Algorithms for Sparsity-Constrained Optimization

Submitted in partial fulfillment of the requirements for
the degree of
Doctor of Philosophy
in
Electrical and Computer Engineering

Sohail Bahmani

B.S., Electrical Engineering, Sharif University of Technology, Iran
M.A.Sc., Engineering Science, Simon Fraser University, Canada

Carnegie Mellon University
Pittsburgh, PA

February, 2013

تقدیم به پدر و مادرم ...

To my parents ...

Acknowledgements

I would like to thank Professor Bhiksha Raj, my advisor, for his continuous support and encouragement during my studies at Carnegie Mellon University. He made every effort to allow me to achieve my goals in research during the course of the PhD studies. I would also like to thank Dr. Petros T. Boufounos for his insightful comments that helped me improve the quality of my work during our collaboration and for serving in my thesis defense committee. I would like to thank Professor José M. F. Moura and Professor Soummya Kar, who also served in the defense committee, for their enlightening advice and comments on my thesis.

Above all, I would like to express my sincere gratitude for my parents who supported me throughout my life in every aspect. I especially thank my mother for giving me motivation and hope that helped me endure and overcome difficulties.

I was partially supported by the John and Claire Bertucci fellowship which I am grateful for. This work is partially supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20065. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI/NBC, or the U.S. Government.

Abstract

Sparsity-constrained optimization has wide applicability in machine learning, statistics, and signal processing problems such as feature selection and Compressive 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. We propose a greedy algorithm, Gradient Support Pursuit (GraSP), to approximate sparse minima of cost functions of arbitrary form. Should the cost function have a "well-behaved" second order variation over the sparse subspaces, we show that 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. We also formulate the 1-bit Compressed Sensing problem as a sparsity-constrained optimization with non-quadratic objective. We show through numerical simulations that the GraSP algorithm with slight modification can show better performance compared to existing algorithms.

Moreover, we study the structured sparsity estimation problems that involve nonlinear statistical models. Previously, several methods have been proposed for these problems

using convex relaxation techniques. These methods usually require a carefully tuned regularization parameter, often a cumbersome or heuristic exercise. Furthermore, the estimate that these methods produce might not belong to the desired sparsity model, albeit accurately approximating the true parameter. Therefore, greedy-type algorithms could be more desirable in estimating structured-sparse parameters. So far, these greedy methods have mostly focused on linear statistical models. We study a non-convex Projected Gradient Descent (PGD) for estimation of parameters with structured sparsity. Similar to the requirements for GraSP, if the cost function has proper second-order variation over the structured subspaces, the PGD algorithm converges to the desired minimizer up to an approximation error. As an example we elaborate on application of the main results to estimation in Generalized Linear Models.

Furthermore, we study the performance of the PGD algorithm for ℓ_p-constrained least squares problems that arise in of Compressed Sensing. Relying on the well-known Restricted Isometry Property, we provide convergence guarantees for this algorithm for the entire range of $0 \leq p \leq 1$, that include and generalize the existing results for the Iterative Hard Thresholding algorithm and provide a new accuracy guarantee for the Iterative Soft Thresholding algorithm as special cases. Our results suggest that in this group of algorithms, as p increases from zero to one, conditions required to guarantee accuracy become stricter and robustness to noise deteriorates.

Contents

List of Algorithms

List of Figures

List of Tables

Notations

$[n]$	the set $\{1, 2, \ldots, n\}$ for any $n \in \mathbb{N}$
$\mathcal{I}$	calligraphic letters denote sets unless stated otherwise
$\mathcal{I}^{\mathrm{c}}$	complement of set $\mathcal{I}$
$2^{\mathcal{I}}$	the set of all subsets (i.e., the powerset) of $\mathcal{I}$
$\mathbf{v}$	bold face small letters denote column vectors
$\|\mathbf{v}\|_0$	the "ℓ_0-norm" of vector $\mathbf{v}$ that merely counts its nonzero entries[1]
$\|\mathbf{v}\|_q$	the ℓ_q-norm of vector $\mathbf{v} \in \mathbb{C}^b$, that is, $\left(\sum_{i=1}^{b} \lvert v_i\rvert^q\right)^{1/q}$, for a real number $q > 0$[fn:quasinorm]
$\mathbf{v}\|_{\mathcal{I}}$	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 $\mathcal{I}^{\mathrm{c}}$ where it is zero
$\mathbf{v}_r$	the best r-term approximation of vector $\mathbf{v}$, unless stated otherwise
$\operatorname{supp}(\mathbf{v})$	the support set (i.e., indices of the non-zero entries) of $\mathbf{v}$
$\mathbf{M}$	bold face capital letters denote matrices
$\mathbf{M}^{\mathrm{T}}$,$\mathbf{M}^{\mathsf{H}}$, $\mathbf{M}^{\dagger}$	transpose, Hermitian transpose, and pseudo-inverse of matrix $\mathbf{M}$, respectively

[1] The term "norm" is used for convenience throughout the thesis. In fact, the ℓ_0 functional violates the positive scalability property of the norms and the ℓ_p functionals with $p \in (0, 1)$ are merely *quasi-norms*.

$\mathbf{M}_{\mathcal{I}}$	restriction of matrix $\mathbf{M}$ to the columns enumerated by $\mathcal{I}$
$\|\mathbf{M}\|$	the operator norm of matrix $\mathbf{M}$ which is equal to $\sqrt{\lambda_{\max}\left(\mathbf{M}^{\mathrm{T}}\mathbf{M}\right)}$
$\mathbf{M} \succcurlyeq \mathbf{M}'$	$\mathbf{M} - \mathbf{M}'$ is positive semidefinite
$\mathbf{I}$	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
$\nabla^2 f\left(\cdot\right)$, $\nabla^2_{\mathcal{I}} f\left(\cdot\right)$	The former denotes the Hessian of the function f, and the latter denotes the the Hessian restricted to rows and columns indexed by $\mathcal{I}$
$\left(x\right)_+$	Positive part of x
$\mathrm{Arg}\left(x\right)$	Argument (phase) of a complex number x
$\Re\left[x\right]$	Real part of a complex number x

Chapter 1

Introduction

Applications that require analysis of high-dimensional data has grown significantly during the past decade. In many of these applications, such as bioinformatics, social networking, and mathematical finance, dimensionality of the data is usually much larger than the number of samples or observations acquired. Therefore statistical inference or data processing would be ill-posed for these *underdetermined* problems. Fortunately, in some applications the data is known *a priori* to have an underlying structure that can be exploited to compensate the deficit of observations. These structures often characterize the signals by a low-dimensional manifold, e.g. the set of *sparse* vectors or the set of *low-rank* matrices, embedded in the high-dimensional ambient space. One of the main goals of high-dimensional data analysis is to design accurate, robust, and computationally efficient algorithms for estimation of these structured signals in underdetermined regimes.

In signal processing, the data acquisition methods are traditionally devised based on the Shannon-Nyquist sampling theorem which ties the number of required observations to the largest frequency content of the signal. However, these acquisition methods deem inefficient and costly for very-high-frequency signals. The drawbacks are particularly pro-

nounced in applications where the signal of interest is sparse with respect to some known basis or frame. To break the limitations of traditional signal acquisition, *Compressed Sensing* (CS)(Donoho, 2006; Candès and Tao, 2006) introduced a novel approach for accurate reconstruction of sparse signals from a relatively small number of linear observations. In addition to the data sampling problem, the mathematical formulation of CS is employed to address a variety of other problems in different fields. For instance, the fact that CS operates at low sampling rates allows shorter acquisition time; a feature that is highly desirable in applications such as tomography and magnetic resonance imaging (MRI) where traditional methods are time consuming or need longer exposure to hazardous radiation.

Sparse linear regression problems studied in statistics and machine learning are similar to CS. These problems usually describe *feature* and *variable selection* problems in high-dimensional linear models. However, the linear models in these problems are slightly different as they are dictated by the observed data; a fact that does not permit many of the assumptions considered about the measurement vectors in CS. Nevertheless, sparse linear regression problems and the algorithms developed to solve them are also studied extensively.

