vectors $\mathbf{x}$ and sets $\mathcal{K} \subseteq [p]$ with $|\mathrm{supp}(\mathbf{x}) \cup \mathcal{K}| \leq k$.

In the special case when the cost function is the squared error as in (3), we can write $\mathbf{H}_f(\mathbf{x}) = \mathbf{A}^{\mathrm{T}}\mathbf{A}$ which is constant. The SRH condition then requires

$$B_k\|\Delta\|_2^2 \leq \|\mathbf{A}\Delta\|_2^2 \leq A_k\|\Delta\|_2^2$$

to hold for all k-sparse vectors Δ with $A_k/B_k \leq \mu_k$. Therefore, in this special case the SRH condition essentially becomes equivalent to the RIP condition.

Remark 2. Note that the functions that satisfy the SRH are convex over canonical sparse subspaces, but they are not necessarily convex everywhere. The following two examples describe some nonconvex functions that have SRH.

Example 1. Let $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x}$, where $\mathbf{Q} = 2 \times \mathbf{1}\mathbf{1}^{\mathrm{T}} - \mathbf{I}$. Obviously, we have $\mathbf{H}_f(\mathbf{x}) = \mathbf{Q}$. Therefore, (5) and (6) determine the extreme eigenvalues across all of the $k \times k$ symmetric submatrices of $\mathbf{Q}$. Note that the diagonal entries of $\mathbf{Q}$ are all equal to one, while its off-diagonal entries are all equal to two. Therefore, for any 1-sparse signal $\mathbf{u}$ we have $\mathbf{u}^{\mathrm{T}}\mathbf{Q}\mathbf{u} = \|\mathbf{u}\|_2^2$, meaning that f has μ_1-SRH with $\mu_1 = 1$. However, for $\mathbf{u} = [1, -1, 0, \ldots, 0]^{\mathrm{T}}$ we have $\mathbf{u}^{\mathrm{T}}\mathbf{Q}\mathrm{u} < 0$, which means that the Hessian of f is not positive semi-definite (i.e., f is not convex).

Example 2. Let $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_2^2 + Cx_1x_2\cdots x_{k+1}$ where the dimensionality of $\mathbf{x}$ is greater than k. It is obvious that this function is convex for k-sparse vectors as $x_1x_2\cdots x_{k+1} = 0$ for any k-sparse vector. So we can easily verify that f satisfies SRH of order k. However, for $x_1 = x_2 = \cdots = x_{k+1} = t$ and $x_i = 0$ for $i > k+1$ the restriction of the Hessian of f to indices in $[k+1]$ (i.e., $\mathbf{P}_{[k+1]}^{\mathrm{T}}\mathbf{H}_f(\mathbf{x})\mathbf{P}_{[k+1]}$) is a matrix with diagonal entries all equal to one and off-diagonal entries all equal to Ct^{k-1}. Let $\mathbf{Q}$ denote this matrix and $\mathbf{u}$ be a unit-norm vector such that $\langle\mathbf{u}, \mathbf{1}\rangle = 0$. Then it is straightforward to verify that $\mathbf{u}^{\mathrm{T}}\mathbf{Q}\mathbf{u} = 1 - Ct^{k-1}$, which can be negative for sufficiently large values of C and t. Therefore, the Hessian of f is not positive semi-definite everywhere, meaning that f is not convex.

To generalize the notion of SRH to the case of nonsmooth functions, first we define the *restricted subgradient* of a function.

Definition 2 (Restricted Subgradient)**.** We say vector $\nabla_f(\mathbf{x})$ is a restricted subgradient of $f : \mathbb{R}^p \mapsto \mathbb{R}$ at point $\mathbf{x}$ if

$$f(\mathbf{x}+\Delta) - f(\mathbf{x}) \geq \langle\nabla_f(\mathbf{x}), \Delta\rangle$$

holds for all k-sparse vectors Δ.

Remark 3. We introduced the notion of restricted subgradient so that the restrictions imposed on f are as minimal as we need. We acknowledge that the existence of restricted subgradients implies convexity in sparse directions, but it does not imply convexity everywhere.

Remark 4. Obviously, if the function f is convex everywhere, then any subgradient of f determines a restricted subgradient of f as well. In general one may need to invoke the axiom of choice to define the restricted subgradient.

Remark 5. We drop the sparsity level from the notation as it can be understood from the context.

With a slight abuse of terminology we call

$$\mathrm{B}_f\left(\mathbf{x}' \parallel \mathbf{x}\right) = f\left(\mathbf{x}'\right) - f\left(\mathbf{x}\right) - \left\langle \nabla_f\left(\mathbf{x}\right), \mathbf{x}' - \mathbf{x}\right\rangle$$

the restricted Bregman divergence of $f : \mathbb{R}^p \mapsto \mathbb{R}$ between points $\mathbf{x}$ and $\mathbf{x}'$ where $\nabla_f(\cdot)$ gives a restricted subgradient of $f(\cdot)$.

Definition 3 (Stable Restricted Linearization)**.** Let $\mathbf{x}$ be a k-sparse vector in $\mathbb{R}^p$. For function $f : \mathbb{R}^p \mapsto \mathbb{R}$ we define the functions

$$\alpha_k\left(\mathbf{x}\right) = \sup\left\{ \frac{1}{\|\Delta\|_2^2} \mathrm{B}_f\left(\mathbf{x} + \Delta \parallel \mathbf{x}\right) \mid \Delta \neq 0 \text{ and } \left|\mathrm{supp}\left(\mathbf{x}\right) \cup \mathrm{supp}\left(\Delta\right)\right| \leq k \right\}$$

and

$$\beta_k\left(\mathbf{x}\right) = \inf\left\{ \frac{1}{\|\Delta\|_2^2} \mathrm{B}_f\left(\mathbf{x} + \Delta \parallel \mathbf{x}\right) \mid \Delta \neq 0 \text{ and } \left|\mathrm{supp}\left(\mathbf{x}\right) \cup \mathrm{supp}\left(\Delta\right)\right| \leq k \right\}.$$

Then $f(\cdot)$ is said to have a Stable Restricted Linearization with constant μ_k, or μ_k-SRL, if $\frac{\alpha_k(\mathbf{x})}{\beta_k(\mathbf{x})} \leq \mu_k$ for all k-sparse vectors $\mathbf{x}$.

Remark 6. The SRH and SRL conditions are similar to various forms of the Restricted Strong Convexity (RSC) and Restricted Strong Smoothness (RSS) conditions (Negahban et al., 2009; Agarwal et al., 2010; Blumensath, 2010; Jalali et al., 2011; Zhang, 2011) in the sense that they all bound the curvature of the objective function over a restricted set. The SRL condition quantifies the curvature in terms of a (restricted) Bregman divergence similar to RSC and RSS. The quadratic form used in SRH can also be converted to the Bregman divergence form used in RSC and RSS and vice-versa using the mean-value theorem. However, compared to various forms of RSC and RSS conditions SRH and SRL have some important distinctions. The main difference is that the bounds in SRH and SRL conditions are not global constants; only their ratio is required to be bounded globally. Furthermore, unlike the SRH and SRL conditions the variants of RSC and RSS, that are used in convex relaxation methods, are required to hold over a set which is strictly larger than the set of canonical k-sparse vectors.

There is also a subtle but important difference regarding the points where the curvature is evaluated at. Since Negahban et al. (2009) analyze a convex program, rather than an iterative algorithm, they only needed to invoke the RSC and RSS at a neighborhood of the true parameter. In contrast, the other variants of RSC and RSS (see, e.g., Agarwal et al., 2010; Jalali et al., 2011), as well as our SRH and SRL conditions, require the curvature bounds to hold uniformly over a larger set of points, thereby they are more stringent.

3.3 Main Theorems

Now we can state our main results regarding approximation of

$$\mathbf{x}^{\star} = \arg\min\ f(\mathbf{x}) \text{ s.t. } \|\mathbf{x}\|_0 \leq s, \tag{7}$$

using the GraSP algorithm.

Theorem 1. *Suppose that f is a twice continuously differentiable function that has μ_{4s}-SRH with $\mu_{4s} \leq \frac{1+\sqrt{3}}{2}$. Furthermore, suppose that for some $\varepsilon > 0$ we have $\varepsilon \leq B_{4s}(\mathbf{x})$ for all $4s$-sparse vectors $\mathbf{x}$. Then $\widehat{\mathbf{x}}^{(i)}$, the estimate at the i-th iteration, satisfies*

$$\left\|\widehat{\mathbf{x}}^{(i)} - \mathbf{x}^{\star}\right\|_2 \leq 2^{-i}\|\mathbf{x}^{\star}\|_2 + \frac{6+2\sqrt{3}}{\varepsilon}\left\|\nabla f(\mathbf{x}^{\star})|_I\right\|_2,$$

where I is the position of the $3s$ largest entries of $\nabla f(\mathbf{x}^{\star})$ in magnitude.

Remark 7. Note that this result indicates that $\nabla f(\mathbf{x}^{\star})$ determines how accurate the estimate can be. In particular, if the sparse minimum $\mathbf{x}^{\star}$ is sufficiently close to an unconstrained minimum of f then the estimation error floor is negligible because $\nabla f(\mathbf{x}^{\star})$ has small magnitude. This result is analogous to accuracy guarantees for estimation from noisy measurements in CS (Candès et al., 2006; Needell and Tropp, 2009).

Remark 8. As the derivations required to prove Theorem 1 show, the provided accuracy guarantee holds for any s-sparse $\mathbf{x}^{\star}$, even if it does not obey (7). Obviously, for arbitrary choices of $\mathbf{x}^{\star}$, $\nabla f(\mathbf{x}^{\star})|_I$ may have a large norm that cannot be bounded properly which implies large errors. In statistical estimation problems, often the true parameter that describes the data is chosen as the target parameter $\mathbf{x}^{\star}$ rather than the minimizer of the average loss function as in (7). In these problems, the approximation error $\|\nabla f(\mathbf{x}^{\star})|_I\|_2$ has statistical interpretation and can determine the statistical precision of the problem. This property is easy to verify in linear regression problems. We will also show this for the logistic loss as an example in Section 4.

Nonsmooth cost functions should be treated differently, since we do not have the luxury of working with Hessian matrices for these type of functions. The following theorem provides guarantees that are similar to those of Theorem 1 for nonsmooth cost functions that satisfy the SRL condition.

Theorem 2. *Suppose that f is a function that is not necessarily smooth, but it satisfies μ_{4s}-SRL with $\mu_{4s} \leq \frac{3+\sqrt{3}}{4}$. Furthermore, suppose that for $\beta_{4s}(\cdot)$ in Definition 3 there exists some $\varepsilon > 0$ such that $\beta_{4s}(\mathbf{x}) \geq \varepsilon$ holds for all $4s$-sparse vectors $\mathbf{x}$. Then $\widehat{\mathbf{x}}^{(i)}$, the estimate at the i-th iteration, satisfies*

$$\left\|\widehat{\mathbf{x}}^{(i)} - \mathbf{x}^{\star}\right\|_2 \leq 2^{-i}\|\mathbf{x}^{\star}\|_2 + \frac{6+2\sqrt{3}}{\varepsilon}\left\|\nabla_f(\mathbf{x}^{\star})|_I\right\|_2,$$

where I is the position of the $3s$ largest entries of $\nabla_f(\mathbf{x}^{\star})$ in magnitude.

Remark 9. Should the SRH or SRL conditions hold for the objective function, it is straightforward to convert the *point accuracy* guarantees of Theorems 1 and 2, into accuracy guarantees in terms of the objective value. First we can use SRH or SRL to bound the Bregman divergence, or its restricted version defined above, for points $\widehat{\mathbf{x}}^{(i)}$ and $\mathbf{x}^{\star}$. Then we can obtain a bound for the accuracy of the objective value by invoking the results of the theorems. This indirect approach, however, might not lead to sharp bounds and thus we do not pursue the detailed analysis in this work.

4. Example: Sparse Minimization of ℓ_2-regularized Logistic Regression

One of the models widely used in machine learning and statistics is the logistic model. In this model the relation between the data, represented by a random vector $\mathbf{a} \in \mathbb{R}^p$, and its associated label, represented by a random binary variable $y \in \{0,1\}$, is determined by the conditional probability

$$\Pr\{y \mid \mathbf{a};\mathbf{x}\} = \frac{\exp\left(y\left\langle \mathbf{a},\mathbf{x}\right\rangle\right)}{1+\exp\left(\left\langle \mathbf{a},\mathbf{x}\right\rangle\right)}, \tag{8}$$

where $\mathbf{x}$ denotes a parameter vector. Then, for a set of n independently drawn data samples $\{(\mathbf{a}_i, y_i)\}_{i=1}^n$ the joint likelihood can be written as a function of $\mathbf{x}$. To find the maximum likelihood estimate one should maximize this likelihood function, or equivalently minimize the negative log-likelihood, the logistic loss,

$$g(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \log\left(1+\exp\left(\left\langle \mathbf{a}_i,\mathbf{x}\right\rangle\right)\right) - y_i\left\langle \mathbf{a}_i,\mathbf{x}\right\rangle.$$

It is well-known that $g(\cdot)$ is strictly convex for $p \leq n$ provided that the associated design matrix, $\mathbf{A} = [\mathbf{a}_1\,\mathbf{a}_2\,\ldots\,\mathbf{a}_n]^{\mathrm{T}}$, is full-rank. However, in many important applications (e.g., feature selection) the problem can be underdetermined (i.e., $n < p$). In these scenarios the logistic loss is merely convex and it does not have a unique minimum. Furthermore, it is possible, especially in underdetermined problems, that the observed data is *linearly separable*. In that case one can achieve arbitrarily small loss values by tending the parameters to infinity along certain directions. To compensate for these drawbacks the logistic loss is usually regularized by some penalty term (Hastie et al., 2009; Bunea, 2008).

One of the candidates for the penalty function is the (squared) ℓ_2-norm of $\mathbf{x}$ (i.e., $\|\mathbf{x}\|_2^2$). Considering a positive penalty coefficient η the regularized loss is

$$f(\mathbf{x}) = g(\mathbf{x}) + \frac{\eta}{2}\|\mathbf{x}\|_2^2.$$

For any convex $g(\cdot)$ this regularized loss is guaranteed to be η-strongly convex, thus it has a unique minimum. Furthermore, the penalty term implicitly bounds the length of the minimizer thereby resolving the aforementioned problems. Nevertheless, the ℓ_2 penalty does not promote sparse solutions. Therefore, it is often desirable to impose an explicit sparsity constraint, in addition to the ℓ_2 regularizer.

4.1 Verifying SRH for ℓ_2-regularized Logistic Loss

It is easy to show that the Hessian of the logistic loss at any point $\mathbf{x}$ is given by $\mathbf{H}_g(\mathbf{x}) = \frac{1}{4n}\mathbf{A}^{\mathrm{T}}\Lambda\mathbf{A}$, where Λ is an $n \times n$ diagonal matrix whose diagonal entries are $\Lambda_{ii} = \operatorname{sech}^2\frac{1}{2}\left\langle \mathbf{a}_i,\mathbf{x}\right\rangle$ with $\operatorname{sech}(\cdot)$ denoting the *hyperbolic secant* function. Note that $\mathbf{0} \preccurlyeq \mathbf{H}_g(\mathbf{x}) \preccurlyeq \frac{1}{4n}\mathbf{A}^{\mathrm{T}}\mathbf{A}$. Therefore, if $\mathbf{H}_\eta(\mathbf{x})$ denotes the Hessian of the ℓ_2-regularized logistic loss, we have

$$\forall \mathbf{x},\Delta \qquad\qquad \eta\|\Delta\|_2^2 \leq \Delta^{\mathrm{T}}\mathbf{H}_\eta(\mathbf{x})\Delta \leq \frac{1}{4n}\|\mathbf{A}\Delta\|_2^2 + \eta\|\Delta\|_2^2. \tag{9}$$

To verify SRH, the upper and lower bounds achieved at k-sparse vectors Δ are of particular interest. It only remains to find an appropriate upper bound for $\|\mathbf{A}\Delta\|_2^2$ in terms of $\|\Delta\|_2^2$. To this end we use the following result on Chernoff bounds for random matrices due to Tropp (2012).