While linear models are widely used to analyze data and systems in variety of fields, there are many applications where non-linear models are better suited. For example, in binary classification problems the relation between the target parameter, data points, and their associated binary labels is generally determined by a non-linear equation. A typical application is the gene selection where among thousands of genes a few genes that are likely to cause a specific type of cancer must be detected based on their *expression level* in tissue samples (Lazar et al., 2012). Also there are variety of *inverse problems* in optics, imaging, and tomography where the observations do not exhibit a linear relation with the underlying signal (Kolehmainen et al., 2000; Boas et al., 2001; Borcea, 2002; Shecht-

man et al., 2011b,a). Despite broad application of non-linear models in high-dimensional regime, they have received relatively less attention compared to their linear counterparts.

1.1 Contributions

The material presented in this thesis consists mostly of our work published in (Bahmani et al., 2011; Bahmani and Raj, 2013; Bahmani et al., 2013, 2012). The main theme of this thesis is sparsity-constrained optimization that arise in certain statistical estimation problems. We present a greedy approximate algorithm for minimization of an objective function subject to sparsity of the optimization variable. To prove accuracy of the proposed algorithm we introduce a few sufficient conditions some of which are shown to hold for certain families of objective functions. We also show how a variant of the proposed algorithm can be applied to the problem of 1-bit Compressed Sensing. We further extend the results by studying minimization of an objective subject to structured-sparsity of the optimization variable. Under sufficient conditions similar to those mentioned above, we prove accuracy of non-convex Projected Gradient Descent algorithm for estimation of parameters with structured sparsity.

In a separate line of work, we also study the problem of ℓ_p-constrained least squares, one of the non-convex formulations of CS. Assuming that one can project any point onto a given ℓ_p-ball, we show that non-convex Projected Gradient Descent converges to the true sparse signal up to an approximation error. We further characterize the necessary conditions for projection of a point an a given ℓ_p-ball.

1.2 Thesis outline

The rest of the thesis is organized as follows. In Chapter 2 we briefly review CS and sparse linear regression. Furthermore, we motivate the main subject of the thesis by describing some applications where non-linear models need to be considered. In Chapter 3 we introduce a non-convex greedy algorithm called GraSP for approximating sparsity-constrained optimization and prove its accuracy under appropriate conditions. The theoretical analysis of this chapter is provided in Appendix A. We cast 1-bit CS as a sparsity-constrained optimization in Chapter 4 and numerically compare the performance of GraSP with the prior work on 1-bit CS. Some of the technical details of this chapter are subsumed to Appendix B. We also study minimization of an objective function subject to *model-based sparsity* constraints in Chapter 5 and consider non-convex Projected Gradient Descent as the approximate algorithm. Derivations of the corresponding accuracy guarantees are provided in Appendix C. We then study the non-convex ℓ_p-constrained least squares problems by analyzing performance of Projected Gradient Descent methods in Chapter 6. The mathematical derivations for this chapter are gathered in Appendix 6. Finally, we conclude the thesis in Chapter 7.

Chapter 2

Preliminaries

2.1 Sparse Linear Regression and Compressed Sensing

Least squares problems occur in various signal processing and statistical inference applications. In these problems the relation between the vector of noisy observations $\mathbf{y} \in \mathbb{R}^m$ and the unknown parameter or signal $\mathbf{x}^\star \in \mathbb{R}^n$ is governed by a linear equation of the form

$$\mathbf{y} = \mathbf{A}\mathbf{x}^\star + \mathbf{e}, \tag{2.1}$$

where $\mathbf{A} \in \mathbb{R}^{m \times n}$ is a matrix that may model a linear system or simply contains a set of collected data. The vector $\mathbf{e} \in \mathbb{R}^m$ represents the additive observation noise. Estimating $\mathbf{x}^\star$ from the observation vector $\mathbf{y}$ is achieved by finding the vector $\mathbf{x}$ that minimizes the squared error $\|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2{}^2$. This least squares approach, however, is well-posed only if the nullspace of matrix $\mathbf{A}$ merely contains the zero vector. The cases in which the nullspace is greater than the singleton $\{\mathbf{0}\}$, as in *underdetermined* scenarios ($m < n$), are more relevant in a variety of applications. To enforce unique least squares solutions in these cases, it becomes necessary to have some prior information about the structure of $\mathbf{x}^\star$.

One of the structural characteristics that describe parameters and signals of interest in a wide range of applications from medical imaging to astronomy is *sparsity*. Study of high-dimensional linear inference problems with sparse parameter has gained significant attention since the introduction of Compressed Sensing, also known as *Compressive Sampling*, (CS) (Donoho, 2006; Candès and Tao, 2006). In standard CS problems the aim is to estimate a sparse vector $\mathbf{x}^\star$ from linear measurements. In the absence of noise (i.e., when $\mathbf{e} = \mathbf{0}$), $\mathbf{x}^\star$ can be determined uniquely form the observation vector $\mathbf{y} = \mathbf{A}\mathbf{x}^\star$ provided that $\text{spark}\,(\mathbf{A}) > 2\|\mathbf{x}^\star\|_0$ (i.e., every $2\|\mathbf{x}^\star\|_0$ columns of $\mathbf{A}$ are linearly independent) (Donoho and Elad, 2003). Then the ideal estimation procedure would be to find the sparsest vector $\mathbf{x}$ that incurs no residual error (i.e., $\|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2 = 0$). This ideal estimation method can be extended to the case of noisy observations as well. Formally, the vector $\mathbf{x}^\star$ can be estimated by solving the ℓ_0-minimization

$$\widehat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\mathbf{x}\|_0 \quad \text{s.t. } \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2 \leq \varepsilon, \tag{2.2}$$

where ε is a given upper bound for $\|\mathbf{e}\|_2$ (Candès et al., 2006). Unfortunately, the ideal solver (2.2) 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 basically returns the solution to the convex optimization problem

$$\widehat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\mathbf{x}\|_1 \quad \text{s.t. } \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2 \leq \varepsilon, \tag{2.3}$$

The required conditions for approximate equivalence of (2.2) and (2.3), however, generally hold only if measurements are collected at a higher rate. Ideally, one merely needs $m = O(s)$ measurements to estimate $\mathbf{x}^\star$, but $m = O(s \log n/s)$ measurements are necessary for

the accuracy of (2.3) to be guaranteed.

The convex program (2.3) 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., Figueiredo et al., 2007; Hale et al., 2008; Beck and Teboulle, 2009; Wen et al., 2010; 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

$$\widehat{\mathbf{x}} = \arg\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 \quad \text{s.t. } \|\mathbf{x}\|_0 \leq s. \tag{2.4}$$

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 (2.3) 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).

Formulation of sparse linear regression problems as well as algorithms used to solve them are virtually identical to CS. However, these problems that are usually studied in statistics and machine learning, have a set-up that distinguishes them from the CS problems. The sensing or sampling problems addressed by CS often do not impose strong restrictions on the choice of the measurement matrix. Matrices drawn from certain ensembles of random matrices (e.g., Gaussian, Rademacher, partial Fourier, etc) can be chosen as the measurement matrix (Candès and Tao, 2006). These types of random matrices allow us to guarantee the required conditions such as RIP, at least in the probabilistic sense. However, the analog of the measurement matrix in sparse linear regression, the *design matrix*, is often dictated by the data under study. In general the entries of the design matrix have unknown distributions and are possibly dependent. In certain scenarios the independence of observations/measurements may not hold either. While it is inevitable to make assumptions about the design matrix for the purpose of theoretical analysis, the considered assumptions are usually weaker compared to the CS assumptions. Consequently, the analysis of sparse linear inference problems is more challenging than in CS problems.

2.2 Nonlinear Inference Problems

To motivate the need for generalization of CS, in this section we describe a few problems and models which involve non-linear observations.

2.2.1 Generalized Linear Models

Generalized Linear Models (GLMs) are among the most commonly used models for parametric estimation in statistics (Dobson and Barnett, 2008). Linear, logistic, Poisson, and gamma models used in corresponding regression problems all belong to the family of GLMs. Because the parameter and the data samples in GLMs are mixed in a linear form, these models are considered among linear models in statistics and machine learning literature. However, as will be seen below, in GLMs the relation between the response variable and the parameters is in general nonlinear.

Given a vector of covariates (i.e., data sample) $\mathbf{a} \in \mathcal{X} \subseteq \mathbb{R}^n$ and a true parameter $\mathbf{x}^\star \in \mathbb{R}^n$, the response variable $y \in \mathcal{Y} \subseteq \mathbb{R}$ in canonical GLMs is assumed to follow an exponential family conditional distribution: $y \mid \mathbf{a}; \mathbf{x}^\star \sim Z(y) \exp\left(y \langle \mathbf{a}, \mathbf{x}^\star \rangle - \psi\left(\langle \mathbf{a}, \mathbf{x}^\star \rangle\right)\right)$, where $Z(y)$ is a positive function, and $\psi : \mathbb{R} \mapsto \mathbb{R}$ is the *log-partition function* that satisfies $\psi(t) = \log \int_{\mathcal{Y}} Z(y) \exp(ty)\, \mathrm{d}y$ for all $t \in \mathbb{R}$. Examples of the log-partition function, which is always convex, include but are not limited to $\psi_{\text{lin}}(t) = t^2/2\sigma^2$, $\psi_{\log}(t) = \log(1 + \exp(t))$, and $\psi_{\text{Pois}}(t) = \exp(t)$ corresponding to linear, logistic, and Poisson models, respectively.

Suppose that m iid covariate-response pairs $\{(\mathbf{a}_i, y_i)\}_{i=1}^m$ are observed in a GLM. As usual, it is assumed that $\mathbf{a}_i$'s do not depend on the true parameter. The joint likelihood function of the observation at parameter $\mathbf{x}$ can be written as $\prod_{i=1}^m p(\mathbf{a}_i)\, p(y_i \mid \mathbf{a}_i; \mathbf{x})$ where $p(y_i \mid \mathbf{a}_i; \mathbf{x})$ is the exponential family distribution mentioned above. In the Maximum Likelihood Estimation (MLE) framework the negative log likelihood is used as a measure of the discrepancy between the true parameter $\mathbf{x}^\star$ and an estimate $\mathbf{x}$ based on the observations.

Because $p(\mathbf{a}_i)$'s do not depend on $\mathbf{x}$ the corresponding terms can be simply ignored. Formally, the average of negative log conditional likelihoods is considered as the empirical loss

$$f(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m} \psi\left(\langle \mathbf{a}_i, \mathbf{x}\rangle\right) - y_i \langle \mathbf{a}_i, \mathbf{x}\rangle ,$$

and the MLE is performed by minimizing $f(\mathbf{x})$ over the set of feasible $\mathbf{x}$. The constant c and $Z(y)$ that appear in the distribution are disregarded as they have no effect in the outcome. We will use the logistic model, a special case of GLMs, in Chapters 3 and 5 as examples where our algorithms apply.

2.2.2 1-bit Compressed Sensing

As mentioned above, the ideal CS formulation allows accurate estimation of sparse signals from relatively small number of linear measurements. However, sometimes certain practical limitations impose non-ideal conditions that must be addressed in order to apply the CS framework. One of these limitations is the fact that in digital signal processing systems the signals and measurements have quantized values. Motivated by this problem, researchers have studied performance of CS with quantized measurements. Of particular interest has been the problem of 1-bit Compressed Sensing (Boufounos and Baraniuk, 2008), in which the CS linear measurements are quantized down to one bit that represents their sign. Namely, for a signal $\mathbf{x}^\star$ and measurement vector $\mathbf{a}$ the observed measurement in 1-bit CS is given by $y = \operatorname{sgn}\left(\langle \mathbf{a}, \mathbf{x}^\star\rangle + e\right)$ where e is an additive noise. As can be seen, the observations and the signal are related by a nonlinear transform. In Chapter 4 we will explain how the problem of estimating $\mathbf{x}^\star$ from a collection of 1-bit measurements can be cast as a sparsity-constrained optimization.

2.2.3 Phase Retrieval

One of the common non-linear inverse problems that arise in applications such as optics and imaging is the problem of *phase retrieval*. In these applications the observations of the object of interest are in the form of phaseless linear measurements. In general, reconstruction of the signal is not possible in these scenarios. However, if the signal is known to be sparse *a priori* then accurate reconstruction can be achieved up to a unit-modulus factor. In particular, *Quadratic Compressed Sensing* is studied in (Shechtman et al., 2011b,a) for phase retrieval problems in sub-wavelength imaging. Using convex relaxation it is shown that the estimator can be formulated as a solution to a Semi-Definite Program (SDP) dubbed *PhaseLift* (Candès et al., 2012; Candès and Li, 2012; Li and Voroninski, 2012).

Chapter 3

Sparsity-Constrained Optimization

3.1 Background

Theoretical and application aspects of sparse estimation in linear models have been studied extensively in areas such as signal processing, machine learning, and statistics. The sparse linear regression and CS algorithms attempt to provide a sparse vector whose consistency with the acquired data is usually measured by the squared error. 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. 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 3.2. 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 CS 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. Also, (Beck and Eldar, 2012) studies the problem of minimizing a generic objective function subject to sparsity constraint from the optimization perspective. By analyzing necessary optimality conditions for the sparse minimizer, a few iterative algorithms are proposed in (Beck and Eldar, 2012) that converge to the sparse minimizer, should the objective satisfies some regularity conditions. Furthermore, 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.

In this chapter we propose a non-convex 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 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 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.

3.2 Convex Methods and Their Required Conditions

The existing studies on sparsity-constrained optimization 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-norm as the means to induce sparsity in the solution. With $f(\mathbf{x})$ denoting the considered loss function and for proper values of $\lambda \geq 0$ and $R \geq 0$, these works study either the accuracy of the ℓ_1-regularized estimator given by

$$\arg\min_{\mathbf{x}} f(\mathbf{x}) + \lambda \|\mathbf{x}\|_1,$$

or that of the ℓ_1-constrained estimator given by

$$\arg\min_{\mathbf{x}} f(\mathbf{x})$$
$$\text{subject to} \|\mathbf{x}\|_1 \leq R.$$

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. Typically in parametric estimation problems a sample loss function $l(\mathbf{x}, \mathbf{a}, y)$ is associated with the covariate-response pair $(\mathbf{a}, y)$ and a parameter $\mathbf{x}$. Given m iid observations the empirical loss is formulated as $\widehat{L}_m(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m} l(\mathbf{x}, \mathbf{a}_i, y_i)$. The estimator under study is often the minimizer of the empirical loss, perhaps considering an extra regularization or constraint for the parameter $\mathbf{x}$. Furthermore, it is known that $\widehat{L}_m(\mathbf{x})$ as an empirical process is a good approximation of the expected loss $L(\mathbf{x}) = \mathbb{E}[l(\mathbf{x}, \mathbf{a}, y)]$ (see Vapnik, 1998, chap. 5 and van de Geer, 2000). Consequently, if for a valid choice of $\mathbf{x}^\star$ the required sufficient condition is not satisfied by $L(\mathbf{x})$, then in general it cannot be satisfied at the same $\mathbf{x}^\star$ by $\widehat{L}_m(\mathbf{x})$ either. In particular, 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 de-

sired 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.

3.3 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 chapter we focus on a generalization of (2.4) where a generic cost function replaces the squared error. Specifically, for the cost function $f : \mathbb{R}^n \mapsto \mathbb{R}$, it is desirable to approximate

$$\arg\min_{\mathbf{x}} f(\mathbf{x}) \quad \text{s.t. } \|\mathbf{x}\|_0 \leq s. \tag{3.1}$$

We propose the Gradient Support Pursuit (GraSP) algorithm, which is inspired by and generalizes the CoSaMP algorithm, to approximate the solution to (3.1) for a broader class of cost functions.