Theorem 3 (Matrix Chernoff (Tropp, 2012)). *Consider a finite sequence $\{\mathbf{M}_i\}$ of $k \times k$, independent, random, self-adjoint matrices that satisfy*

$$\mathbf{M}_i \succcurlyeq \mathbf{0} \quad \textit{and} \quad \lambda_{\max}\left(\mathbf{M}_i\right) \le R \quad \textit{almost surely.}$$

Let $\theta_{\max} := \lambda_{\max}\left(\sum_i \mathbb{E}\left[\mathbf{M}_i\right]\right)$. *Then for* $\tau \ge 0$,

$$\Pr\left\{\lambda_{\max}\left(\sum_i \mathbf{M}_i\right) \ge (1+\tau)\,\theta_{\max}\right\} \le k \exp\left(\frac{\theta_{\max}}{R}\left(\tau - (1+\tau)\log(1+\tau)\right)\right).$$

As stated before, in a standard logistic model data samples $\{\mathbf{a}_i\}$ are supposed to be independent instances of a random vector $\mathbf{a}$. In order to apply Theorem 3 we need to make the following extra assumptions:

Assumption. For every $\mathcal{J} \subseteq [p]$ with $|\mathcal{J}| = k$,

(i) we have $\left\|\mathbf{a}|_{\mathcal{J}}\right\|_2^2 \le R$ almost surely, and

(ii) none of the matrices $\mathbf{P}_{\mathcal{J}}^{\mathrm{T}} \mathbb{E}\left[\mathbf{a}\mathbf{a}^{\mathrm{T}}\right] \mathbf{P}_{\mathcal{J}}$ is the zero matrix.

We define $\theta_{\max}^{\mathcal{J}} := \lambda_{\max}\left(\mathbf{P}_{\mathcal{J}}^{\mathrm{T}} \mathbf{C} \mathbf{P}_{\mathcal{J}}\right)$, where $\mathbf{C} = \mathbb{E}\left[\mathbf{a}\mathbf{a}^{\mathrm{T}}\right]$, and let

$$\overline{\theta} := \max_{\mathcal{J} \subseteq [p]\,,\,|\mathcal{J}|=k} \theta_{\max}^{\mathcal{J}} \quad \text{and} \quad \widetilde{\theta} := \min_{\mathcal{J} \subseteq [p]\,,\,|\mathcal{J}|=k} \theta_{\max}^{\mathcal{J}}.$$

To simplify the notation henceforth we let $h(\tau) = (1+\tau)\log(1+\tau) - \tau$.

Corollary 1. *With the above assumptions, if* $n \ge \frac{R}{\widetilde{\theta} h(\tau)}\left(\log k + k\left(1 + \log \frac{p}{k}\right) - \log \varepsilon\right)$ *for some* $\tau > 0$ *and* $\varepsilon \in (0,1)$, *then with probability at least* $1 - \varepsilon$ *the* ℓ_2*-regularized logistic loss has* μ_k*-SRH with* $\mu_k \le 1 + \frac{1+\tau}{4\eta}\overline{\theta}$.

Proof For any set of k indices $\mathcal{J}$ let $\mathbf{M}_i^{\mathcal{J}} = \mathbf{a}_i|_{\mathcal{J}}\,\mathbf{a}_i|_{\mathcal{J}}^{\mathrm{T}} = \mathbf{P}_{\mathcal{J}}^{\mathrm{T}} \mathbf{a}_i \mathbf{a}_i^{\mathrm{T}} \mathbf{P}_{\mathcal{J}}$. The independence of the vectors $\mathbf{a}_i$ implies that the matrix

$$\begin{aligned}
\mathbf{A}_{\mathcal{J}}^{\mathrm{T}} \mathbf{A}_{\mathcal{J}} &= \sum_{i=1}^{n} \mathbf{a}_i|_{\mathcal{J}}\,\mathbf{a}_i|_{\mathcal{J}}^{\mathrm{T}} \\
&= \sum_{i=1}^{n} \mathbf{M}_i^{\mathcal{J}}
\end{aligned}$$

is a sum of n independent, random, self-adjoint matrices. Assumption (i) implies that $\lambda_{\max}\left(\mathbf{M}_i^{\mathcal{J}}\right) = \left\|\mathbf{a}_i|_{\mathcal{J}}\right\|_2^2 \le R$ almost surely. Furthermore, we have

$$\begin{aligned}
\lambda_{\max}\left(\sum_{i=1}^{n} \mathbb{E}\left[\mathbf{M}_i^{\mathcal{J}}\right]\right) &= \lambda_{\max}\left(\sum_{i=1}^{n} \mathbb{E}\left[\mathbf{P}_{\mathcal{J}}^{\mathrm{T}} \mathbf{a}_i \mathbf{a}_i^{\mathrm{T}} \mathbf{P}_{\mathcal{J}}\right]\right) \\
&= \lambda_{\max}\left(\sum_{i=1}^{n} \mathbf{P}_{\mathcal{J}}^{\mathrm{T}} \mathbb{E}\left[\mathbf{a}_i \mathbf{a}_i^{\mathrm{T}}\right] \mathbf{P}_{\mathcal{J}}\right) \\
&= \lambda_{\max}\left(\sum_{i=1}^{n} \mathbf{P}_{\mathcal{J}}^{\mathrm{T}} \mathbf{C} \mathbf{P}_{\mathcal{J}}\right) \\
&= n\lambda_{\max}\left(\mathbf{P}_{\mathcal{J}}^{\mathrm{T}} \mathbf{C} \mathbf{P}_{\mathcal{J}}\right) \\
&= n\theta_{\max}^{\mathcal{J}}.
\end{aligned}$$

Hence, for any fixed index set $\mathcal{J}$ with $|\mathcal{J}| = k$ we may apply Theorem 3 for $\mathbf{M}_i = \mathbf{M}_i^{\mathcal{J}}$, $\theta_{\max} = n\theta_{\max}^{\mathcal{J}}$, and $\tau > 0$ to obtain

$$\Pr\left\{\lambda_{\max}\left(\sum_{i=1}^{n}\mathbf{M}_i^{\mathcal{J}}\right) \geq (1+\tau)\, n\theta_{\max}^{\mathcal{J}}\right\} \leq k\exp\left(-\frac{n\theta_{\max}^{\mathcal{J}} h(\tau)}{R}\right).$$

Furthermore, we can write

$$\begin{aligned}\Pr\left\{\lambda_{\max}\left(\mathbf{A}_{\mathcal{J}}^{\mathrm{T}}\mathbf{A}_{\mathcal{J}}\right) \geq (1+\tau)\, n\overline{\theta}\right\} &= \Pr\left\{\lambda_{\max}\left(\sum_{i=1}^{n}\mathbf{M}_i^{\mathcal{J}}\right) \geq (1+\tau)\, n\overline{\theta}\right\}\\ &\leq \Pr\left\{\lambda_{\max}\left(\sum_{i=1}^{n}\mathbf{M}_i^{\mathcal{J}}\right) \geq (1+\tau)\, n\theta_{\max}^{\mathcal{J}}\right\}\\ &\leq k\exp\left(-\frac{n\theta_{\max}^{\mathcal{J}} h(\tau)}{R}\right)\\ &\leq k\exp\left(-\frac{n\widetilde{\theta} h(\tau)}{R}\right). \end{aligned} \qquad (10)$$

Note that Assumption (ii) guarantees that $\widetilde{\theta} > 0$, and thus the above probability bound will not be vacuous for sufficiently large n. To ensure a uniform guarantee for all $\binom{p}{k}$ possible choices of $\mathcal{J}$ we can use the union bound to obtain

$$\begin{aligned}\Pr\left\{\bigvee_{\substack{\mathcal{J}\subseteq[p]\\ |\mathcal{J}|=k}} \lambda_{\max}\left(\mathbf{A}_{\mathcal{J}}^{\mathrm{T}}\mathbf{A}_{\mathcal{J}}\right) \geq (1+\tau)\, n\overline{\theta}\right\} &\leq \sum_{\substack{\mathcal{J}\subseteq[p]\\ |\mathcal{J}|=k}} \Pr\left\{\lambda_{\max}\left(\mathbf{A}_{\mathcal{J}}^{\mathrm{T}}\mathbf{A}_{\mathcal{J}}\right) \geq (1+\tau)\, n\overline{\theta}\right\}\\ &\leq k\binom{p}{k}\exp\left(-\frac{n\widetilde{\theta} h(\tau)}{R}\right)\\ &\leq k\left(\frac{pe}{k}\right)^{k}\exp\left(-\frac{n\widetilde{\theta} h(\tau)}{R}\right)\\ &= \exp\left(\log k + k + k\log\frac{p}{k} - \frac{n\widetilde{\theta} h(\tau)}{R}\right). \end{aligned}$$

Therefore, for $\varepsilon \in (0,1)$ and $n \geq R\left(\log k + k\left(1+\log\frac{p}{k}\right) - \log\varepsilon\right) / \left(\widetilde{\theta} h(\tau)\right)$ it follows from (9) that for any $\mathbf{x}$ and any k-sparse Δ,

$$\eta\,\|\Delta\|_2^2 \leq \Delta^{\mathrm{T}}\mathbf{H}_\eta(\mathbf{x})\,\Delta \leq \left(\eta + \frac{1+\tau}{4}\overline{\theta}\right)\|\Delta\|_2^2$$

holds with probability at least $1-\varepsilon$. Thus, the ℓ_2-regularized logistic loss has an SRH constant $\mu_k \leq 1 + \frac{1+\tau}{4\eta}\overline{\theta}$ with probability $1-\varepsilon$. ■

Remark 10. One implication of this result is that for a regime in which k and p grow sufficiently large while $\frac{p}{k}$ remains constant one can achieve small failure rates provided that $n = \Omega\left(Rk\log\frac{p}{k}\right)$. Note that R is deliberately included in the argument of the order function because in general R depends on k. In other words, the above analysis may require $n = \Omega\left(k^2\log\frac{p}{k}\right)$ as the sufficient number of observations. This bound is a consequence of using Theorem 3, but to the best of our knowledge, other results regarding the extreme eigenvalues of the average of independent random PSD matrices also yield an n of the same order. If matrix $\mathbf{A}$ has certain additional properties (e.g., independent and sub-Gaussian entries), however, a better rate of $n = \Omega\left(k\log\frac{p}{k}\right)$ can be achieved without using the techniques mentioned above.

Remark 11. The analysis provided here is not specific to the ℓ_2-regularized logistic loss and can be readily extended to any other ℓ_2-regularized GLM loss whose log-partition function has a Lipschitz-continuous derivative.

4.2 Bounding the Approximation Error

We are going to bound $\|\nabla f(\mathbf{x}^\star)|_I\|_2$ which controls the approximation error in the statement of Theorem 1. In the case of case of ℓ_2-regularized logistic loss considered in this section we have

$$\nabla f(\mathbf{x}) = \sum_{i=1}^{n}\left(\frac{1}{1+\exp\left(-\langle\mathbf{a}_i,\mathbf{x}\rangle\right)} - y_i\right)\mathbf{a}_i + \eta\mathbf{x}.$$

Denoting $\frac{1}{1+\exp(-\langle\mathbf{a}_i,\mathbf{x}^\star\rangle)} - y_i$ by v_i for $i = 1,2,\ldots,n$ then we can deduce

$$\begin{aligned}\|\nabla f(\mathbf{x}^\star)|_I\|_2 &= \left\|\frac{1}{n}\sum_{i=1}^{n} v_i\,\mathbf{a}_i|_I + \eta\,\mathbf{x}^\star|_I\right\|_2\\ &= \left\|\frac{1}{n}\mathbf{A}_I^{\mathrm{T}}\mathbf{v} + \eta\,\mathbf{x}^\star|_I\right\|_2\\ &\leq \frac{1}{n}\left\|\mathbf{A}_I^{\mathrm{T}}\right\|\|\mathbf{v}\|_2 + \eta\,\|\mathbf{x}^\star|_I\|_2\\ &\leq \frac{1}{\sqrt{n}}\|\mathbf{A}_I\|\sqrt{\frac{1}{n}\sum_{i=1}^{n} v_i^2} + \eta\,\|\mathbf{x}^\star|_I\|_2,\end{aligned}$$

where $\mathbf{v} = [v_1\,v_2\ldots v_n]^{\mathrm{T}}$. Note that v_i's are n independent copies of the random variable $v = \frac{1}{1+\exp(-\langle\mathbf{a},\mathbf{x}^\star\rangle)} - y$ that is zero-mean and always lie in the interval $[-1,1]$. Therefore, applying the Hoeffding's inequality yields

$$\Pr\left\{\frac{1}{n}\sum_{i=1}^{n} v_i^2 \geq (1+c)\,\sigma_v^2\right\} \leq \exp\left(-2nc^2\sigma_v^4\right),$$

where $\sigma_v^2 = \mathbb{E}\left[v^2\right]$ is the variance of v. Furthermore, using the logistic model (8) we can deduce

$$\begin{aligned}
\sigma_v^2 &= \mathbb{E}\left[v^2\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[v^2 \mid \mathbf{a}\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left(y - \mathbb{E}\left[y \mid \mathbf{a}\right]\right)^2 \mid \mathbf{a}\right]\right] \\
&= \mathbb{E}\left[\operatorname{var}\left(y \mid \mathbf{a}\right)\right] \\
&= \mathbb{E}\left[\frac{1}{1+\exp\left(\langle \mathbf{a}, \mathbf{x}^\star \rangle\right)} \times \frac{\exp\left(\langle \mathbf{a}, \mathbf{x}^\star \rangle\right)}{1+\exp\left(\langle \mathbf{a}, \mathbf{x}^\star \rangle\right)}\right] && \text{(because } y \mid \mathbf{a} \sim \text{Bernoulli as in (8))} \\
&= \mathbb{E}\left[\frac{1}{2+\exp\left(\langle \mathbf{a}, \mathbf{x}^\star \rangle\right) + \exp\left(-\langle \mathbf{a}, \mathbf{x}^\star \rangle\right)}\right] \\
&\leq \frac{1}{4} && \text{(because } \exp(t) + \exp(-t) \geq 2\text{)}.
\end{aligned}$$

Therefore, we have $\frac{1}{n}\sum_{i=1}^n v_i^2 < \frac{1}{4}$ with high probability. As in the previous subsection one can also bound $\frac{1}{\sqrt{n}}\left\|\mathbf{A}_\mathcal{I}\right\| = \sqrt{\frac{1}{n}\lambda_{\max}\left(\mathbf{A}_\mathcal{I}^\mathrm{T}\mathbf{A}_\mathcal{I}\right)}$ using (10) with $k = |\mathcal{I}| = 3s$. Hence, with high probability we have

$$\left\|\nabla f\left(\mathbf{x}^\star\right)|_\mathcal{I}\right\|_2 \leq \frac{1}{2}\sqrt{(1+\tau)\,\overline{\theta}} + \eta\left\|\mathbf{x}^\star\right\|_2.$$

Interestingly, this analysis can also be extended to the GLMs whose log-partition function $\psi(\cdot)$ obeys $0 \leq \psi''(t) \leq C$ for all t with C being a positive constant. For these models the approximation error can be bounded in terms of the variance of $v_\psi = \psi'\left(\langle \mathbf{a}, \mathbf{x}^\star \rangle\right) - y$.

5. Experimental Results

Algorithms that are used for sparsity-constrained estimation or optimization often induce sparsity using different types of regularizations or constraints. Therefore, the *optimized* objective function may vary from one algorithm to another, even though all of these algorithms try to estimate the same sparse parameter and sparsely optimize the same original objective. Because of the discrepancy in the optimized objective functions it is generally difficult to compare performance of these algorithms. Applying algorithms on real data generally produces even less reliable results because of the unmanageable or unknown characteristics of the real data. Nevertheless, we evaluated performance of GraSP for variable selection in the logistic model both on synthetic and real data.

5.1 Synthetic Data

In our simulations the sparse parameter of interest $\mathbf{x}^\star$ is a $p = 1000$ dimensional vector that has $s = 10$ nonzero entries drawn independently from the standard Gaussian distribution. An intercept $c \in \mathbb{R}$ is also considered which is drawn independently of the other parameters according to the standard Gaussian distribution. Each data sample is an independent instance of the random vector $\mathbf{a} = \left[a_1, a_2, \ldots, a_p\right]^\mathrm{T}$ generated by an autoregressive process (Hamilton, 1994) determined by

$$a_{j+1} = \rho a_j + \sqrt{1-\rho^2} z_j, \qquad \text{for all } j \in [p-1]$$

with $a_1 \sim \mathcal{N}(0,1)$, $z_j \sim \mathcal{N}(0,1)$, and $\rho \in [0,1]$ being the correlation parameter. The data model we describe and use above is identical to the experimental model used in Agarwal et al. (2010), except that we adjusted the coefficients to ensure that $\mathbb{E}\left[a_j^2\right] = 1$ for all $j \in [p]$. The data labels, $y \in \{0,1\}$ are then drawn randomly according to the Bernoulli distribution with

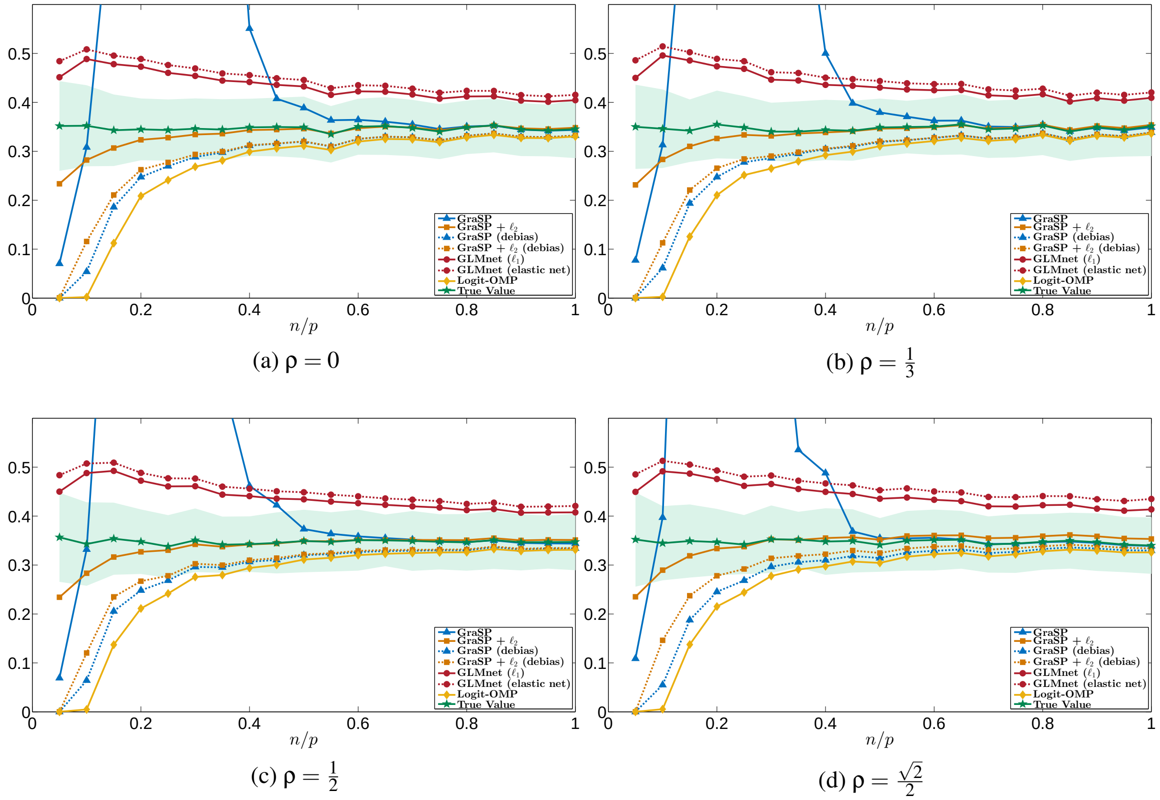
$$\Pr\{y = 0 \mid \mathbf{a}\} = 1/\left(1 + \exp\left(\langle \mathbf{a}, \mathbf{x}^{\star}\rangle + c\right)\right).$$

We compared GraSP to the LASSO algorithm implemented in the GLMnet package (Friedman et al., 2010), as well as the Orthogonal Matching Pursuit method dubbed Logit-OMP (Lozano et al., 2011). To isolate the effect of ℓ_2-regularization, both LASSO and the basic implementation of GraSP did not consider additional ℓ_2-regularization terms. To analyze the effect of an additional ℓ_2-regularization we also evaluated the performance of GraSP with ℓ_2-regularized logistic loss, as well as the logistic regression with elastic net (i.e., mixed ℓ_1-ℓ_2) penalty also available in the GLMnet package. We configured the GLMnet software to produce s-sparse solutions for a fair comparison. For the elastic net penalty $(1-\omega)\|\mathbf{x}\|_2^2/2 + \omega\|\mathbf{x}\|_1$ we considered the "mixing parameter" ω to be 0.8. For the ℓ_2-regularized logistic loss we considered $\eta = (1-\omega)\sqrt{\frac{\log p}{n}}$. For each choice of the number of measurements n between 50 and 1000 in steps of size 50, and ρ in the set $\left\{0, \frac{1}{3}, \frac{1}{2}, \frac{\sqrt{2}}{2}\right\}$ we generate the data and the associated labels and apply the algorithms. The average performance is measured over 200 trials for each pair of (n, ρ).

Figure 1 compares the average value of the empirical logistic loss achieved by each of the considered algorithms for a wide range of "sampling ratio" n/p. For GraSP, the curves labelled by GraSP and GraSP + ℓ_2 corresponding to the cases where the algorithm is applied to unregularized and ℓ_2-regularized logistic loss, respectively. Furthermore, the results of GLMnet for the LASSO and the elastic net regularization are labelled by GLMnet (ℓ_1) and GLMnet (elastic net), respectively. The simulation result of the Logit-OMP algorithm is also included. To contrast the obtained results we also provided the average of empirical logistic loss evaluated at the true parameter and one standard deviation above and below this average on the plots. Furthermore, we evaluated performance of GraSP with the debiasing procedure described in Section 3.1.

As can be seen from the figure at lower values of the sampling ratio GraSP is not accurate and does not seem to be converging. This behavior can be explained by the fact that without regularization at low sampling ratios the training data is linearly separable or has very few mislabelled samples. In either case, the value of the loss can vary significantly even in small neighborhoods. Therefore, the algorithm can become too sensitive to the pruning step at the end of each iteration. At larger sampling ratios, however, the loss from GraSP begins to decrease rapidly, becoming effectively identical to the loss at the true parameter for $n/p > 0.7$. The results show that unlike GraSP, Logit-OMP performs gracefully at lower sampling ratios. At higher sampling ratios, however, GraSP appears to yield smaller bias in the loss value. Furthermore, the difference between the loss obtained by the LASSO and the loss at the true parameter never drops below a certain threshold, although the convex method exhibits a more stable behaviour at low sampling ratios.

Interestingly, GraSP becomes more stable at low sampling ratios when the logistic loss is regularized with the ℓ_2-norm. However, this stability comes at the cost of a bias in the loss value at high sampling ratios that is particularly pronounced in Figure 1d. Nevertheless, for all of the tested values of ρ, at low sampling ratios GraSP+ℓ_2 and at high sampling ratios GraSP are consistently closer to the true loss value compared to the other methods. Debiasing the iterates of GraSP also



(a) $\rho = 0$

(b) $\rho = \frac{1}{3}$

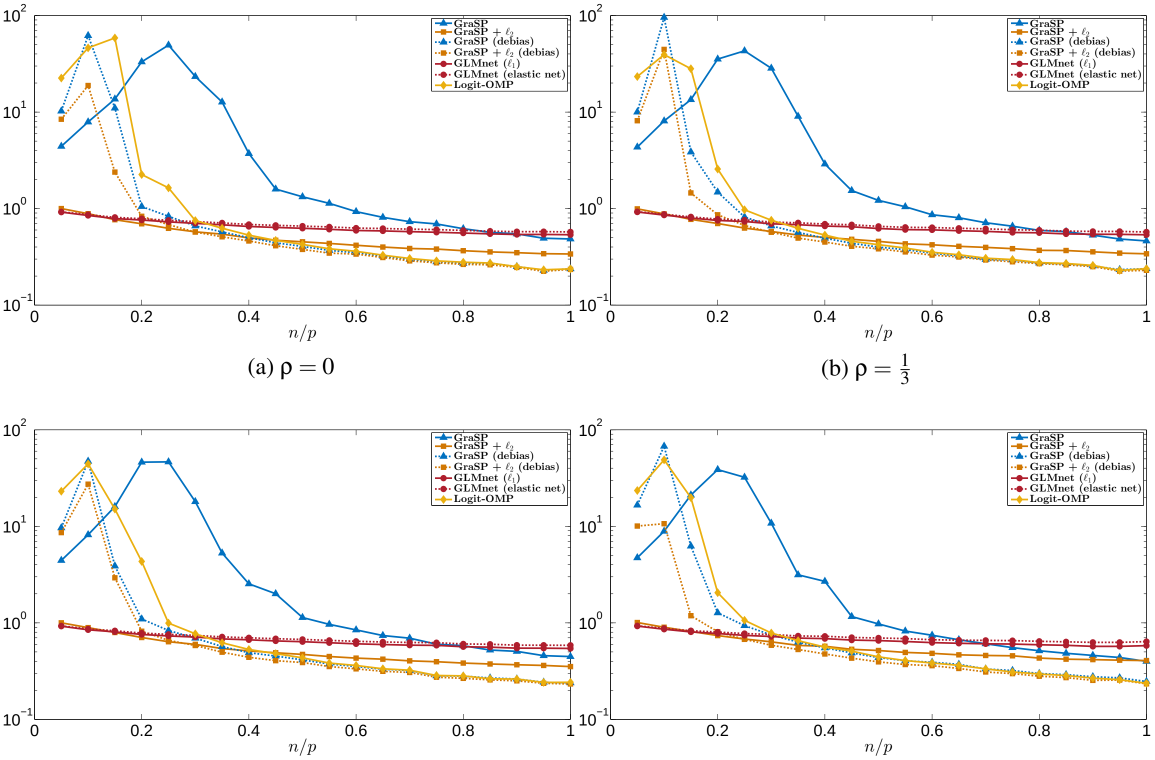
(c) $\rho = \frac{1}{2}$

(d) $\rho = \frac{\sqrt{2}}{2}$

Figure 1: Comparison of the average (empirical) logistic loss at solutions obtained via GraSP, GraSP with ℓ_2-penalty, LASSO, the elastic-net regularization, and Logit-OMP. The results of both GraSP methods with "debiasing" are also included. The average loss at the true parameter and one standard deviation interval around it are plotted as well.

appears to have a stabilizing effect at lower sampling ratios. For GraSP with ℓ_2 regularized cost, the debiasing particularly reduced the undesirable bias at $\rho = \frac{\sqrt{2}}{2}$.

Figure 2 illustrates the performance of the same algorithms in terms of the relative error $\frac{\|\widehat{\mathbf{x}}-\mathbf{x}^\star\|_2}{\|\mathbf{x}^\star\|_2}$ where $\widehat{\mathbf{x}}$ denotes the estimate that the algorithms produce. Not surprisingly, none of the algorithms attain an arbitrarily small relative error. Furthermore, the parameter ρ does not appear to affect the performance of the algorithms significantly. Without the ℓ_2-regularization, at high sampling ratios GraSP provides an estimate that has a comparable error versus the ℓ_1-regularization method. However, for mid to high sampling ratios both GraSP and GLMnet methods are outperformed by Logit-OMP. At low to mid sampling ratios, GraSP is unstable and does not converge to an estimate close to the true parameter. Logit-OMP shows similar behavior at lower sampling ratios. Performance of GraSP changes dramatically once we consider the ℓ_2-regularization and/or the debiasing procedure. With ℓ_2-regularization, GraSP achieves better relative error compared to GLMnet and ordinary GraSP for almost the entire range of tested sampling ratios. Applying the debiasing procedure has improved the performance of both GraSP methods except at very low sampling ratios.



(a) $\rho = 0$

(b) $\rho = \frac{1}{3}$

(c) $\rho = \frac{1}{2}$

(d) $\rho = \frac{\sqrt{2}}{2}$

Figure 2: Comparison of the average relative error (i.e., $\frac{\|\hat{\mathbf{x}}-\mathbf{x}^\star\|_2}{\|\mathbf{x}^\star\|_2}$) in logarithmic scale at solutions obtained via GraSP, GraSP with ℓ_2-penalty, LASSO, the elastic-net regularization, and Logit-OMP. The results of both GraSP methods with "debiasing" are also included.

These variants of GraSP appear to perform better than Logit-OMP for almost the entire range of n/p.

5.2 Real Data

We also conducted the same simulation on some of the data sets used in NIPS 2003 Workshop on feature extraction (Guyon et al., 2005), namely the ARCENE and DEXTER data sets. The logistic loss values at obtained estimates are reported in Tables 2 and 3. For each data set we applied the sparse logistic regression for a range of sparsity level s. The columns indicated by "G" correspond to different variants of GraSP. Suffixes ℓ_2 and "d" indicate the ℓ_2-regularization and the debiasing are applied, respectively. The columns indicated by ℓ_1 and E-net correspond to the results of the ℓ_1-regularization and the elastic-net regularization methods that are performed using the GLMnet package. The last column contains the result of the Logit-OMP algorithm.

The results for DEXTER data set show that GraSP variants without debiasing and the convex methods achieve comparable loss values in most cases, whereas the convex methods show significantly better performance on the ARCENE data set. Nevertheless, except for a few instances where

s	G	Gd	Gℓ_2	Gℓ_2d	ℓ_1	E-net	Logit-OMP
5	5.89E+01	5.75E-01	2.02E+01	5.24E-01	5.59E-01	6.43E-01	**2.23E-01**
10	3.17E+02	5.43E-01	3.71E+01	4.53E-01	5.10E-01	5.98E-01	**5.31E-07**
15	3.38E+02	6.40E-07	5.94E+00	**1.42E-07**	4.86E-01	5.29E-01	5.31E-07
20	1.21E+02	3.44E-07	8.82E+00	**3.08E-08**	4.52E-01	5.19E-01	5.31E-07
25	9.87E+02	1.13E-07	4.46E+01	**1.35E-08**	4.18E-01	4.96E-01	5.31E-07

Table 2: ARCENE

s	G	Gd	Gℓ_2	Gℓ_2d	ℓ_1	E-net	Logit-OMP
5	7.58E+00	3.28E-01	3.30E+00	2.80E-01	5.75E-01	6.08E-01	**2.64E-01**
10	1.08E+00	1.79E-01	4.33E-01	**1.28E-01**	5.23E-01	5.33E-01	1.79E-01
15	6.06E+00	1.71E-01	3.35E-01	1.17E-01	4.88E-01	4.98E-01	**1.16E-01**
20	1.30E+00	8.84E-02	1.79E-01	8.19E-02	4.27E-01	4.36E-01	**4.60E-02**
25	1.17E+00	**2.51E-07**	2.85E-01	1.17E-02	3.94E-01	4.12E-01	4.62E-03
30	3.04E-01	5.83E-07	2.65E-01	**1.77E-07**	3.70E-01	3.88E-01	2.88E-07
35	6.22E-01	2.08E-07	2.68E-01	**1.19E-07**	3.47E-01	3.72E-01	2.14E-07
40	5.38E-01	2.01E-07	6.30E-02	**1.27E-07**	3.31E-01	3.56E-01	2.14E-07
45	3.29E-01	2.11E-07	1.05E-01	**1.47E-07**	3.16E-01	3.41E-01	2.14E-07
50	2.06E-01	**1.31E-07**	5.66E-02	1.46E-07	2.87E-01	3.11E-01	2.14E-07
55	3.61E-02	**1.20E-07**	8.40E-02	1.31E-07	2.80E-01	2.89E-01	2.14E-07
60	1.18E-01	2.46E-07	5.70E-02	**1.09E-07**	2.66E-01	2.82E-01	2.14E-07
65	1.18E-01	**7.86E-08**	2.87E-02	9.47E-08	2.59E-01	2.75E-01	2.14E-07
70	8.92E-02	1.17E-07	2.23E-02	**8.15E-08**	2.52E-01	2.69E-01	2.14E-07
75	1.03E-01	8.54E-08	3.93E-02	**7.94E-08**	2.45E-01	2.69E-01	2.14E-07

Table 3: DEXTER

Logit-OMP has the best performance, the smallest loss values in both data sets are attained by GraSP methods with debiasing step.