Of course, even for a simple quadratic objective, (3.1) 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 3.3.1. 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.

3.3.1 Algorithm Description

Algorithm 1: The GraSP algorithm

input : $f(\cdot)$ and s
output: $\widehat{\mathbf{x}}$

initialize: $\widehat{\mathbf{x}} = 0$
repeat
1. **compute local gradient:** $\mathbf{z} = \nabla f(\widehat{\mathbf{x}})$
2. **identify directions:** $\mathcal{Z} = \operatorname{supp}(\mathbf{z}_{2s})$
3. **merge supports:** $\mathcal{T} = \mathcal{Z} \cup \operatorname{supp}(\widehat{\mathbf{x}})$
4. **minimize over support:** $\mathbf{b} = \arg\min f(\mathbf{x})$ s.t. $\mathbf{x}|_{\mathcal{T}^c} = \mathbf{0}$
5. **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 3.3.2. 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, e.g., 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 chapter 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 $\mathbf{b}$ in line 5 of the algorithm, 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 $\mathbf{b}$ in line 4, we can set $\mathbf{b}|_{\mathcal{T}^c} = \mathbf{0}$ and take a restricted Newton step as

$$\mathbf{b}|_{\mathcal{T}} = \widehat{\mathbf{x}}|_{\mathcal{T}} - \kappa \left(\nabla^2_{\mathcal{T}} f\left(\widehat{\mathbf{x}}\right)\right)^{-1} \widehat{\mathbf{x}}|_{\mathcal{T}},$$

where $\kappa > 0$ is a step-size. Of course, here we are assuming that the restricted Hessian, $\nabla^2_{\mathcal{T}} f\left(\widehat{\mathbf{x}}\right)$, is invertible.

3. *Restricted Gradient Descent*: The minimization step in line 4 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

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

Since $\mathcal{T}$ contains both the support set of $\widehat{\mathbf{x}}$ and the $2s$-largest entries of $\nabla f\left(\widehat{\mathbf{x}}\right)$, 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.

3.3.2 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 3.1 (Stable Restricted Hessian)**.** Suppose that f is a twice continuously differentiable function whose Hessian is denoted by $\nabla^2 f(\cdot)$. Furthermore, let

$$A_k(\mathbf{x}) = \sup\left\{\boldsymbol{\Delta}^{\mathrm{T}}\nabla^2 f(\mathbf{x})\,\boldsymbol{\Delta} \;\middle|\; |\mathrm{supp}(\mathbf{x}) \cup \mathrm{supp}(\boldsymbol{\Delta})| \leq k, \|\boldsymbol{\Delta}\|_2 = 1\right\} \tag{3.2}$$

and

$$B_k(\mathbf{x}) = \inf\left\{\boldsymbol{\Delta}^{\mathrm{T}}\nabla^2 f(\mathbf{x})\,\boldsymbol{\Delta} \;\middle|\; |\mathrm{supp}(\mathbf{x}) \cup \mathrm{supp}(\boldsymbol{\Delta})| \leq k, \|\boldsymbol{\Delta}\|_2 = 1\right\}, \tag{3.3}$$

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 3.1. Since the Hessian of f is symmetric, an equivalent for Definition 3.1 is that a twice continuously differentiable function f has μ_k-SRH if the condition number of $\nabla^2_{\mathcal{K}} f(\mathbf{x})$ is not greater than μ_k for all k-sparse vectors $\mathbf{x}$ and sets $\mathcal{K} \subseteq [n]$ with $|\mathrm{supp}(\mathbf{x}) \cup \mathcal{K}| \leq k$.

In the special case when the cost function is the squared error as in (2.4), we can write $\nabla^2 f(\mathbf{x}) = \mathbf{A}^{\mathrm{T}}\mathbf{A}$ which is constant. The SRH condition then requires

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

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

Remark 3.2. 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 non-convex functions that have SRH.

Example 3.1. 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 $\nabla^2 f(\mathbf{x}) = \mathbf{Q}$. Therefore, (3.2) and (3.3) 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 3.2. 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., $\mathbf{P}_{[k+1]}^{\mathrm{T}}\nabla^2 f(\mathbf{x})\,\mathbf{P}_{[k+1]}$) 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 3.2 (Restricted Subgradient)**.** We say vector $\nabla_f(\mathbf{x})$ is a restricted subgradient of $f : \mathbb{R}^n \mapsto \mathbb{R}$ at point $\mathbf{x}$ if

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

holds for all k-sparse vectors $\boldsymbol{\Delta}$.

Remark 3.3. 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 3.4. 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 3.5. 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 $f : \mathbb{R}^n \mapsto \mathbb{R}$ between points $\mathbf{x}$ and $\mathbf{x}'$ where $\nabla_f\left(\cdot\right)$ gives a restricted subgradient of $f\left(\cdot\right)$.

Definition 3.3 (Stable Restricted Linearization)**.** Let $\mathbf{x}$ be a k-sparse vector in $\mathbb{R}^n$. For function $f : \mathbb{R}^n \mapsto \mathbb{R}$ we define the functions

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

and

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

respectively. Then $f\left(\cdot\right)$ 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 3.6. 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.

3.3.3 Main Theorems

Now we can state our main results regarding approximation of

$$\mathbf{x}^\star = \arg\min \; f(\mathbf{x}) \text{ s.t. } \|\mathbf{x}\|_0 \leq s, \tag{3.4}$$

using the GraSP algorithm.

Theorem 3.1. *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 $\epsilon > 0$ we have $\|\nabla f(\mathbf{x}^\star)|_{\mathcal{I}}\|_2 \leq \epsilon\, B_{4s}(\mathbf{x})$ for all $4s$-sparse $\mathbf{x}$, where $\mathcal{I}$ is the position of the $3s$ largest entries of $\nabla f(\mathbf{x}^\star)$ in magnitude.*

Then $\widehat{\mathbf{x}}^{(i)}$, the estimate at the i-th iteration, satisfies

$$\left\|\widehat{\mathbf{x}}^{(i)} - \mathbf{x}^\star\right\|_2 \leq 2^{-i}\|\mathbf{x}^\star\|_2 + \left(6 + 2\sqrt{3}\right)\epsilon.$$

Remark 3.7. 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 3.8. As the derivations required to prove Theorem 3.1 show, the provided accuracy guarantee holds for any s-sparse $\mathbf{x}^\star$, even if it does not obey (3.4). Obviously, for arbitrary choices of $\mathbf{x}^\star$, $\nabla f(\mathbf{x}^\star)|_{\mathcal{I}}$ may have a large norm that cannot be bounded properly which implies large values for ϵ and thus large approximation 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 (3.4). In these problems, the approximation error $\|\nabla f(\mathbf{x}^\star)|_{\mathcal{I}}\|_2$ 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 3.4.

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 3.1 for nonsmooth cost functions that satisfy the SRL condition.

Theorem 3.2. *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 3.3 there exists some $\epsilon > 0$ such that $\left\|\nabla_f(\mathbf{x}^\star)|_{\mathcal{I}}\right\|_2 \leq \epsilon\beta_{4s}(\mathbf{x})$ holds for all $4s$-sparse vectors $\mathbf{x}$, where $\mathcal{I}$ is the position of the $3s$ largest entries of $\nabla_f(\mathbf{x}^\star)$ in magnitude. Then $\widehat{\mathbf{x}}^{(i)}$, the estimate at the i-th iteration,*

satisfies

$$\left\|\widehat{\mathbf{x}}^{(i)} - \mathbf{x}^\star\right\|_2 \leq 2^{-i}\|\mathbf{x}^\star\|_2 + \left(6 + 2\sqrt{3}\right)\epsilon.$$

Remark 3.9. Should the SRH or SRL conditions hold for the objective function, it is straightforward to convert the *point accuracy* guarantees of Theorems 3.1 and 3.2, 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.