6. Discussion and Conclusion

In many applications understanding high dimensional data or systems that involve these types of data can be reduced to identification of a sparse parameter. For example, in gene selection problems researchers are interested in locating a few genes among thousands of genes that cause or contribute to a particular disease. These problems can usually be cast as sparsity-constrained optimizations. In this paper we introduce a greedy algorithm called the Gradient Support Pursuit(GraSP) as an approximate solver for a wide range of sparsity-constrained optimization problems.

We provide theoretical convergence guarantees based on the notions of a Stable Restricted Hessian (SRH) for smooth cost functions and a Stable Restricted Linearization (SRL) for non-smooth cost functions, both of which are introduced in this paper. Our algorithm generalizes the well-established sparse recovery algorithm CoSaMP that merely applies in linear models with squared error loss. The SRH and SRL also generalize the well-known Restricted Isometry Property for sparse recovery to the case of cost functions other than the squared error. To provide a concrete example we studied the requirements of GraSP for ℓ_2-regularized logistic loss. Using a similar approach one can verify SRH condition for loss functions that have Lipschitz-continuous gradient that incorporates a broad family of loss functions.

At medium- and large-scale problems computational cost of the GraSP algorithm is mostly affected by the inner convex optimization step whose complexity is polynomial in s. On the other hand, for very large-scale problems, especially with respect to the dimension of the input, p, the running time of the GraSP algorithm will be dominated by evaluation of the function and its gradient, whose computational cost grows with p. This problem is common in algorithms that only have deterministic steps; even ordinary coordinate-descent methods have this limitation (Nesterov, 2012). Similar to improvements gained by using randomization in coordinate-descent methods (Nesterov, 2012), introducing randomization in the GraSP algorithm could reduce its computational complexity at large-scale problems. This extension, however, is beyond the scope of this paper and we leave it for future work.

Appendix A. Iteration Analysis For Smooth Cost Functions

To analyze our algorithm we first establish a series of results on how the algorithm operates on its current estimate, leading to an iteration invariant property on the estimation error. Propositions 1 and 2 are used to prove Lemmas 1 and 2. These Lemmas then are used to prove Lemma 3 that provides an iteration invariant which in turn yields the main result.

Proposition 1. *Let $\mathbf{M}(t)$ be a matrix-valued function such that for all $t \in [0,1]$, $\mathbf{M}(t)$ is symmetric and its eigenvalues lie in interval $[B(t), A(t)]$ with $B(t) > 0$. Then for any vector $\mathbf{v}$ we have*

$$\left(\int_0^1 B(t)\mathrm{d}t\right)\|\mathbf{v}\|_2 \leq \left\|\left(\int_0^1 \mathbf{M}(t)\mathrm{d}t\right)\mathbf{v}\right\|_2 \leq \left(\int_0^1 A(t)\mathrm{d}t\right)\|\mathbf{v}\|_2.$$

Proof Let $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ denote the smallest and largest eigenvalue functions defined over the set of symmetric positive-definite matrices, respectively. These functions are in order concave and convex. Therefore, Jensen's inequality yields

$$\lambda_{\min}\left(\int_0^1 \mathbf{M}(t)\mathrm{d}t\right) \geq \int_0^1 \lambda_{\min}(\mathbf{M}(t))\,\mathrm{d}t \geq \int_0^1 B(t)\mathrm{d}t$$

and

$$\lambda_{\max}\left(\int_0^1 \mathbf{M}(t)\mathrm{d}t\right) \leq \int_0^1 \lambda_{\max}(\mathbf{M}(t))\,\mathrm{d}t \leq \int_0^1 A(t)\mathrm{d}t,$$

which imply the desired result. ■

Proposition 2. *Let $\mathbf{M}(t)$ be a matrix-valued function such that for all $t \in [0,1]$ $\mathbf{M}(t)$ is symmetric and its eigenvalues lie in interval $[B(t), A(t)]$ with $B(t) > 0$. If Γ is a subset of row/column indices of $\mathbf{M}(\cdot)$ then for any vector $\mathbf{v}$ we have*

$$\left\| \left(\int_0^1 \mathbf{P}_\Gamma^{\mathrm{T}} \mathbf{M}(t) \mathbf{P}_{\Gamma^c} \mathrm{d}t \right) \mathbf{v} \right\|_2 \leq \int_0^1 \frac{A(t) - B(t)}{2} \mathrm{d}t \, \|\mathbf{v}\|_2 .$$

Proof Since $\mathbf{M}(t)$ is symmetric, it is also diagonalizable. Thus, for any vector $\mathbf{v}$ we may write

$$B(t) \|\mathbf{v}\|_2^2 \leq \mathbf{v}^{\mathrm{T}} \mathbf{M}(t) \mathbf{v} \leq A(t) \|\mathbf{v}\|_2^2 ,$$

and thereby

$$-\frac{A(t)-B(t)}{2} \leq \frac{\mathbf{v}^{\mathrm{T}} \left(\mathbf{M}(t) - \frac{A(t)+B(t)}{2} \mathbf{I} \right) \mathbf{v}}{\|\mathbf{v}\|^2} \leq \frac{A(t)-B(t)}{2}.$$

Since $\mathbf{M}(t) - \frac{A(t)+B(t)}{2}\mathbf{I}$ is also diagonalizable, it follows from the above inequality that

$$\left\| \mathbf{M}(t) - \frac{A(t)+B(t)}{2} \mathbf{I} \right\| \leq \frac{A(t)-B(t)}{2}.$$

Let $\widetilde{\mathbf{M}}(t) = \mathbf{P}_\Gamma^{\mathrm{T}} \mathbf{M}(t) \mathbf{P}_{\Gamma^c}$. Since $\widetilde{\mathbf{M}}(t)$ is a submatrix of $\mathbf{M}(t) - \frac{A(t)+B(t)}{2}\mathbf{I}$ we should have

$$\left\| \widetilde{\mathbf{M}}(t) \right\| \leq \left\| \mathbf{M}(t) - \frac{A(t)+B(t)}{2} \mathbf{I} \right\| \leq \frac{A(t)-B(t)}{2}. \tag{11}$$

Finally, it follows from the convexity of the operator norm, Jensen's inequality, and (11) that

$$\left\| \int_0^1 \widetilde{\mathbf{M}}(t) \, \mathrm{d}t \right\| \leq \int_0^1 \left\| \widetilde{\mathbf{M}}(t) \right\| \mathrm{d}t \leq \int_0^1 \frac{A(t)-B(t)}{2} \mathrm{d}t.$$

■

To simplify notation we introduce functions

$$\begin{aligned}
\alpha_k(\mathbf{p},\mathbf{q}) &= \int_0^1 A_k(t\mathbf{q} + (1-t)\mathbf{p}) \, \mathrm{d}t \\
\beta_k(\mathbf{p},\mathbf{q}) &= \int_0^1 B_k(t\mathbf{q} + (1-t)\mathbf{p}) \, \mathrm{d}t \\
\gamma_k(\mathbf{p},\mathbf{q}) &= \alpha_k(\mathbf{p},\mathbf{q}) - \beta_k(\mathbf{p},\mathbf{q}),
\end{aligned}$$

where $A_k(\cdot)$ and $B_k(\cdot)$ are defined by (5) and (6), respectively.

Lemma 1. *Let $\mathcal{R}$ denote the set* $\operatorname{supp}(\widehat{\mathbf{x}}-\mathbf{x}^\star)$. *The current estimate* $\widehat{\mathbf{x}}$ *then satisfies*

$$\left\|(\widehat{\mathbf{x}}-\mathbf{x}^\star)|_{\mathcal{Z}^c}\right\|_2 \leq \frac{\gamma_{4s}(\widehat{\mathbf{x}},\mathbf{x}^\star)+\gamma_{2s}(\widehat{\mathbf{x}},\mathbf{x}^\star)}{2\beta_{2s}(\widehat{\mathbf{x}},\mathbf{x}^\star)}\|\widehat{\mathbf{x}}-\mathbf{x}^\star\|_2+\frac{\left\|\nabla f(\mathbf{x}^\star)|_{\mathcal{R}\backslash\mathcal{Z}}\right\|_2+\left\|\nabla f(\mathbf{x}^\star)|_{\mathcal{Z}\backslash\mathcal{R}}\right\|_2}{\beta_{2s}(\widehat{\mathbf{x}},\mathbf{x}^\star)}.$$

Proof Since $\mathcal{Z}=\operatorname{supp}(\mathbf{z}_{2s})$ and $|\mathcal{R}|\leq 2s$ we have $\left\|\mathbf{z}|_{\mathcal{R}}\right\|_2\leq\|\mathbf{z}|_{\mathcal{Z}}\|_2$ and thereby

$$\left\|\mathbf{z}|_{\mathcal{R}\backslash\mathcal{Z}}\right\|_2\leq\left\|\mathbf{z}|_{\mathcal{Z}\backslash\mathcal{R}}\right\|_2. \tag{12}$$

Furthermore, because $\mathbf{z}=\nabla f(\widehat{\mathbf{x}})$ we can write

$$\begin{aligned}
\left\|\mathbf{z}|_{\mathcal{R}\backslash\mathcal{Z}}\right\|_2 &\geq \left\|\nabla f(\widehat{\mathbf{x}})|_{\mathcal{R}\backslash\mathcal{Z}}-\nabla f(\mathbf{x}^\star)|_{\mathcal{R}\backslash\mathcal{Z}}\right\|_2-\left\|\nabla f(\mathbf{x}^\star)|_{\mathcal{R}\backslash\mathcal{Z}}\right\|_2\\
&=\left\|\left(\int_0^1\mathbf{P}^{\mathrm{T}}_{\mathcal{R}\backslash\mathcal{Z}}\mathbf{H}_f(t\widehat{\mathbf{x}}+(1-t)\mathbf{x}^\star)\,\mathrm{d}t\right)(\widehat{\mathbf{x}}-\mathbf{x}^\star)\right\|_2-\left\|\nabla f(\mathbf{x}^\star)|_{\mathcal{R}\backslash\mathcal{Z}}\right\|_2\\
&\geq\left\|\left(\int_0^1\mathbf{P}^{\mathrm{T}}_{\mathcal{R}\backslash\mathcal{Z}}\mathbf{H}_f(t\widehat{\mathbf{x}}+(1-t)\mathbf{x}^\star)\,\mathbf{P}_{\mathcal{R}\backslash\mathcal{Z}}\mathrm{d}t\right)(\widehat{\mathbf{x}}-\mathbf{x}^\star)|_{\mathcal{R}\backslash\mathcal{Z}}\right\|_2-\left\|\nabla f(\mathbf{x}^\star)|_{\mathcal{R}\backslash\mathcal{Z}}\right\|_2\\
&\quad-\left\|\left(\int_0^1\mathbf{P}^{\mathrm{T}}_{\mathcal{R}\backslash\mathcal{Z}}\mathbf{H}_f(t\widehat{\mathbf{x}}+(1-t)\mathbf{x}^\star)\,\mathbf{P}_{\mathcal{Z}\cap\mathcal{R}}\mathrm{d}t\right)(\widehat{\mathbf{x}}-\mathbf{x}^\star)|_{\mathcal{Z}\cap\mathcal{R}}\right\|_2,
\end{aligned}$$

where we split the active coordinates (i.e., $\mathcal{R}$) into the sets $\mathcal{R}\backslash\mathcal{Z}$ and $\mathcal{Z}\cap\mathcal{R}$ to apply the triangle inequality and obtain the last expression. Applying Propositions 1 and 2 yields

$$\begin{aligned}
\left\|\mathbf{z}|_{\mathcal{R}\backslash\mathcal{Z}}\right\|_2 &\geq \beta_{2s}(\widehat{\mathbf{x}},\mathbf{x}^\star)\left\|(\widehat{\mathbf{x}}-\mathbf{x}^\star)|_{\mathcal{R}\backslash\mathcal{Z}}\right\|_2-\frac{\gamma_{2s}(\widehat{\mathbf{x}},\mathbf{x}^\star)}{2}\left\|(\widehat{\mathbf{x}}-\mathbf{x}^\star)|_{\mathcal{Z}\cap\mathcal{R}}\right\|_2-\left\|\nabla f(\mathbf{x}^\star)|_{\mathcal{R}\backslash\mathcal{Z}}\right\|_2\\
&\geq \beta_{2s}(\widehat{\mathbf{x}},\mathbf{x}^\star)\left\|(\widehat{\mathbf{x}}-\mathbf{x}^\star)|_{\mathcal{R}\backslash\mathcal{Z}}\right\|_2-\frac{\gamma_{2s}(\widehat{\mathbf{x}},\mathbf{x}^\star)}{2}\|\widehat{\mathbf{x}}-\mathbf{x}^\star\|_2-\left\|\nabla f(\mathbf{x}^\star)|_{\mathcal{R}\backslash\mathcal{Z}}\right\|_2.
\end{aligned}\tag{13}$$

Similarly, we have

$$\begin{aligned}
\left\|\mathbf{z}|_{\mathcal{Z}\backslash\mathcal{R}}\right\|_2 &\leq \left\|\nabla f(\widehat{\mathbf{x}})|_{\mathcal{Z}\backslash\mathcal{R}}-\nabla f(\mathbf{x}^\star)|_{\mathcal{Z}\backslash\mathcal{R}}\right\|_2+\left\|\nabla f(\mathbf{x}^\star)|_{\mathcal{Z}\backslash\mathcal{R}}\right\|_2\\
&=\left\|\left(\int_0^1\mathbf{P}^{\mathrm{T}}_{\mathcal{Z}\backslash\mathcal{R}}\mathbf{H}_f(t\widehat{\mathbf{x}}+(1-t)\mathbf{x}^\star)\,\mathbf{P}_{\mathcal{R}}\mathrm{d}t\right)(\widehat{\mathbf{x}}-\mathbf{x}^\star)|_{\mathcal{R}}\right\|_2+\left\|\nabla f(\mathbf{x}^\star)|_{\mathcal{Z}\backslash\mathcal{R}}\right\|_2\\
&\leq\frac{\gamma_{4s}(\widehat{\mathbf{x}},\mathbf{x}^\star)}{2}\left\|(\widehat{\mathbf{x}}-\mathbf{x}^\star)|_{\mathcal{R}}\right\|_2+\left\|\nabla f(\mathbf{x}^\star)|_{\mathcal{Z}\backslash\mathcal{R}}\right\|_2\\
&=\frac{\gamma_{4s}(\widehat{\mathbf{x}},\mathbf{x}^\star)}{2}\|\widehat{\mathbf{x}}-\mathbf{x}^\star\|_2+\left\|\nabla f(\mathbf{x}^\star)|_{\mathcal{Z}\backslash\mathcal{R}}\right\|_2.
\end{aligned}\tag{14}$$

Combining (12), (13), and (14) we obtain

$$\begin{aligned}
\frac{\gamma_{4s}(\widehat{\mathbf{x}},\mathbf{x}^\star)}{2}\|\widehat{\mathbf{x}}-\mathbf{x}^\star\|_2+\left\|\nabla f(\mathbf{x}^\star)|_{\mathcal{Z}\backslash\mathcal{R}}\right\|_2 &\geq\left\|\mathbf{z}|_{\mathcal{Z}\backslash\mathcal{R}}\right\|_2\\
&\geq\left\|\mathbf{z}|_{\mathcal{R}\backslash\mathcal{Z}}\right\|_2\\
&\geq\beta_{2s}(\widehat{\mathbf{x}},\mathbf{x}^\star)\left\|(\widehat{\mathbf{x}}-\mathbf{x}^\star)|_{\mathcal{R}\backslash\mathcal{Z}}\right\|_2-\frac{\gamma_{2s}(\widehat{\mathbf{x}},\mathbf{x}^\star)}{2}\|\widehat{\mathbf{x}}-\mathbf{x}^\star\|_2\\
&\quad-\left\|\nabla f(\mathbf{x}^\star)|_{\mathcal{R}\backslash\mathcal{Z}}\right\|_2.
\end{aligned}$$