3.4 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 $\mathbf{a} \in \mathbb{R}^n$, and its associated label, represented by a random binary variable $y \in \{0, 1\}$, is determined by the conditional probability

$$\Pr\{y \mid \mathbf{a}; \mathbf{x}\} = \frac{\exp\left(y \left\langle \mathbf{a}, \mathbf{x} \right\rangle\right)}{1 + \exp\left(\left\langle \mathbf{a}, \mathbf{x} \right\rangle\right)}, \tag{3.5}$$

where $\mathbf{x}$ denotes a parameter vector. Then, for a set of m independently drawn data samples $\{(\mathbf{a}_i, y_i)\}_{i=1}^m$ 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,

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

It is well-known that $g(\cdot)$ is strictly convex for $n \leq m$ provided that the associated design matrix, $\mathbf{A} = [\mathbf{a}_1\ \mathbf{a}_2\ \ldots\ \mathbf{a}_m]^{\mathrm{T}}$, is full-rank. However, in many important applications (e.g., feature selection) the problem can be underdetermined (i.e., $m < n$). 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

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

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.

3.4.1 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

$$\nabla^2 g(\mathbf{x}) = \frac{1}{4m}\mathbf{A}^{\mathrm{T}}\mathbf{\Lambda}\mathbf{A},$$

where $\mathbf{\Lambda}$ is an $m \times m$ diagonal matrix whose diagonal entries are $\Lambda_{ii} = \operatorname{sech}^2 \frac{1}{2} \langle \mathbf{a}_i, \mathbf{x} \rangle$ with $\operatorname{sech}(\cdot)$ denoting the *hyperbolic secant* function. Note that $\mathbf{0} \preccurlyeq \nabla^2 g(\mathbf{x}) \preccurlyeq \frac{1}{4m} \mathbf{A}^{\mathrm{T}} \mathbf{A}$. Therefore, if $\nabla^2 f_\eta(\mathbf{x})$ denotes the Hessian of the ℓ_2-regularized logistic loss, we have

$$\forall \mathbf{x}, \mathbf{\Delta} \qquad \eta \|\mathbf{\Delta}\|_2^2 \leq \mathbf{\Delta}^{\mathrm{T}} \nabla^2 f_\eta(\mathbf{x}) \, \mathbf{\Delta} \leq \frac{1}{4m} \|\mathbf{A}\mathbf{\Delta}\|_2^2 + \eta \|\mathbf{\Delta}\|_2^2. \tag{3.6}$$

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

Theorem 3.3 (Matrix Chernoff (Tropp, 2012))**.** *Consider a finite sequence* $\{\mathbf{M}_i\}$ *of* $k \times k$*, independent, random, self-adjoint matrices that satisfy*

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

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

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

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 3.3 we need to make the following extra assumptions:

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

(i) we have $\left\|\mathbf{a}|_{\mathcal{J}}\right\|_2^2 \leq R$ 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

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

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

Corollary 3.1. *With the above assumptions, if*

$$m \geq R\left(\log k + k\left(1 + \log\frac{n}{k}\right) - \log\varepsilon\right) / \left(\widetilde{\theta}h\left(\tau\right)\right)$$

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 $\mu_k \leq 1 + \frac{1+\tau}{4\eta}\overline{\theta}$.

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

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

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

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

Hence, for any fixed index set $\mathcal{J}$ with $|\mathcal{J}| = k$ we may apply Theorem 3.3 for $\mathbf{M}_i = \mathbf{M}_i^{\mathcal{J}}$, $\theta_{\max} = m\theta_{\max}^{\mathcal{J}}$, and $\tau > 0$ to obtain

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

Furthermore, we can write

$$\begin{aligned}
\Pr\left\{\lambda_{\max}\left(\mathbf{A}_{\mathcal{J}}^{\mathrm{T}}\mathbf{A}_{\mathcal{J}}\right) \geq (1+\tau)\, m\overline{\theta}\right\} &= \Pr\left\{\lambda_{\max}\left(\sum_{i=1}^{m}\mathbf{M}_i^{\mathcal{J}}\right) \geq (1+\tau)\, m\overline{\theta}\right\} \\
&\leq \Pr\left\{\lambda_{\max}\left(\sum_{i=1}^{m}\mathbf{M}_i^{\mathcal{J}}\right) \geq (1+\tau)\, m\theta_{\max}^{\mathcal{J}}\right\} \\
&\leq k\exp\left(-\frac{m\theta_{\max}^{\mathcal{J}} h\left(\tau\right)}{R}\right) \\
&\leq k\exp\left(-\frac{m\widetilde{\theta} h\left(\tau\right)}{R}\right). \qquad (3.7)
\end{aligned}$$

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

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

Therefore, for $\varepsilon \in (0,1)$ and $m \geq R\left(\log k + k\left(1+\log\frac{n}{k}\right) - \log\varepsilon\right) / \left(\widetilde{\theta} h\left(\tau\right)\right)$ it follows from

(3.6) that for any $\mathbf{x}$ and any k-sparse $\mathbf{\Delta}$,

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

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 3.10. One implication of this result is that for a regime in which k and n grow sufficiently large while $\frac{n}{k}$ remains constant one can achieve small failure rates provided that $m = \Omega\left(Rk\log\frac{n}{k}\right)$. 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 $m = \Omega\left(k^2\log\frac{n}{k}\right)$ as the sufficient number of observations. This bound is a consequence of using Theorem 3.3, but to the best of our knowledge, other results regarding the extreme eigenvalues of the average of independent random PSD matrices also yield an m of the same order. If matrix $\mathbf{A}$ has certain additional properties (e.g., independent and sub-Gaussian entries), however, a better rate of $m = \Omega\left(k\log\frac{n}{k}\right)$ can be achieved without using the techniques mentioned above.

Remark 3.11. 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.

3.4.2 Bounding the approximation error

We are going to bound $\left\|\nabla f_\eta(\mathbf{x}^\star)|_{\mathcal{I}}\right\|_2$ which controls the approximation error in the statement of Theorem 3.1. In the case of case of ℓ_2-regularized logistic loss considered in this section we have

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

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

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

where $\mathbf{v} = [v_1 \, v_2 \, \ldots \, v_m]^{\mathrm{T}}$. Note that v_i's are m 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

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

where $\sigma_v^2 = \mathbb{E}\left[v^2 \right]$ is the variance of v. Furthermore, using the logistic model (3.5) we can deduce

$$\begin{aligned}
\sigma_v^2 &= \mathbb{E}\left[v^2 \right] \\
&= \mathbb{E}\left[\mathbb{E}\left[v^2 \mid \mathbf{a} \right] \right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left(y - \mathbb{E}\left[y \mid \mathbf{a} \right] \right)^2 \mid \mathbf{a} \right] \right] \\
&= \mathbb{E}\left[\mathrm{var}\left(y \mid \mathbf{a} \right) \right] \\
&= \mathbb{E}\left[\frac{1}{1 + \exp\left(\langle \mathbf{a}, \mathbf{x}^\star \rangle \right)} \times \frac{\exp\left(\langle \mathbf{a}, \mathbf{x}^\star \rangle \right)}{1 + \exp\left(\langle \mathbf{a}, \mathbf{x}^\star \rangle \right)} \right] \quad \text{(because } y \mid \mathbf{a} \sim \text{Bernoulli as in (3.5))} \\
&= \mathbb{E}\left[\frac{1}{2 + \exp\left(\langle \mathbf{a}, \mathbf{x}^\star \rangle \right) + \exp\left(-\langle \mathbf{a}, \mathbf{x}^\star \rangle \right)} \right] \\
&\leq \frac{1}{4} \qquad \text{(because } \exp(t) + \exp(-t) \geq 2\text{).}
\end{aligned}$$

Therefore, we have $\frac{1}{m}\sum_{i=1}^{m} v_i^2 < \frac{1}{4}$ with high probability. As in the previous subsection one can also bound $\frac{1}{\sqrt{m}}\|\mathbf{A}_{\mathcal{I}}\| = \sqrt{\frac{1}{m}\lambda_{\max}\left(\mathbf{A}_{\mathcal{I}}^{\mathrm{T}}\mathbf{A}_{\mathcal{I}}\right)}$ using (3.7) with $k = |\mathcal{I}| = 3s$. Hence, with high probability we have

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

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

3.5 Simulations

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 the performance of GraSP for variable selection in the logistic model both on synthetic and real data.