Since $\mathcal{R} = \mathrm{supp}\,(\widehat{\mathbf{x}}-\mathbf{x}^\star)$, we have $\left\|(\widehat{\mathbf{x}}-\mathbf{x}^\star)\,|_{\mathcal{R}\backslash\mathcal{Z}}\right\|_2 = \|(\widehat{\mathbf{x}}-\mathbf{x}^\star)\,|_{\mathcal{Z}^c}\|_2$. Hence,

$$\|(\widehat{\mathbf{x}}-\mathbf{x}^\star)\,|_{\mathcal{Z}^c}\|_2 \leq \frac{\gamma_{4s}\,(\widehat{\mathbf{x}},\mathbf{x}^\star)+\gamma_{2s}\,(\widehat{\mathbf{x}},\mathbf{x}^\star)}{2\beta_{2s}\,(\widehat{\mathbf{x}},\mathbf{x}^\star)}\,\|\widehat{\mathbf{x}}-\mathbf{x}^\star\|_2+\frac{\left\|\nabla f\,(\mathbf{x}^\star)\,|_{\mathcal{R}\backslash\mathcal{Z}}\right\|_2+\left\|\nabla f\,(\mathbf{x}^\star)\,|_{\mathcal{Z}\backslash\mathcal{R}}\right\|}{\beta_{2s}\,(\widehat{\mathbf{x}},\mathbf{x}^\star)}.$$

■

Lemma 2. *The vector* $\mathbf{b}$ *given by*

$$\mathbf{b} = \arg\min f\,(\mathbf{x}) \;\text{ s.t. } \mathbf{x}|_{\mathcal{T}^c} = 0 \tag{15}$$

satisfies

$$\|\,\mathbf{x}^\star|_{\mathcal{T}}-\mathbf{b}\|_2 \leq \frac{\|\nabla f\,(\mathbf{x}^\star)\,|_{\mathcal{T}}\|_2}{\beta_{4s}\,(\mathbf{b},\mathbf{x}^\star)}+\frac{\gamma_{4s}\,(\mathbf{b},\mathbf{x}^\star)}{2\beta_{4s}\,(\mathbf{b},\mathbf{x}^\star)}\,\|\mathbf{x}^\star|_{\mathcal{T}^c}\|_2\,.$$

Proof We have

$$\nabla f\,(\mathbf{x}^\star)-\nabla f\,(\mathbf{b}) = \int_0^1 \mathbf{H}_f\,(t\mathbf{x}^\star+(1-t)\,\mathbf{b})\,\mathrm{d}t\;(\mathbf{x}^\star-\mathbf{b})\,.$$

Furthermore, since $\mathbf{b}$ is the solution to (15) we must have $\nabla f\,(\mathbf{b})\,|_{\mathcal{T}} = 0$. Therefore,

$$\begin{aligned}\nabla f\,(\mathbf{x}^\star)\,|_{\mathcal{T}} &= \left(\int_0^1 \mathbf{P}_{\mathcal{T}}^{\mathrm{T}}\mathbf{H}_f\,(t\mathbf{x}^\star+(1-t)\,\mathbf{b})\,\mathrm{d}t\right)\,(\mathbf{x}^\star-\mathbf{b})\\ &= \left(\int_0^1 \mathbf{P}_{\mathcal{T}}^{\mathrm{T}}\mathbf{H}_f\,(t\mathbf{x}^\star+(1-t)\,\mathbf{b})\,\mathbf{P}_{\mathcal{T}}\mathrm{d}t\right)\,(\mathbf{x}^\star-\mathbf{b})\,|_{\mathcal{T}}\\ &\quad+\left(\int_0^1 \mathbf{P}_{\mathcal{T}}^{\mathrm{T}}\mathbf{H}_f(t\mathbf{x}^\star+(1-t)\,\mathbf{b})\,\mathbf{P}_{\mathcal{T}^c}\mathrm{d}t\right)\,(\mathbf{x}^\star-\mathbf{b})\,|_{\mathcal{T}^c}.\end{aligned} \tag{16}$$

Since f has μ_{4s}-SRH and $|\mathcal{T}\cup\mathrm{supp}\,(t\mathbf{x}^\star+(1-t)\,\mathbf{b})| \leq 4s$ for all $t \in [0,1]$, functions $A_{4s}\,(\cdot)$ and $B_{4s}\,(\cdot)$, defined using (5) and (6), exist such that we have

$$B_{4s}\,(t\mathbf{x}^\star+(1-t)\,\mathbf{b}) \leq \lambda_{\min}\left(\mathbf{P}_{\mathcal{T}}^{\mathrm{T}}\mathbf{H}_f\,(t\mathbf{x}^\star+(1-t)\,\mathbf{b})\,\mathbf{P}_{\mathcal{T}}\right)$$

and

$$A_{4s}\,(t\mathbf{x}^\star+(1-t)\,\mathbf{b}) \geq \lambda_{\max}\left(\mathbf{P}_{\mathcal{T}}^{\mathrm{T}}\mathbf{H}_f\,(t\mathbf{x}^\star+(1-t)\,\mathbf{b})\,\mathbf{P}_{\mathcal{T}}\right).$$

Thus, from Proposition 1 we obtain

$$\beta_{4s}\,(\mathbf{b},\mathbf{x}^\star) \leq \lambda_{\min}\left(\int_0^1 \mathbf{P}_{\mathcal{T}}^{\mathrm{T}}\mathbf{H}_f\,(t\mathbf{x}^\star+(1-t)\,\mathbf{b})\,\mathbf{P}_{\mathcal{T}}\mathrm{d}t\right)$$

and

$$\alpha_{4s}\left(\mathbf{b},\mathbf{x}^{\star}\right) \geq \lambda_{\max}\left(\int_0^1 \mathbf{P}_{\mathcal{T}}^{\mathrm{T}}\mathbf{H}_f\left(t\mathbf{x}^{\star}+\left(1-t\right)\mathbf{b}\right)\mathbf{P}_{\mathcal{T}}\mathrm{d}t\right).$$

This result implies that the matrix $\int_0^1 \mathbf{P}_{\mathcal{T}}^{\mathrm{T}}\mathbf{H}_f\left(t\mathbf{x}^{\star}+\left(1-t\right)\mathbf{b}\right)\mathbf{P}_{\mathcal{T}}\mathrm{d}t$, henceforth denoted by $\mathbf{W}$, is invertible and

$$\frac{1}{\alpha_{4s}\left(\mathbf{b},\mathbf{x}^{\star}\right)} \leq \lambda_{\min}\left(\mathbf{W}^{-1}\right) \leq \lambda_{\max}\left(\mathbf{W}^{-1}\right) \leq \frac{1}{\beta_{4s}\left(\mathbf{b},\mathbf{x}^{\star}\right)}, \tag{17}$$

where we used the fact that $\lambda_{\max}\left(\mathbf{M}\right)\lambda_{\min}\left(\mathbf{M}^{-1}\right)=1$ for any positive-definite matrix $\mathbf{M}$, particularly for $\mathbf{W}$ and $\mathbf{W}^{-1}$. Therefore, by multiplying both sides of (16) by $\mathbf{W}^{-1}$ obtain

$$\mathbf{W}^{-1}\nabla f\left(\mathbf{x}^{\star}\right)|_{\mathcal{T}} = \left(\mathbf{x}^{\star}-\mathbf{b}\right)|_{\mathcal{T}} + \mathbf{W}^{-1}\left(\int_0^1 \mathbf{P}_{\mathcal{T}}^{\mathrm{T}}\mathbf{H}_f(t\mathbf{x}^{\star}+(1-t)\mathbf{b})\mathbf{P}_{\mathcal{T}^c}\mathrm{d}t\right)\mathbf{x}^{\star}|_{\mathcal{T}^c},$$

where we also used the fact that $\left(\mathbf{x}^{\star}-\mathbf{b}\right)|_{\mathcal{T}^c}=\mathbf{x}^{\star}|_{\mathcal{T}^c}$. With $\mathcal{S}^{\star}=\mathrm{supp}\left(\mathbf{x}^{\star}\right)$, using triangle inequality, (17), and Proposition 2 then we obtain

$$\begin{aligned}
\|\mathbf{x}^{\star}|_{\mathcal{T}}-\mathbf{b}\|_2 &= \|\left(\mathbf{x}^{\star}-\mathbf{b}\right)|_{\mathcal{T}}\|_2 \\
&\leq \left\|\mathbf{W}^{-1}\left(\int_0^1 \mathbf{P}_{\mathcal{T}}^{\mathrm{T}}\mathbf{H}_f(t\mathbf{x}^{\star}+(1-t)\mathbf{b})\mathbf{P}_{\mathcal{T}^c\cap\mathcal{S}^{\star}}\mathrm{d}t\right)\mathbf{x}^{\star}|_{\mathcal{T}^c\cap\mathcal{S}^{\star}}\right\|_2 + \left\|\mathbf{W}^{-1}\nabla f(\mathbf{x}^{\star})|_{\mathcal{T}}\right\|_2 \\
&\leq \frac{\|\nabla f\left(\mathbf{x}^{\star}\right)|_{\mathcal{T}}\|_2}{\beta_{4s}\left(\mathbf{b},\mathbf{x}^{\star}\right)} + \frac{\gamma_{4s}\left(\mathbf{b},\mathbf{x}^{\star}\right)}{2\beta_{4s}\left(\mathbf{b},\mathbf{x}^{\star}\right)}\|\mathbf{x}^{\star}|_{\mathcal{T}^c}\|_2,
\end{aligned}$$

as desired. $\blacksquare$

Lemma 3 (Iteration Invariant)**.** *The estimation error in the current iteration,* $\|\widehat{\mathbf{x}}-\mathbf{x}^{\star}\|_2$*, and that in the next iteration,* $\|\mathbf{b}_s-\mathbf{x}^{\star}\|_2$*, are related by the inequality:*

$$\begin{aligned}
\|\mathbf{b}_s-\mathbf{x}^{\star}\|_2 \leq{}& \frac{\gamma_{4s}\left(\widehat{\mathbf{x}},\mathbf{x}^{\star}\right)+\gamma_{2s}\left(\widehat{\mathbf{x}},\mathbf{x}^{\star}\right)}{2\beta_{2s}\left(\widehat{\mathbf{x}},\mathbf{x}^{\star}\right)}\left(1+\frac{\gamma_{4s}\left(\mathbf{b},\mathbf{x}^{\star}\right)}{\beta_{4s}\left(\mathbf{b},\mathbf{x}^{\star}\right)}\right)\|\widehat{\mathbf{x}}-\mathbf{x}^{\star}\|_2 \\
&+\left(1+\frac{\gamma_{4s}\left(\mathbf{b},\mathbf{x}^{\star}\right)}{\beta_{4s}\left(\mathbf{b},\mathbf{x}^{\star}\right)}\right)\frac{\left\|\nabla f\left(\mathbf{x}^{\star}\right)|_{\mathcal{R}\backslash\mathcal{Z}}\right\|_2+\left\|\nabla f\left(\mathbf{x}^{\star}\right)|_{\mathcal{Z}\backslash\mathcal{R}}\right\|_2}{\beta_{2s}\left(\widehat{\mathbf{x}},\mathbf{x}^{\star}\right)}+\frac{2\|\nabla f\left(\mathbf{x}^{\star}\right)|_{\mathcal{T}}\|_2}{\beta_{4s}\left(\mathbf{b},\mathbf{x}^{\star}\right)}.
\end{aligned}$$

Proof Because $\mathcal{Z}\subseteq\mathcal{T}$ we must have $\mathcal{T}^c\subseteq\mathcal{Z}^c$. Therefore, we can write $\|\mathbf{x}^{\star}|_{\mathcal{T}^c}\|_2=\|\left(\widehat{\mathbf{x}}-\mathbf{x}^{\star}\right)|_{\mathcal{T}^c}\|_2\leq\|\left(\widehat{\mathbf{x}}-\mathbf{x}^{\star}\right)|_{\mathcal{Z}^c}\|_2$. Then using Lemma 1 we obtain

$$\|\mathbf{x}^{\star}|_{\mathcal{T}^c}\|_2 \leq \frac{\gamma_{4s}(\widehat{\mathbf{x}},\mathbf{x}^{\star})+\gamma_{2s}(\widehat{\mathbf{x}},\mathbf{x}^{\star})}{2\beta_{2s}(\widehat{\mathbf{x}},\mathbf{x}^{\star})}\|\widehat{\mathbf{x}}-\mathbf{x}^{\star}\|_2+\frac{\left\|\nabla f(\mathbf{x}^{\star})|_{\mathcal{R}\backslash\mathcal{Z}}\right\|_2+\left\|\nabla f(\mathbf{x}^{\star})|_{\mathcal{Z}\backslash\mathcal{R}}\right\|_2}{\beta_{2s}(\widehat{\mathbf{x}},\mathbf{x}^{\star})}. \tag{18}$$

Furthermore,

$$\begin{aligned}
\|\mathbf{b}_s-\mathbf{x}^{\star}\|_2 &\leq \|\mathbf{b}_s-\mathbf{x}^{\star}|_{\mathcal{T}}\|_2+\|\mathbf{x}^{\star}|_{\mathcal{T}^c}\|_2 \\
&\leq \|\mathbf{x}^{\star}|_{\mathcal{T}}-\mathbf{b}\|_2+\|\mathbf{b}_s-\mathbf{b}\|_2+\|\mathbf{x}^{\star}|_{\mathcal{T}^c}\|_2 \leq 2\|\mathbf{x}^{\star}|_{\mathcal{T}}-\mathbf{b}\|_2+\|\mathbf{x}^{\star}|_{\mathcal{T}^c}\|_2,
\end{aligned} \tag{19}$$

where the last inequality holds because $\|\mathbf{x}^\star|_{\mathcal{T}}\|_0 \le s$ and $\mathbf{b}_s$ is the best s-term approximation of $\mathbf{b}$. Therefore, using Lemma 2,

$$\|\mathbf{b}_s-\mathbf{x}^\star\|_2 \le \frac{2}{\beta_{4s}(\mathbf{b},\mathbf{x}^\star)}\|\nabla f(\mathbf{x}^\star)|_{\mathcal{T}}\|_2 + \left(1+\frac{\gamma_{4s}(\mathbf{b},\mathbf{x}^\star)}{\beta_{4s}(\mathbf{b},\mathbf{x}^\star)}\right)\|\mathbf{x}^\star|_{\mathcal{T}^c}\|_2 . \tag{20}$$

Combining (18) and (20) we obtain

$$\begin{aligned}\|\mathbf{b}_s-\mathbf{x}^\star\|_2 \le & \frac{\gamma_{4s}(\widehat{\mathbf{x}},\mathbf{x}^\star)+\gamma_{2s}(\widehat{\mathbf{x}},\mathbf{x}^\star)}{2\beta_{2s}(\widehat{\mathbf{x}},\mathbf{x}^\star)}\left(1+\frac{\gamma_{4s}(\mathbf{b},\mathbf{x}^\star)}{\beta_{4s}(\mathbf{b},\mathbf{x}^\star)}\right)\|\widehat{\mathbf{x}}-\mathbf{x}^\star\|_2 \\ & + \left(1+\frac{\gamma_{4s}(\mathbf{b},\mathbf{x}^\star)}{\beta_{4s}(\mathbf{b},\mathbf{x}^\star)}\right)\frac{\left\|\nabla f(\mathbf{x}^\star)|_{\mathcal{R}\setminus\mathcal{Z}}\right\|_2+\left\|\nabla f(\mathbf{x}^\star)|_{\mathcal{Z}\setminus\mathcal{R}}\right\|_2}{\beta_{2s}(\widehat{\mathbf{x}},\mathbf{x}^\star)} + \frac{2\|\nabla f(\mathbf{x}^\star)|_{\mathcal{T}}\|_2}{\beta_{4s}(\mathbf{b},\mathbf{x}^\star)}.\end{aligned}$$

■

Using the results above, we can now prove Theorem 1.

***Proof of Theorem* 1.** Using definition 1 it is easy to verify that for $k \le k'$ and any vector $\mathbf{u}$ we have $A_k(\mathbf{u}) \le A_{k'}(\mathbf{u})$ and $B_k(\mathbf{u}) \ge B_{k'}(\mathbf{u})$. Consequently, for $k \le k'$ and any pair of vectors $\mathbf{p}$ and $\mathbf{q}$ we have $\alpha_k(\mathbf{p},\mathbf{q}) \le \alpha_{k'}(\mathbf{p},\mathbf{q})$, $\beta_k(\mathbf{p},\mathbf{q}) \ge \beta_{k'}(\mathbf{p},\mathbf{q})$, and $\mu_k \le \mu_{k'}$. Furthermore, for any function that satisfies μ_k-SRH we can write

$$\frac{\alpha_k(\mathbf{p},\mathbf{q})}{\beta_k(\mathbf{p},\mathbf{q})} = \frac{\int_0^1 A_k(t\mathbf{q}+(1-t)\mathbf{p})\,\mathrm{d}t}{\int_0^1 B_k(t\mathbf{q}+(1-t)\mathbf{p})\,\mathrm{d}t} \le \frac{\int_0^1 \mu_k B_k(t\mathbf{q}+(1-t)\mathbf{p})\,\mathrm{d}t}{\int_0^1 B_k(t\mathbf{q}+(1-t)\mathbf{p})\,\mathrm{d}t} = \mu_k,$$

and thereby $\frac{\gamma_k(\mathbf{p},\mathbf{q})}{\beta_k(\mathbf{p},\mathbf{q})} \le \mu_k - 1$. Therefore, applying Lemma 3 to the estimate in the i-th iterate of the algorithm shows that

$$\begin{aligned}\left\|\widehat{\mathbf{x}}^{(i)}-\mathbf{x}^\star\right\|_2 \le & (\mu_{4s}-1)\mu_{4s}\left\|\widehat{\mathbf{x}}^{(i-1)}-\mathbf{x}^\star\right\|_2 + \frac{2\|\nabla f(\mathbf{x}^\star)|_{\mathcal{T}}\|_2}{\beta_{4s}(\mathbf{b},\mathbf{x}^\star)} \\ & + \mu_{4s}\frac{\left\|\nabla f(\mathbf{x}^\star)|_{\mathcal{R}\setminus\mathcal{Z}}\right\|_2+\left\|\nabla f(\mathbf{x}^\star)|_{\mathcal{Z}\setminus\mathcal{R}}\right\|_2}{\beta_{2s}\left(\widehat{\mathbf{x}}^{(i-1)},\mathbf{x}^\star\right)} \\ \le & \left(\mu_{4s}^2-\mu_{4s}\right)\left\|\widehat{\mathbf{x}}^{(i-1)}-\mathbf{x}^\star\right\|_2 + \frac{2}{\varepsilon}\|\nabla f(\mathbf{x}^\star)|_{\mathcal{I}}\|_2 + \frac{2\mu_{4s}}{\varepsilon}\|\nabla f(\mathbf{x}^\star)|_{\mathcal{I}}\|_2 .\end{aligned}$$

Applying the assumption $\mu_{4s} \le \frac{1+\sqrt{3}}{2}$ then yields

$$\left\|\widehat{\mathbf{x}}^{(i)}-\mathbf{x}^\star\right\|_2 \le \frac{1}{2}\left\|\widehat{\mathbf{x}}^{(i-1)}-\mathbf{x}^\star\right\|_2 + \frac{3+\sqrt{3}}{\varepsilon}\|\nabla f(\mathbf{x}^\star)|_{\mathcal{I}}\|_2 .$$

The theorem follows using this inequality recursively. ■

Appendix B. Iteration Analysis For Non-Smooth Cost Functions

In this part we provide analysis of GraSP for non-smooth functions. Definition 3 basically states that for any k-sparse vector $\mathbf{x} \in \mathbb{R}^n$, $\alpha_k(\mathbf{x})$ and $\beta_k(\mathbf{x})$ are in order the smallest and largest values for which

$$\beta_k(\mathbf{x})\|\Delta\|_2^2 \le \mathrm{B}_f(\mathbf{x}+\Delta \parallel \mathbf{x}) \le \alpha_k(\mathbf{x})\|\Delta\|_2^2 \tag{21}$$

holds for all vectors $\Delta \in \mathbb{R}^n$ that satisfy $|\mathrm{supp}(\mathbf{x}) \cup \mathrm{supp}(\Delta)| \leq k$. By interchanging $\mathbf{x}$ and $\mathbf{x}+\Delta$ in (21) and using the fact that

$$\mathrm{B}_f(\mathbf{x}+\Delta \parallel \mathbf{x}) + \mathrm{B}_f(\mathbf{x} \parallel \mathbf{x}+\Delta) = \left\langle \nabla_f(\mathbf{x}+\Delta) - \nabla_f(\mathbf{x}), \Delta \right\rangle$$

one can easily deduce

$$[\beta_k(\mathbf{x}+\Delta)+\beta_k(\mathbf{x})]\|\Delta\|_2^2 \leq \left\langle \nabla_f(\mathbf{x}+\Delta)-\nabla_f(\mathbf{x}), \Delta \right\rangle \leq [\alpha_k(\mathbf{x}+\Delta)+\alpha_k(\mathbf{x})]\|\Delta\|_2^2. \tag{22}$$

Propositions 3, 4, and 5 establish some basic inequalities regarding the restricted Bregman divergence under SRL assumption. Using these inequalities we prove Lemmas 4 and 5. These two Lemmas are then used to prove an iteration invariant result in Lemma 6 which in turn is used to prove Theorem 2.

Note In Propositions 3, 4, and 5 we assume $\mathbf{x}_1$ and $\mathbf{x}_2$ are two vectors in $\mathbb{R}^n$ such that $|\mathrm{supp}(\mathbf{x}_1) \cup \mathrm{supp}(\mathbf{x}_2)| \leq r$. Furthermore, we use the shorthand $\Delta = \mathbf{x}_1 - \mathbf{x}_2$ and denote $\mathrm{supp}(\Delta)$ by $\mathcal{R}$. We also denote $\nabla_f(\mathbf{x}_1) - \nabla_f(\mathbf{x}_2)$ by Δ'. To simplify the notation further the shorthands $\overline{\alpha}_l$, $\overline{\beta}_l$, and $\overline{\gamma}_l$ are used for $\overline{\alpha}_l(\mathbf{x}_1,\mathbf{x}_2) := \alpha_l(\mathbf{x}_1) + \alpha_l(\mathbf{x}_2)$, $\overline{\beta}_l(\mathbf{x}_1,\mathbf{x}_2) := \beta_l(\mathbf{x}_1)+\beta_l(\mathbf{x}_2)$, and $\overline{\gamma}_l(\mathbf{x}_1,\mathbf{x}_2) := \overline{\alpha}_l(\mathbf{x}_1,\mathbf{x}_2) - \overline{\beta}_l(\mathbf{x}_1,\mathbf{x}_2)$, respectively.

Proposition 3. *Let $\mathcal{R}'$ be a subset of $\mathcal{R}$. Then the following inequalities hold.*

$$\left|\overline{\alpha}_r \left\|\Delta|_{\mathcal{R}'}\right\|_2^2 - \left\langle \Delta', \Delta|_{\mathcal{R}'}\right\rangle\right| \leq \overline{\gamma}_r \left\|\Delta|_{\mathcal{R}'}\right\|_2 \|\Delta\|_2 \tag{23}$$

$$\left|\overline{\beta}_r \left\|\Delta|_{\mathcal{R}'}\right\|_2^2 - \left\langle \Delta', \Delta|_{\mathcal{R}'}\right\rangle\right| \leq \overline{\gamma}_r \left\|\Delta|_{\mathcal{R}'}\right\|_2 \|\Delta\|_2$$

Proof Using (21) we can write

$$\beta_r(\mathbf{x}_1)\left\|\Delta|_{\mathcal{R}'}\right\|_2^2 t^2 \leq \mathrm{B}_f\left(\mathbf{x}_1 - t\,\Delta|_{\mathcal{R}'} \parallel \mathbf{x}_1\right) \leq \alpha_r(\mathbf{x}_1)\left\|\Delta|_{\mathcal{R}'}\right\|_2^2 t^2 \tag{24}$$

$$\beta_r(\mathbf{x}_2)\left\|\Delta|_{\mathcal{R}'}\right\|_2^2 t^2 \leq \mathrm{B}_f\left(\mathbf{x}_2 - t\,\Delta|_{\mathcal{R}'} \parallel \mathbf{x}_2\right) \leq \alpha_r(\mathbf{x}_2)\left\|\Delta|_{\mathcal{R}'}\right\|_2^2 t^2 \tag{25}$$

and

$$\beta_r(\mathbf{x}_1)\left\|\Delta - t\,\Delta|_{\mathcal{R}'}\right\|_2^2 \leq \mathrm{B}_f\left(\mathbf{x}_2 + t\,\Delta|_{\mathcal{R}'} \parallel \mathbf{x}_1\right) \leq \alpha_r(\mathbf{x}_1)\left\|\Delta - t\,\Delta|_{\mathcal{R}'}\right\|_2^2 \tag{26}$$

$$\beta_r(\mathbf{x}_2)\left\|\Delta - t\,\Delta|_{\mathcal{R}'}\right\|_2^2 \leq \mathrm{B}_f\left(\mathbf{x}_1 - t\,\Delta|_{\mathcal{R}'} \parallel \mathbf{x}_2\right) \leq \alpha_r(\mathbf{x}_2)\left\|\Delta - t\,\Delta|_{\mathcal{R}'}\right\|_2^2, \tag{27}$$

where t is an arbitrary real number. Using the definition of the Bregman divergence we can add (24) and (25) to obtain

$$\begin{aligned}\overline{\beta}_r \left\|\Delta|_{\mathcal{R}'}\right\|_2^2 t^2 &\leq f\left(\mathbf{x}_1 - t\,\Delta|_{\mathcal{R}'}\right) - f(\mathbf{x}_1) + f\left(\mathbf{x}_2 + t\,\Delta|_{\mathcal{R}'}\right) - f(\mathbf{x}_2) + \left\langle \Delta', \Delta|_{\mathcal{R}'}\right\rangle t \\ &\leq \overline{\alpha}_r \left\|\Delta|_{\mathcal{R}'}\right\|_2^2 t^2.\end{aligned} \tag{28}$$

Similarly, (26) and (27) yield

$$\begin{aligned}\overline{\beta}_r \left\|\Delta - t\,\Delta|_{\mathcal{R}'}\right\|_2^2 &\leq f\left(\mathbf{x}_1 - t\,\Delta|_{\mathcal{R}'}\right) - f(\mathbf{x}_1) + f\left(\mathbf{x}_2 + t\,\Delta|_{\mathcal{R}'}\right) - f(\mathbf{x}_2) + \left\langle \Delta', \Delta - t\,\Delta|_{\mathcal{R}'}\right\rangle \\ &\leq \overline{\alpha}_r \left\|\Delta - t\,\Delta|_{\mathcal{R}'}\right\|_2^2.\end{aligned} \tag{29}$$

Expanding the quadratic bounds of (29) and using (28) then we obtain

$$0 \leq \overline{\gamma}_r \left\| \Delta|_{\mathcal{R}'} \right\|_2^2 t^2 + 2\left(\overline{\beta}_r \left\| \Delta|_{\mathcal{R}'} \right\|_2^2 - \left\langle \Delta, \Delta|_{\mathcal{R}'} \right\rangle\right) t - \overline{\beta}_r \|\Delta\|_2^2 + \left\langle \Delta', \Delta \right\rangle \tag{30}$$

$$0 \leq \overline{\gamma}_r \left\| \Delta|_{\mathcal{R}'} \right\|_2^2 t^2 - 2\left(\overline{\alpha}_r \left\| \Delta|_{\mathcal{R}'} \right\|_2^2 - \left\langle \Delta, \Delta|_{\mathcal{R}'} \right\rangle\right) t + \overline{\alpha}_r \|\Delta\|_2^2 - \left\langle \Delta', \Delta \right\rangle. \tag{31}$$

It follows from (22), (30), and (31) that

$$\begin{aligned} 0 &\leq \overline{\gamma}_r \left\| \Delta|_{\mathcal{R}'} \right\|_2^2 t^2 + 2\left(\overline{\beta}_r \left\| \Delta|_{\mathcal{R}'} \right\|_2^2 - \left\langle \Delta, \Delta|_{\mathcal{R}'} \right\rangle\right) t + \overline{\gamma}_r \|\Delta\|_2^2 \\ 0 &\leq \overline{\gamma}_r \left\| \Delta|_{\mathcal{R}'} \right\|_2^2 t^2 - 2\left(\overline{\alpha}_r \left\| \Delta|_{\mathcal{R}'} \right\|_2^2 - \left\langle \Delta, \Delta|_{\mathcal{R}'} \right\rangle\right) t + \overline{\gamma}_r \|\Delta\|_2^2. \end{aligned}$$

These two quadratic inequalities hold for any $t \in \mathbb{R}$ thus their discriminants are not positive, that is,

$$\begin{aligned} \left(\overline{\beta}_r \left\| \Delta|_{\mathcal{R}'} \right\|_2^2 - \left\langle \Delta', \Delta|_{\mathcal{R}'} \right\rangle\right)^2 - \overline{\gamma}_r^2 \left\| \Delta|_{\mathcal{R}'} \right\|_2^2 \|\Delta\|_2^2 &\leq 0 \\ \left(\overline{\alpha}_r \left\| \Delta|_{\mathcal{R}'} \right\|_2^2 - \left\langle \Delta', \Delta|_{\mathcal{R}'} \right\rangle\right)^2 - \overline{\gamma}_r^2 \left\| \Delta|_{\mathcal{R}'} \right\|_2^2 \|\Delta\|_2^2 &\leq 0, \end{aligned}$$

which yield the desired result. ■

Proposition 4. *The following inequalities hold for $\mathcal{R}' \subseteq \mathcal{R}$.*

$$\left| \left\| \Delta'|_{\mathcal{R}'} \right\|_2^2 - \overline{\alpha}_r \left\langle \Delta', \Delta|_{\mathcal{R}'} \right\rangle \right| \leq \overline{\gamma}_r \left\| \Delta|_{\mathcal{R}'} \right\|_2 \|\Delta\|_2 \tag{32}$$

$$\left| \left\| \Delta'|_{\mathcal{R}'} \right\|_2^2 - \overline{\beta}_r \left\langle \Delta', \Delta|_{\mathcal{R}'} \right\rangle \right| \leq \overline{\gamma}_r \left\| \Delta|_{\mathcal{R}'} \right\|_2 \|\Delta\|_2$$

Proof From (21) we have

$$\beta_r(\mathbf{x}_1) \left\| \Delta'|_{\mathcal{R}'} \right\|_2^2 t^2 \leq \mathrm{B}_f\left(\mathbf{x}_1 - t\,\Delta'|_{\mathcal{R}'} \parallel \mathbf{x}_1\right) \leq \alpha_r(\mathbf{x}_1) \left\| \Delta'|_{\mathcal{R}'} \right\|_2^2 t^2 \tag{33}$$

$$\beta_r(\mathbf{x}_2) \left\| \Delta'|_{\mathcal{R}'} \right\|_2^2 t^2 \leq \mathrm{B}_f\left(\mathbf{x}_2 + t\,\Delta'|_{\mathcal{R}'} \parallel \mathbf{x}_2\right) \leq \alpha_r(\mathbf{x}_2) \left\| \Delta'|_{\mathcal{R}'} \right\|_2^2 t^2 \tag{34}$$

and

$$\beta_r(\mathbf{x}_1) \left\| \Delta - t\,\Delta'|_{\mathcal{R}'} \right\|_2^2 \leq \mathrm{B}_f\left(\mathbf{x}_2 + t\,\Delta'|_{\mathcal{R}'} \parallel \mathbf{x}_1\right) \leq \alpha_r(\mathbf{x}_1) \left\| \Delta - t\,\Delta'|_{\mathcal{R}'} \right\|_2^2 \tag{35}$$

$$\beta_r(\mathbf{x}_2) \left\| \Delta - t\,\Delta'|_{\mathcal{R}'} \right\|_2^2 \leq \mathrm{B}_f\left(\mathbf{x}_1 - t\,\Delta'|_{\mathcal{R}'} \parallel \mathbf{x}_2\right) \leq \alpha_r(\mathbf{x}_2) \left\| \Delta - t\,\Delta'|_{\mathcal{R}'} \right\|_2^2, \tag{36}$$

for any $t \in \mathbb{R}$. By subtracting the sum of (35) and (36) from that of (33) and (34) we obtain

$$\begin{aligned} \overline{\beta}_r \left\| \Delta'|_{\mathcal{R}'} \right\|_2^2 t^2 - \overline{\alpha}_r \left\| \Delta - t\,\Delta'|_{\mathcal{R}'} \right\|_2^2 &\leq 2\left\langle \Delta', \Delta'|_{\mathcal{R}'} \right\rangle t - \left\langle \Delta', \Delta \right\rangle \\ &\leq \overline{\alpha}_r \left\| \Delta'|_{\mathcal{R}'} \right\|_2^2 t^2 - \overline{\beta}_r \left\| \Delta - t\,\Delta'|_{\mathcal{R}'} \right\|_2^2. \end{aligned} \tag{37}$$