Synthetic Data

In our simulations the sparse parameter of interest $\mathbf{x}^\star$ is a $n = 1000$ 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 inde-

pendent instance of the random vector $\mathbf{a} = [a_1, a_2, \ldots, a_n]^{\mathrm{T}}$ generated by an autoregressive process (Hamilton, 1994) determined by

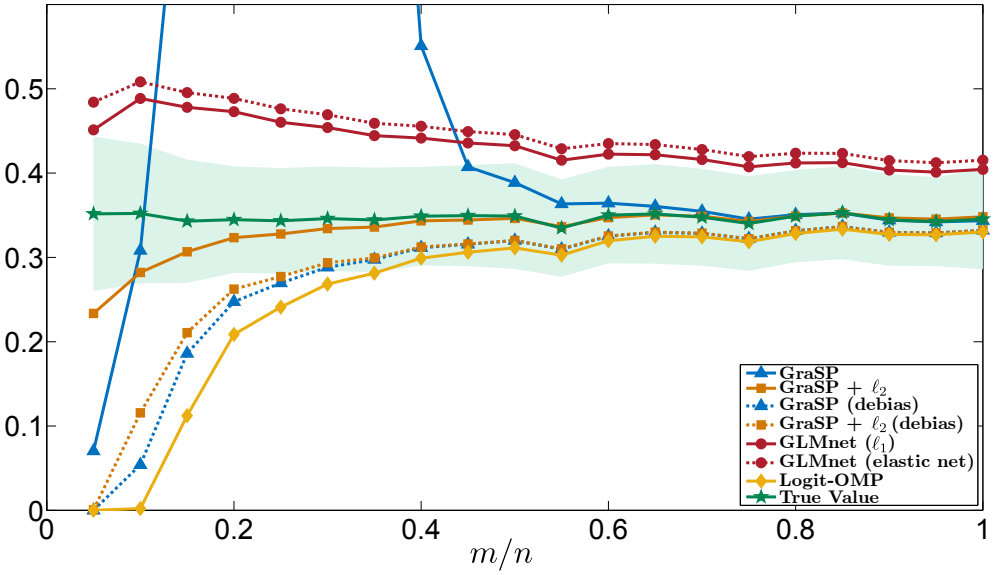
$$a_{j+1} = \rho a_j + \sqrt{1-\rho^2} z_j, \qquad \text{for all } j \in [p-1]$$

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 $j \in [n]$. The data labels, $y \in \{0,1\}$ are then drawn randomly according to the Bernoulli distribution with

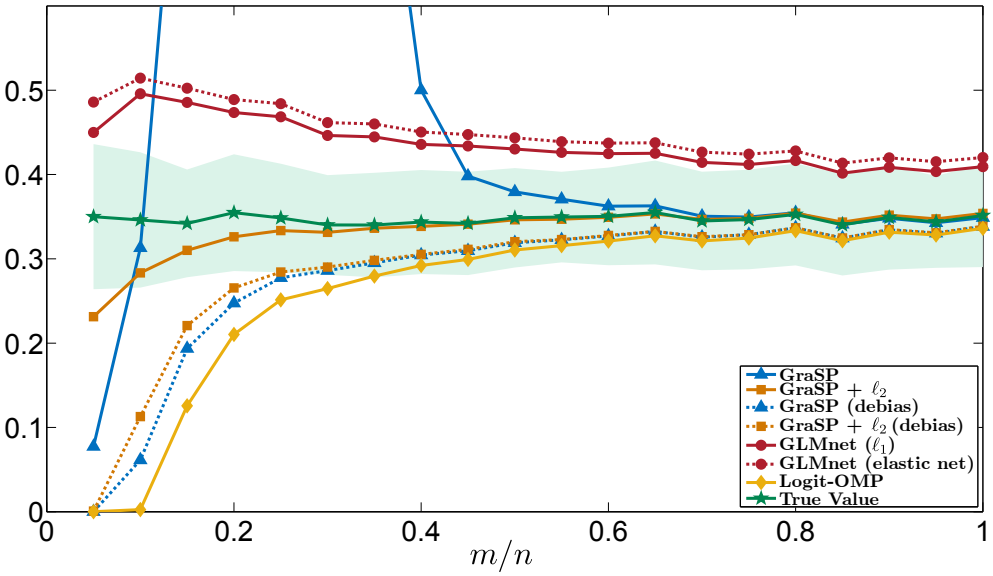
$$\Pr\{y = 0 \mid \mathbf{a}\} = 1/\left(1 + \exp\left(\langle \mathbf{a}, \mathbf{x}^\star \rangle + c\right)\right).$$

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 $\eta = (1-\omega)\sqrt{\frac{\log n}{m}}$. For each choice of the number of measurements m 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 (m, ρ).

Fig. 3.1 compares the average value of the empirical logistic loss achieved by each of the

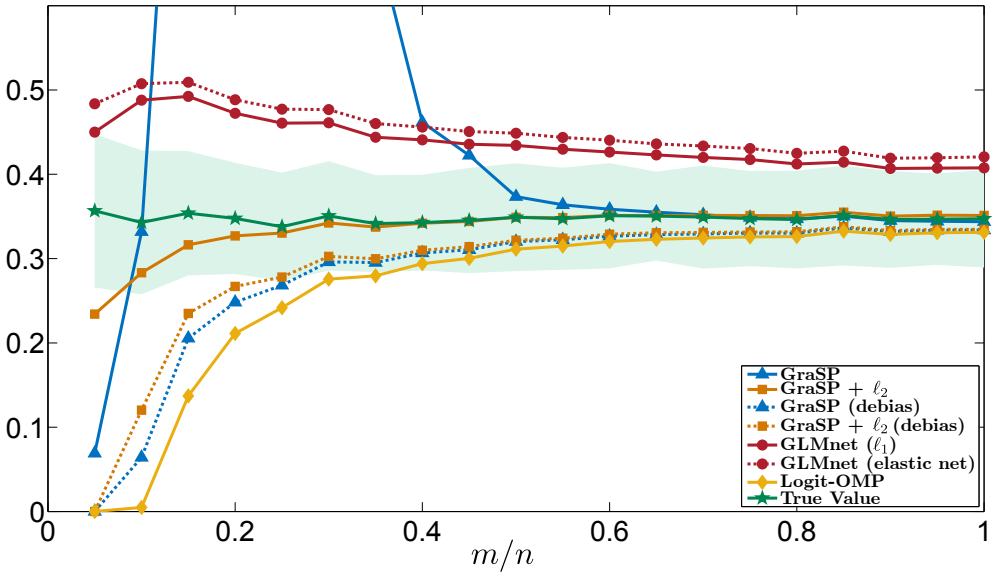


(a) $\rho = 0$

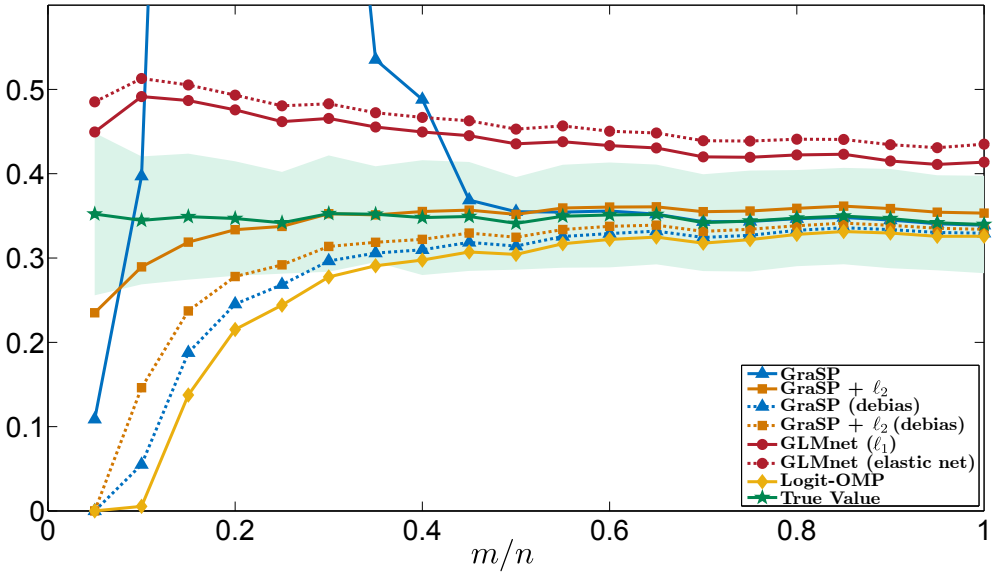


(b) $\rho = 1/3$

Figure 3.1: 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.



(c) $\rho = 1/2$



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

Figure 3.1: continued from the previous page

considered algorithms for a wide range of "sampling ratio" m/n. 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 3.3.1.

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 $m/n > 0.7$. 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 behavior at low sampling ratios.

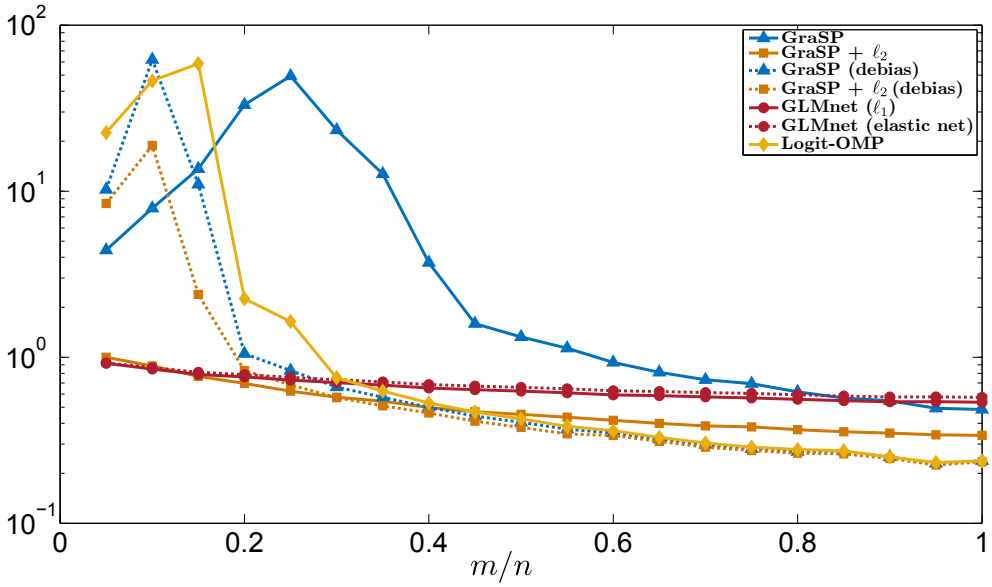
Interestingly, GraSP becomes more stable at low sampling ratios when the logistic loss is regularized with the ℓ_2-norm. However, this stability comes at the cost of a bias in the loss value at high sampling ratios that is particularly pronounced in Fig. 3.1d. Nevertheless, for all of the tested values 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 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 $\rho = \frac{\sqrt{2}}{2}$.

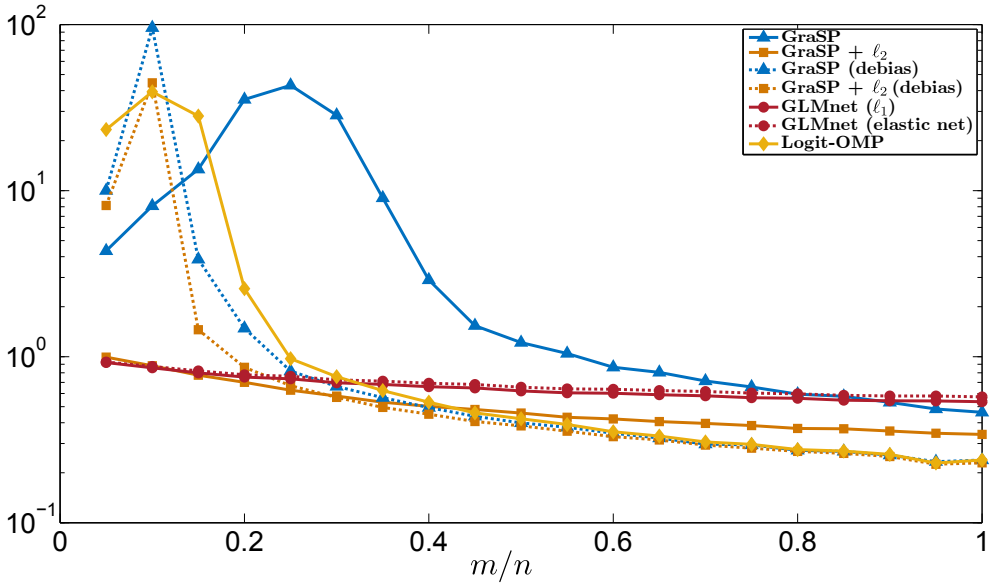
Fig. 3.2 illustrates the performance of the same algorithms in terms of the relative error $\|\widehat{\mathbf{x}} - \mathbf{x}^\star\|_2 / \|\mathbf{x}^\star\|_2$ where $\widehat{\mathbf{x}}$ denotes the estimate that the algorithms produce. Not surprisingly, none of the algorithms attain an arbitrarily 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. These variants of GraSP appear to perform better than Logit-OMP for almost the entire range of m/n.

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. (2004), namely the ARCENE and DEXTER data sets. The logistic loss values at obtained estimates are reported in Tables 3.1 and 3.2. For each data set we applied the sparse logistic regression for a range of sparsity level s. The