Expanding the bounds of (37) then yields

$$0 \leq \bar{\gamma}_r \left\| \Delta'|_{\mathcal{R}'} \right\|_2^2 t^2 + 2\left(\left\langle \Delta', \Delta'|_{\mathcal{R}'} \right\rangle - \bar{\alpha}_r \left\langle \Delta, \Delta'|_{\mathcal{R}'} \right\rangle \right) t + \bar{\alpha}_r \left\| \Delta \right\|_2^2 - \left\langle \Delta', \Delta \right\rangle$$
$$0 \leq \bar{\gamma}_r \left\| \Delta'|_{\mathcal{R}'} \right\|_2^2 t^2 - 2\left(\left\langle \Delta', \Delta'|_{\mathcal{R}'} \right\rangle - \bar{\beta}_r \left\langle \Delta, \Delta'|_{\mathcal{R}'} \right\rangle \right) t - \bar{\beta}_r \left\| \Delta \right\|_2^2 + \left\langle \Delta', \Delta \right\rangle .$$

Note that $\left\langle \Delta', \Delta'|_{\mathcal{R}'} \right\rangle = \left\| \Delta'|_{\mathcal{R}'} \right\|_2^2$ and $\left\langle \Delta, \Delta'|_{\mathcal{R}'} \right\rangle = \left\langle \Delta|_{\mathcal{R}'}, \Delta' \right\rangle$. Therefore, using (22) we obtain

$$0 \leq \bar{\gamma}_r \left\| \Delta'|_{\mathcal{R}'} \right\|_2^2 t^2 + 2\left(\left\| \Delta'|_{\mathcal{R}'} \right\|_2^2 - \bar{\alpha}_r \left\langle \Delta', \Delta|_{\mathcal{R}'} \right\rangle \right) t + \bar{\gamma}_r \left\| \Delta \right\|_2^2 \tag{38}$$
$$0 \leq \bar{\gamma}_r \left\| \Delta'|_{\mathcal{R}'} \right\|_2^2 t^2 - 2\left(\left\| \Delta'|_{\mathcal{R}'} \right\|_2^2 - \bar{\beta}_r \left\langle \Delta', \Delta|_{\mathcal{R}'} \right\rangle \right) t + \bar{\gamma}_r \left\| \Delta \right\|_2^2 . \tag{39}$$

Since the right-hand sides of (38) and (39) are quadratics in t and always non-negative for all values of $t \in \mathbb{R}$, their discriminants cannot be positive. Thus we have

$$\left(\left\| \Delta'|_{\mathcal{R}'} \right\|_2^2 - \bar{\alpha}_r \left\langle \Delta', \Delta|_{\mathcal{R}'} \right\rangle \right)^2 - \bar{\gamma}_r^2 \left\| \Delta'|_{\mathcal{R}'} \right\|_2^2 \left\| \Delta \right\|^2 \leq 0$$
$$\left(\left\| \Delta'|_{\mathcal{R}'} \right\|_2^2 - \bar{\beta}_r \left\langle \Delta', \Delta|_{\mathcal{R}'} \right\rangle \right)^2 - \bar{\gamma}_r^2 \left\| \Delta'|_{\mathcal{R}'} \right\|_2^2 \left\| \Delta \right\|^2 \leq 0,$$

which yield the desired result. ■

Corollary 2. *The inequality*

$$\left\| \Delta'|_{\mathcal{R}'} \right\|_2 \geq \bar{\beta}_r \left\| \Delta|_{\mathcal{R}'} \right\|_2 - \bar{\gamma}_r \left\| \Delta|_{\mathcal{R} \backslash \mathcal{R}'} \right\|_2 ,$$

holds for $\mathcal{R}' \subseteq \mathcal{R}$.

Proof It follows from (32) and (23) that

$$\begin{aligned} -\left\| \Delta'|_{\mathcal{R}'} \right\|_2^2 + \bar{\alpha}_r^2 \left\| \Delta|_{\mathcal{R}'} \right\|_2^2 &= -\left\| \Delta'|_{\mathcal{R}'} \right\|_2^2 + \bar{\alpha}_r \left\langle \Delta', \Delta|_{\mathcal{R}'} \right\rangle + \bar{\alpha}_r \left[\bar{\alpha}_r \left\| \Delta|_{\mathcal{R}'} \right\|_2^2 - \left\langle \Delta', \Delta|_{\mathcal{R}'} \right\rangle \right] \\ &\leq \bar{\gamma}_r \left\| \Delta'|_{\mathcal{R}'} \right\|_2 \left\| \Delta \right\|_2 + \bar{\alpha}_r \bar{\gamma}_r \left\| \Delta|_{\mathcal{R}'} \right\|_2 \left\| \Delta \right\|_2 . \end{aligned}$$

Therefore, after straightforward calculations we get

$$\begin{aligned} \left\| \Delta'|_{\mathcal{R}'} \right\|_2 &\geq \frac{1}{2} \left(-\bar{\gamma}_r \left\| \Delta \right\|_2 + \left| 2\bar{\alpha}_r \left\| \Delta|_{\mathcal{R}'} \right\|_2 - \bar{\gamma}_r \left\| \Delta \right\|_2 \right| \right) \\ &\geq \bar{\alpha}_r \left\| \Delta|_{\mathcal{R}'} \right\|_2 - \bar{\gamma}_r \left\| \Delta \right\|_2 \\ &\geq \bar{\beta}_r \left\| \Delta|_{\mathcal{R}'} \right\|_2 - \bar{\gamma}_r \left\| \Delta|_{\mathcal{R} \backslash \mathcal{R}'} \right\|_2 . \end{aligned}$$

■

Proposition 5. *Suppose that $\mathcal{K}$ is a subset of $\mathcal{R}^c$ with at most k elements. Then we have*

$$\left\|\Delta'|_{\mathcal{K}}\right\|_2 \leq \overline{\gamma}_{k+r}\left\|\Delta\right\|_2.$$

Proof Using (21) for any $t \in \mathbb{R}$ we can write

$$\beta_{k+r}(\mathbf{x}_1)\left\|\Delta'|_{\mathcal{K}}\right\|_2^2 t^2 \leq \mathrm{B}_f\left(\mathbf{x}_1 + t\,\Delta'|_{\mathcal{K}} \parallel \mathbf{x}_1\right) \leq \alpha_{k+r}(\mathbf{x}_1)\left\|\Delta'|_{\mathcal{K}}\right\|_2^2 t^2 \tag{40}$$

$$\beta_{k+r}(\mathbf{x}_2)\left\|\Delta'|_{\mathcal{K}}\right\|_2^2 t^2 \leq \mathrm{B}_f\left(\mathbf{x}_2 - t\,\Delta'|_{\mathcal{K}} \parallel \mathbf{x}_2\right) \leq \alpha_{k+r}(\mathbf{x}_2)\left\|\Delta'|_{\mathcal{K}}\right\|_2^2 t^2 \tag{41}$$

and similarly

$$\beta_{k+r}(\mathbf{x}_1)\left\|\Delta + t\,\Delta'|_{\mathcal{K}}\right\|_2^2 \leq \mathrm{B}_f\left(\mathbf{x}_2 - t\,\Delta'|_{\mathcal{K}} \parallel \mathbf{x}_1\right) \leq \alpha_{k+r}(\mathbf{x}_1)\left\|\Delta + t\,\Delta'|_{\mathcal{K}}\right\|_2^2 \tag{42}$$

$$\beta_{k+r}(\mathbf{x}_2)\left\|\Delta + t\,\Delta'|_{\mathcal{K}}\right\|_2^2 \leq \mathrm{B}_f\left(\mathbf{x}_1 + t\,\Delta'|_{\mathcal{K}} \parallel \mathbf{x}_2\right) \leq \alpha_{k+r}(\mathbf{x}_2)\left\|\Delta + t\,\Delta'|_{\mathcal{K}}\right\|_2^2. \tag{43}$$

By subtracting the sum of (42) and (43) from that of (40) and (41) we obtain

$$\begin{aligned}\overline{\beta}_{k+r}\left\|\Delta'|_{\mathcal{K}}\right\|_2^2 t^2 - \overline{\alpha}_{k+r}\left\|\Delta + t\,\Delta'|_{\mathcal{K}}\right\|_2^2 &\leq -2t\left\langle \Delta', \Delta'|_{\mathcal{K}}\right\rangle - \left\langle \Delta', \Delta\right\rangle \\ &\leq \overline{\alpha}_{k+r}\left\|\Delta'|_{\mathcal{K}}\right\|_2^2 t^2 - \overline{\beta}_{k+r}\left\|\Delta + t\,\Delta'|_{\mathcal{K}}\right\|_2^2.\end{aligned} \tag{44}$$

Note that $\left\langle \Delta', \Delta'|_{\mathcal{K}}\right\rangle = \left\|\Delta'|_{\mathcal{K}}\right\|_2^2$ and $\left\langle \Delta, \Delta'|_{\mathcal{K}}\right\rangle = 0$. Therefore, (22) and (44) imply

$$0 \leq \overline{\gamma}_{k+r}\left\|\Delta'|_{\mathcal{K}}\right\|_2^2 t^2 \pm 2\left\|\Delta'|_{\mathcal{K}}\right\|_2^2 t + \overline{\gamma}_{k+r}\left\|\Delta\right\|_2^2 \tag{45}$$

hold for all $t \in \mathbb{R}$. Hence, as quadratic functions of t, the right-hand side of (45) cannot have a positive discriminant. Thus we must have

$$\left\|\Delta'|_{\mathcal{K}}\right\|_2^4 - \overline{\gamma}_{k+r}^2\left\|\Delta\right\|_2^2\left\|\Delta'|_{\mathcal{K}}\right\|_2^2 \leq 0,$$

which yields the desired result. ■

Lemma 4. *Let $\mathcal{R}$ denote* $\mathrm{supp}\left(\widehat{\mathbf{x}} - \mathbf{x}^\star\right)$. *Then we have*

$$\left\|\left(\widehat{\mathbf{x}} - \mathbf{x}^\star\right)|_{\mathcal{Z}^c}\right\|_2 \leq \frac{\overline{\gamma}_{2s}\left(\widehat{\mathbf{x}}, \mathbf{x}^\star\right) + \overline{\gamma}_{4s}\left(\widehat{\mathbf{x}}, \mathbf{x}^\star\right)}{\overline{\beta}_{2s}\left(\widehat{\mathbf{x}}, \mathbf{x}^\star\right)}\left\|\widehat{\mathbf{x}} - \mathbf{x}^\star\right\|_2 + \frac{\left\|\nabla_f\left(\mathbf{x}^\star\right)|_{\mathcal{R}\backslash\mathcal{Z}}\right\|_2 + \left\|\nabla_f\left(\mathbf{x}^\star\right)|_{\mathcal{Z}\backslash\mathcal{R}}\right\|_2}{\overline{\beta}_{2s}\left(\widehat{\mathbf{x}}, \mathbf{x}^\star\right)}.$$

Proof Given that $\mathcal{Z} = \mathrm{supp}\left(\mathbf{z}_{2s}\right)$ and $|\mathcal{R}| \leq 2s$ we have $\left\|\mathbf{z}|_{\mathcal{R}}\right\|_2 \leq \left\|\mathbf{z}|_{\mathcal{Z}}\right\|_2$. Hence

$$\left\|\mathbf{z}|_{\mathcal{R}\backslash\mathcal{Z}}\right\|_2 \leq \left\|\mathbf{z}|_{\mathcal{Z}\backslash\mathcal{R}}\right\|_2. \tag{46}$$

Furthermore, using Corollary 2 we can write

$$\begin{aligned}\left\|\mathbf{z}|_{\mathcal{R}\setminus\mathcal{Z}}\right\|_2 &= \left\|\nabla_f\left(\widehat{\mathbf{x}}\right)|_{\mathcal{R}\setminus\mathcal{Z}}\right\|_2\\ &\geq \left\|\left(\nabla_f\left(\widehat{\mathbf{x}}\right)-\nabla_f\left(\mathbf{x}^\star\right)\right)|_{\mathcal{R}\setminus\mathcal{Z}}\right\|_2 - \left\|\nabla_f\left(\mathbf{x}^\star\right)|_{\mathcal{R}\setminus\mathcal{Z}}\right\|_2\\ &\geq \overline{\beta}_{2s}\left(\widehat{\mathbf{x}},\mathbf{x}^\star\right)\left\|\left(\widehat{\mathbf{x}}-\mathbf{x}^\star\right)|_{\mathcal{R}\setminus\mathcal{Z}}\right\|_2 - \overline{\gamma}_{2s}\left(\widehat{\mathbf{x}},\mathbf{x}^\star\right)\left\|\left(\widehat{\mathbf{x}}-\mathbf{x}^\star\right)|_{\mathcal{R}\cap\mathcal{Z}}\right\|_2 - \left\|\nabla_f\left(\mathbf{x}^\star\right)|_{\mathcal{R}\setminus\mathcal{Z}}\right\|_2\\ &\geq \overline{\beta}_{2s}\left(\widehat{\mathbf{x}},\mathbf{x}^\star\right)\left\|\left(\widehat{\mathbf{x}}-\mathbf{x}^\star\right)|_{\mathcal{R}\setminus\mathcal{Z}}\right\|_2 - \overline{\gamma}_{2s}\left(\widehat{\mathbf{x}},\mathbf{x}^\star\right)\left\|\widehat{\mathbf{x}}-\mathbf{x}^\star\right\|_2 - \left\|\nabla_f\left(\mathbf{x}^\star\right)|_{\mathcal{R}\setminus\mathcal{Z}}\right\|_2. \end{aligned} \tag{47}$$

Similarly, using Proposition 5 we have

$$\begin{aligned}\left\|\mathbf{z}|_{\mathcal{Z}\setminus\mathcal{R}}\right\|_2 = \left\|\nabla_f\left(\widehat{\mathbf{x}}\right)|_{\mathcal{Z}\setminus\mathcal{R}}\right\|_2 &\leq \left\|\left(\nabla_f\left(\widehat{\mathbf{x}}\right)-\nabla_f\left(\mathbf{x}^\star\right)\right)|_{\mathcal{Z}\setminus\mathcal{R}}\right\|_2 + \left\|\nabla_f\left(\mathbf{x}^\star\right)|_{\mathcal{Z}\setminus\mathcal{R}}\right\|_2\\ &\leq \overline{\gamma}_{4s}\left(\widehat{\mathbf{x}},\mathbf{x}^\star\right)\left\|\widehat{\mathbf{x}}-\mathbf{x}^\star\right\|_2 + \left\|\nabla_f\left(\mathbf{x}^\star\right)|_{\mathcal{Z}\setminus\mathcal{R}}\right\|_2. \end{aligned} \tag{48}$$

Combining (46), (47), and (48) then yields

$$\begin{aligned}\overline{\gamma}_{4s}\left(\widehat{\mathbf{x}},\mathbf{x}^\star\right)\left\|\widehat{\mathbf{x}}-\mathbf{x}^\star\right\|_2 + \left\|\nabla_f\left(\mathbf{x}^\star\right)|_{\mathcal{Z}\setminus\mathcal{R}}\right\|_2 \geq & -\overline{\gamma}_{2s}\left(\widehat{\mathbf{x}},\mathbf{x}^\star\right)\left\|\left(\widehat{\mathbf{x}}-\mathbf{x}^\star\right)|_{\mathcal{R}\cap\mathcal{Z}}\right\|_2\\ & +\overline{\beta}_{2s}\left(\widehat{\mathbf{x}},\mathbf{x}^\star\right)\left\|\left(\widehat{\mathbf{x}}-\mathbf{x}^\star\right)|_{\mathcal{R}\setminus\mathcal{Z}}\right\|_2 - \left\|\nabla_f\left(\mathbf{x}^\star\right)|_{\mathcal{R}\setminus\mathcal{Z}}\right\|_2. \end{aligned}$$

Note that $\left(\widehat{\mathbf{x}}-\mathbf{x}^\star\right)|_{\mathcal{R}\setminus\mathcal{Z}} = \left(\widehat{\mathbf{x}}-\mathbf{x}^\star\right)|_{\mathcal{Z}^c}$. Therefore, we have

$$\left\|\left(\widehat{\mathbf{x}}-\mathbf{x}^\star\right)|_{\mathcal{Z}^c}\right\|_2 \leq \frac{\overline{\gamma}_{2s}\left(\widehat{\mathbf{x}},\mathbf{x}^\star\right)+\overline{\gamma}_{4s}\left(\widehat{\mathbf{x}},\mathbf{x}^\star\right)}{\overline{\beta}_{2s}\left(\widehat{\mathbf{x}},\mathbf{x}^\star\right)}\left\|\widehat{\mathbf{x}}-\mathbf{x}^\star\right\|_2 + \frac{\left\|\nabla_f\left(\mathbf{x}^\star\right)|_{\mathcal{R}\setminus\mathcal{Z}}\right\|_2 + \left\|\nabla_f\left(\mathbf{x}^\star\right)|_{\mathcal{Z}\setminus\mathcal{R}}\right\|_2}{\overline{\beta}_{2s}\left(\widehat{\mathbf{x}},\mathbf{x}^\star\right)}.$$

■

Lemma 5. *The vector* **b** *given by*

$$\mathbf{b} = \arg\min_{\mathbf{x}}\ f\left(\mathbf{x}\right) \quad \text{s.t. } \mathbf{x}|_{\mathcal{T}^c} = \mathbf{0} \tag{49}$$

satisfies $\left\|\mathbf{x}^\star|_{\mathcal{T}} - \mathbf{b}\right\|_2 \leq \frac{\left\|\nabla_f\left(\mathbf{x}^\star\right)|_{\mathcal{T}}\right\|_2}{\overline{\beta}_{4s}\left(\mathbf{x}^\star,\mathbf{b}\right)} + \left(1+\frac{\overline{\gamma}_{4s}\left(\mathbf{x}^\star,\mathbf{b}\right)}{\overline{\beta}_{4s}\left(\mathbf{x}^\star,\mathbf{b}\right)}\right)\left\|\mathbf{x}^\star|_{\mathcal{T}^c}\right\|_2$.

Proof Since **b** satisfies (49) we must have $\nabla_f\left(\mathbf{b}\right)|_{\mathcal{T}} = \mathbf{0}$. Then it follows from Corollary 2 that

$$\begin{aligned}\left\|\mathbf{x}^\star|_{\mathcal{T}} - \mathbf{b}\right\|_2 &= \left\|\left(\mathbf{x}^\star - \mathbf{b}\right)|_{\mathcal{T}}\right\|_2\\ &\leq \frac{\left\|\nabla_f\left(\mathbf{x}^\star\right)|_{\mathcal{T}}\right\|_2}{\overline{\beta}_{4s}\left(\mathbf{x}^\star,\mathbf{b}\right)} + \frac{\overline{\gamma}_{4s}\left(\mathbf{x}^\star,\mathbf{b}\right)}{\overline{\beta}_{4s}\left(\mathbf{x}^\star,\mathbf{b}\right)}\left\|\mathbf{x}^\star|_{\mathcal{T}^c}\right\|_2. \end{aligned}$$

■

Lemma 6. *The estimation error of the current iterate* (*i.e.*, $\|\widehat{\mathbf{x}}-\mathbf{x}^\star\|_2$) *and that of the next iterate* (*i.e.*, $\|\mathbf{b}_s-\mathbf{x}^\star\|_2$) *are related by the inequality:*

$$\begin{aligned}\|\mathbf{b}_s-\mathbf{x}^\star\|_2 \leq &\left(1+\frac{2\overline{\gamma}_{4s}\left(\mathbf{x}^\star,\mathbf{b}\right)}{\overline{\beta}_{4s}\left(\mathbf{x}^\star,\mathbf{b}\right)}\right)\frac{\overline{\gamma}_{2s}\left(\widehat{\mathbf{x}},\mathbf{x}^\star\right)+\overline{\gamma}_{4s}\left(\widehat{\mathbf{x}},\mathbf{x}^\star\right)}{\overline{\beta}_{2s}\left(\widehat{\mathbf{x}}^i,\mathbf{x}^\star\right)}\|\widehat{\mathbf{x}}-\mathbf{x}^\star\|_2+\frac{2\left\|\nabla_f\left(\mathbf{x}^\star\right)|_{\mathcal{T}}\right\|_2}{\overline{\beta}_{4s}\left(\mathbf{x}^\star,\mathbf{b}\right)}\\ &+\left(1+\frac{2\overline{\gamma}_{4s}\left(\mathbf{x}^\star,\mathbf{b}\right)}{\overline{\beta}_{4s}\left(\mathbf{x}^\star,\mathbf{b}\right)}\right)\frac{\left\|\nabla_f\left(\mathbf{x}^\star\right)|_{\mathcal{R}\backslash\mathcal{Z}}\right\|_2+\left\|\nabla_f\left(\mathbf{x}^\star\right)|_{\mathcal{Z}\backslash\mathcal{R}}\right\|_2}{\overline{\beta}_{2s}\left(\widehat{\mathbf{x}},\mathbf{x}^\star\right)}.\end{aligned}$$

Proof Since $\mathcal{T}^c\subseteq\mathcal{Z}^c$ we have $\|\mathbf{x}^\star|_{\mathcal{T}^c}\|_2=\|\left(\widehat{\mathbf{x}}-\mathbf{x}^\star\right)|_{\mathcal{T}^c}\|_2\leq\|\left(\widehat{\mathbf{x}}-\mathbf{x}^\star\right)|_{\mathcal{Z}^c}\|_2$. Therefore, applying Lemma 4 yields

$$\|\mathbf{x}^\star|_{\mathcal{T}^c}\|_2\leq\frac{\overline{\gamma}_{2s}(\widehat{\mathbf{x}},\mathbf{x}^\star)+\overline{\gamma}_{4s}(\widehat{\mathbf{x}},\mathbf{x}^\star)}{\overline{\beta}_{2s}(\widehat{\mathbf{x}},\mathbf{x}^\star)}\|\widehat{\mathbf{x}}-\mathbf{x}^\star\|_2+\frac{\left\|\nabla_f(\mathbf{x}^\star)|_{\mathcal{R}\backslash\mathcal{Z}}\right\|_2+\left\|\nabla_f(\mathbf{x}^\star)|_{\mathcal{Z}\backslash\mathcal{R}}\right\|_2}{\overline{\beta}_{2s}(\widehat{\mathbf{x}},\mathbf{x}^\star)}. \tag{50}$$

Furthermore, as showed by (19) during the proof of Lemma 3, we again have

$$\|\mathbf{b}_s-\mathbf{x}^\star\|_2\leq2\|\mathbf{x}^\star|_{\mathcal{T}}-\mathbf{b}\|_2+\|\mathbf{x}^\star|_{\mathcal{T}^c}\|_2.$$

Hence, it follows from Lemma 5 that

$$\|\mathbf{b}_s-\mathbf{x}^\star\|_2\leq\frac{2\left\|\nabla_f\left(\mathbf{x}^\star\right)|_{\mathcal{T}}\right\|_2}{\overline{\beta}_{4s}\left(\mathbf{x}^\star,\mathbf{b}\right)}+\left(1+\frac{2\overline{\gamma}_{4s}\left(\mathbf{x}^\star,\mathbf{b}\right)}{\overline{\beta}_{4s}\left(\mathbf{x}^\star,\mathbf{b}\right)}\right)\|\mathbf{x}^\star|_{\mathcal{T}^c}\|_2. \tag{51}$$

Combining (50) and (51) yields

$$\begin{aligned}\|\mathbf{b}_s-\mathbf{x}^\star\|_2 \leq &\left(1+\frac{2\overline{\gamma}_{4s}\left(\mathbf{x}^\star,\mathbf{b}\right)}{\overline{\beta}_{4s}\left(\mathbf{x}^\star,\mathbf{b}\right)}\right)\frac{\overline{\gamma}_{2s}\left(\widehat{\mathbf{x}},\mathbf{x}^\star\right)+\overline{\gamma}_{4s}\left(\widehat{\mathbf{x}},\mathbf{x}^\star\right)}{\overline{\beta}_{2s}\left(\widehat{\mathbf{x}},\mathbf{x}^\star\right)}\|\widehat{\mathbf{x}}-\mathbf{x}^\star\|_2+\frac{2\left\|\nabla_f\left(\mathbf{x}^\star\right)|_{\mathcal{T}}\right\|_2}{\overline{\beta}_{4s}\left(\mathbf{x}^\star,\mathbf{b}\right)}\\ &+\left(1+\frac{2\overline{\gamma}_{4s}\left(\mathbf{x}^\star,\mathbf{b}\right)}{\overline{\beta}_{4s}\left(\mathbf{x}^\star,\mathbf{b}\right)}\right)\frac{\left\|\nabla_f\left(\mathbf{x}^\star\right)|_{\mathcal{R}\backslash\mathcal{Z}}\right\|_2+\left\|\nabla_f\left(\mathbf{x}^\star\right)|_{\mathcal{Z}\backslash\mathcal{R}}\right\|_2}{\overline{\beta}_{2s}\left(\widehat{\mathbf{x}},\mathbf{x}^\star\right)}.\end{aligned}$$

■

***Proof of Theorem* 2.** Let the vectors involved in the j-th iteration of the algorithm be denoted by superscript (j). Given that $\mu_{4s}\leq\frac{3+\sqrt{3}}{4}$ we have

$$\frac{\overline{\gamma}_{4s}\left(\widehat{\mathbf{x}}^{(j)},\mathbf{x}^\star\right)}{\overline{\beta}_{4s}\left(\widehat{\mathbf{x}}^{(j)},\mathbf{x}^\star\right)}\leq\frac{\sqrt{3}-1}{4}\quad\text{and}\quad1+\frac{2\overline{\gamma}_{4s}\left(\mathbf{x}^\star,\mathbf{b}^{(j)}\right)}{\overline{\beta}_{4s}\left(\mathbf{x}^\star,\mathbf{b}^{(j)}\right)}\leq\frac{1+\sqrt{3}}{2},$$

that yield,

$$\begin{aligned}\left(1+\frac{2\overline{\gamma}_{4s}\left(\mathbf{x}^\star,\mathbf{b}\right)}{\overline{\beta}_{4s}\left(\mathbf{x}^\star,\mathbf{b}\right)}\right)\frac{\overline{\gamma}_{2s}\left(\widehat{\mathbf{x}}^{(j)},\mathbf{x}^\star\right)+\overline{\gamma}_{4s}\left(\widehat{\mathbf{x}}^{(j)},\mathbf{x}^\star\right)}{\overline{\beta}_{2s}\left(\widehat{\mathbf{x}}^{(j)},\mathbf{x}^\star\right)} &\leq\frac{1+\sqrt{3}}{2}\times\frac{2\overline{\gamma}_{4s}\left(\widehat{\mathbf{x}}^{(j)},\mathbf{x}^\star\right)}{\overline{\beta}_{4s}\left(\widehat{\mathbf{x}}^{(j)},\mathbf{x}^\star\right)}\\ &\leq\frac{1+\sqrt{3}}{2}\times\frac{\sqrt{3}-1}{2}\\ &=\frac{1}{2}.\end{aligned}$$

Therefore, it follows from Lemma 6 that

$$\left\|\widehat{\mathbf{x}}^{(j+1)}-\mathbf{x}^{\star}\right\|_2 \leq \frac{1}{2}\left\|\widehat{\mathbf{x}}^{(j)}-\mathbf{x}^{\star}\right\|_2+\frac{3+\sqrt{3}}{\varepsilon}\left\|\nabla_f\left(\mathbf{x}^{\star}\right)\big|_I\right\|_2.$$

Applying this inequality recursively for $j=0,1,\cdots,i-1$ then yields

$$\|\widehat{\mathbf{x}}-\mathbf{x}^{\star}\|_2 \leq 2^{-i}\|\mathbf{x}^{\star}\|_2+\frac{6+2\sqrt{3}}{\varepsilon}\left\|\nabla_f\left(\mathbf{x}^{\star}\right)\big|_I\right\|_2,$$

which is the desired result. ■

References

A. Agarwal, S. Negahban, and M. Wainwright. Fast global convergence rates of gradient methods for high-dimensional statistical recovery. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 37–45. 2010. long version available at `arXiv:1104.4824v1 [stat.ML]`.

S. Bahmani, P. Boufounos, and B. Raj. Greedy sparsity-constrained optimization. In *Conference Record of the Forty-Fifth Asilomar Conference on Signals, Systems, and Computers*, pages 1148–1152, 2011.

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

P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

T. Blumensath. Compressed sensing with nonlinear observations. Preprint, 2010. URL `http://users.fmrib.ox.ac.uk/~tblumens/papers/B_Nonlinear.pdf`.

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

F. Bunea. Honest variable selection in linear and logistic regression models via ℓ_1 and $\ell_1+\ell_2$ penalization. *Electronic Journal of Statistics*, 2:1153–1194, 2008.

E. J. Candès. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, 346(9-10):589–592, 2008.

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

V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849, Dec. 2012.

A. Cohen, W. Dahmen, and R. DeVore. Compressed sensing and best k-term approximation. *American Mathematical Society*, 22(1):211–231, 2009.

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

A. J. Dobson and A. Barnett. *An Introduction to Generalized Linear Models*. Chapman and Hall/CRC, 3rd edition, May 2008.

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

D. L. Donoho and M. Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization. *Proceedings of the National Academy of Sciences*, 100(5):2197–2202, 2003.

D. L. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Transactions on Information Theory*, 47(7):2845–2862, 2001.

S. Foucart. Sparse recovery algorithms: sufficient conditions in terms of restricted isometry constants. In *Approximation Theory XIII: San Antonio 2010*, volume 13 of *Springer Proceedings in Mathematics*, pages 65–77. Springer New York, 2012.

J. H. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2 2010. ISSN 1548-7660. Software available online at `http://www-stat.stanford.edu/∼tibs/glmnet-matlab/`.

I. Guyon, S. Gunn, A. Ben-Hur, and G. Dror. Result analysis of the NIPS 2003 feature selection challenge. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 545–552. MIT-Press, Cambridge, MA, 2005.

J. D. Hamilton. *Time Series Analysis*. Princeton University Press, Princeton, NJ, 1994.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Verlag, 2009.

A. Jalali, C. C. Johnson, and P. K. Ravikumar. On learning discrete graphical models using greedy methods. In J. Shawe-Taylor, R. S. Zemel, P. Bartlett, F. C. N. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 1935–1943. 2011.

S. M. Kakade, O. Shamir, K. Sridharan, and A. Tewari. Learning exponential families in high-dimensions: strong convexity and sparsity. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, volume 9 of *JMLR W&CP*, pages 381–388, 2010.

J. Liu, J. Chen, and J. Ye. Large-scale sparse logistic regression. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 547–556, New York, NY, USA, 2009. ACM.

A. Lozano, G. Swirszcz, and N. Abe. Group orthogonal matching pursuit for logistic regression. In G. Gordon, D. Dunson, and M. Dudik, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *JMLR W&CP*, pages 452–460, 2011.

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

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

S. Negahban, P. Ravikumar, M. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1348–1356. 2009. long version available at `arXiv:1010.2731v1 [math.ST]`.

Y. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, Jan. 2012.

Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad. Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition. In *Conference Record of the Twenty-Seventh Asilomar Conference on Signals, Systems and Computers*, volume 1, pages 40–44, 1993.

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

A. Tewari, P. K. Ravikumar, and I. S. Dhillon. Greedy algorithms for structurally constrained high dimensional problems. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 882–890. 2011.

J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, Aug. 2012.

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

S. A. van de Geer. *Empirical Processes in M-estimation*. Cambridge University Press, Cambridge, UK, 2000.

S. A. van de Geer. High-dimensional generalized linear models and the Lasso. *The Annals of Statistics*, 36(2):614–645, 2008.

V. Vapnik. *Statistical Learning Theory*. Wiley, New York, NY, 1998.

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