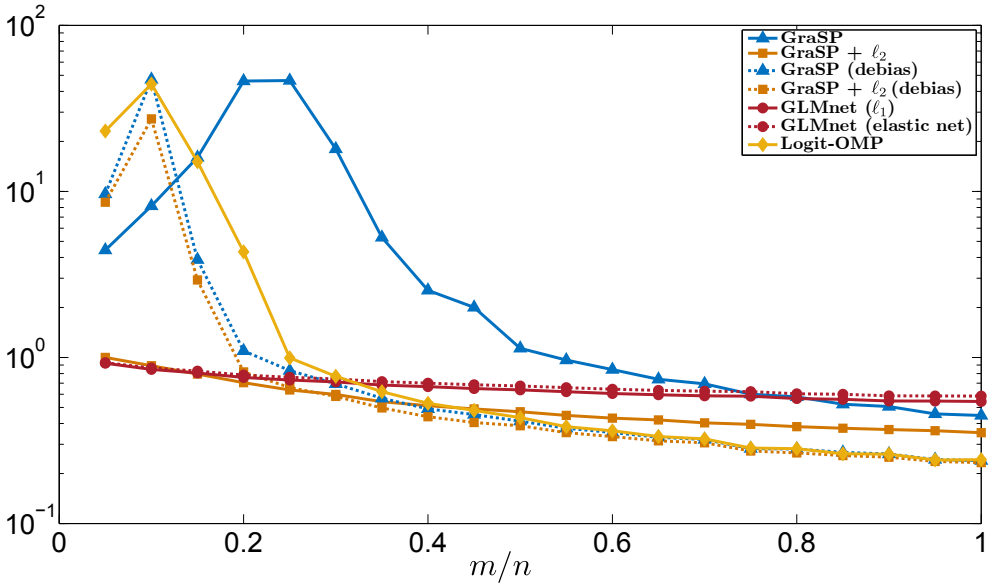


(a) $\rho = 0$

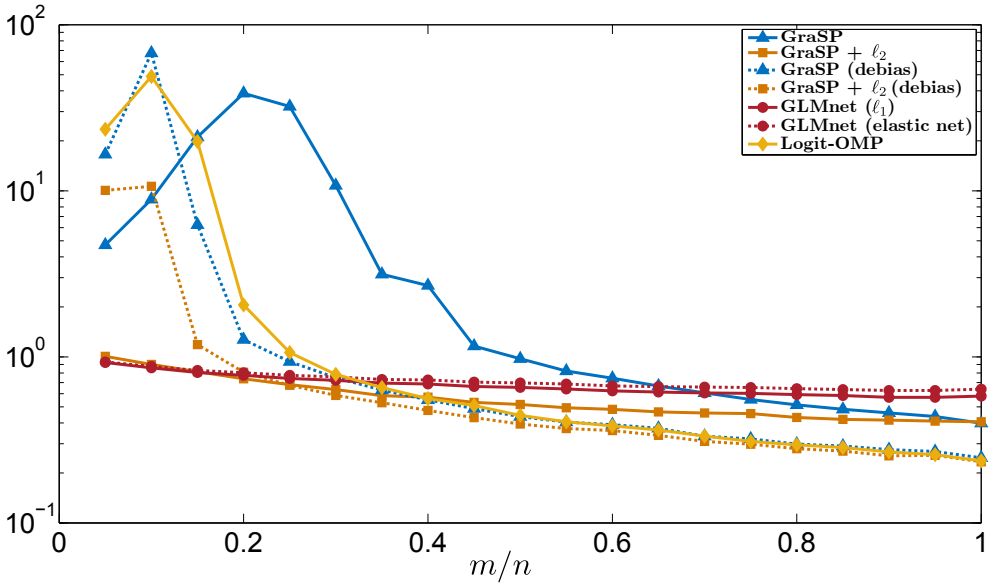


(b) $\rho = 1/3$

Figure 3.2: Comparison of the average relative error (i.e., $\|\widehat{\mathbf{x}} - \mathbf{x}^\star\|_2 / \|\mathbf{x}^\star\|_2$) 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.



(c) $\rho = 1/2$



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

Figure 3.2: continued from the previous page.

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

3.6 Summary and Discussion

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. We introduced a greedy algorithm called the Gradient Support Pursuit(GraSP) as an approximate solver for a wide range of sparsity-constrained optimization problems.

Table 3.1: ARCENE

s	G	Gd	Gℓ_2	Gℓ_2d	ℓ_1	E-net	Logit-OMP
5	5.89E+1	5.75E-1	2.02E+1	5.24E-1	5.59E-1	6.43E-1	2.23E-1
10	3.17E+2	5.43E-1	3.71E+1	4.53E-1	5.10E-1	5.98E-1	5.31E-7
15	3.38E+2	6.40E-7	5.94	1.42E-7	4.86E-1	5.29E-1	5.31E-7
20	1.21E+2	3.44E-7	8.82	3.08E-8	4.52E-1	5.19E-1	5.31E-7
25	9.87E+2	1.13E-7	4.46E+1	1.35E-8	4.18E-1	4.96E-1	5.31E-7

Table 3.2: DEXTER

s	G	Gd	Gℓ_2	Gℓ_2d	ℓ_1	E-net	Logit-OMP
5	7.58	3.28E-1	3.30	2.80E-1	5.75E-1	6.08E-1	2.64E-1
10	1.08	1.79E-1	4.33E-1	1.28E-1	5.23E-1	5.33E-1	1.79E-1
15	6.06	1.71E-1	3.35E-1	1.17E-1	4.88E-1	4.98E-1	1.16E-1
20	1.30	8.84E-2	1.79E-1	8.19E-2	4.27E-1	4.36E-1	4.60E-2
25	1.17	2.51E-7	2.85E-1	1.17E-2	3.94E-1	4.12E-1	4.62E-3
30	3.04E-1	5.83E-7	2.65E-1	1.77E-7	3.70E-1	3.88E-1	2.88E-7
35	6.22E-1	2.08E-7	2.68E-1	1.19E-7	3.47E-1	3.72E-1	2.14E-7
40	5.38E-1	2.01E-7	6.30E-2	1.27E-7	3.31E-1	3.56E-1	2.14E-7
45	3.29E-1	2.11E-7	1.05E-1	1.47E-7	3.16E-1	3.41E-1	2.14E-7
50	2.06E-1	1.31E-7	5.66E-2	1.46E-7	2.87E-1	3.11E-1	2.14E-7
55	3.61E-2	1.20E-7	8.40E-2	1.31E-7	2.80E-1	2.89E-1	2.14E-7
60	1.18E-1	2.46E-7	5.70E-2	1.09E-7	2.66E-1	2.82E-1	2.14E-7
65	1.18E-1	7.86E-8	2.87E-2	9.47E-8	2.59E-1	2.75E-1	2.14E-7
70	8.92E-2	1.17E-7	2.23E-2	8.15E-8	2.52E-1	2.69E-1	2.14E-7
75	1.03E-1	8.54E-8	3.93E-2	7.94E-8	2.45E-1	2.69E-1	2.14E-7

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 chapter. 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, n, the running time of the GraSP algorithm will be dominated by evaluation of the function and its gradient, whose computational cost grows with n. 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 is an interesting research topic for future work.

Chapter 4

1-bit Compressed Sensing

4.1 Background

Quantization is an indispensable part of digital signal processing and digital communications systems. To incorporate CS methods in these systems, it is thus necessary to analyze and evaluate them considering the effect of measurement quantization. There has been a growing interest in quantized CS in the literature (Laska et al., 2009; Dai et al., 2009; Sun and Goyal, 2009; Zymnis et al., 2010; Jacques et al., 2011; Laska et al., 2011b), particularly the extreme case of quantization to a single bit dubbed 1-bit Compressed Sensing (Boufounos and Baraniuk, 2008). As mentioned in Chapter 2, in 1-bit CS problems only the sign of linear measurements are recorded. The advantage of this acquisition scheme is that it can be implemented using simple hardware that is not expensive and can operate at very high sampling rates.

As in standard CS, the algorithms proposed for the 1-bit CS problem can be categorized into convex methods and non-convex greedy methods. Boufounos and Baraniuk (2008) proposed an algorithm for 1-bit CS reconstruction that induces sparsity through

the ℓ_1-norm while penalizes incosistency with the 1-bit sign measurements via a convex regularization term. In a noise-free scenario, the 1-bit measurements do not convey any information about the length of the signal. Therefore, the algorithm in (Boufounos and Baraniuk, 2008), as well as other 1-bit CS algorithms, aim at accurate estimation of the normalized signal. Requiring the 1-bit CS estimate to lie on the surface of the unit-ball imposes a non-convex constraint in methods that perform an (approximate) optimization, even those that use the convex ℓ_1-norm to induce sparsity. Among greedy 1-bit CS algorithms, an algorithm called Matching Sign Pursuit (MSP) is proposed in (Boufounos, 2009) based on the CoSaMP algorithm (Needell and Tropp, 2009). This algorithm is empirically shown to perform better than standard CoSaMP algorithm for estimation of the normalized sparse signal. Laska et al. (2011a) propose the Restricted-Step Shrinkage (RSS) algorithm for 1-bit CS problems. This algorithm, which is similar to *trust-region* algorithms in non-convex optimization, is shown to converge to a stationary point of the objective function regardless of the initialization. More recently, Jacques et al. (2013) derived a lower bound on the best achievable reconstruction error of any 1-bit CS algorithm in noise-free scenarios. Furthermore, using the notion of "binary stable embeddings", they have shown that Gaussian measurement matrices can be used for 1-bit CS problems both in noisy and noise-free regime. The Binary Iterative Hard Thresholding (BIHT) algorithm is also proposed in (Jacques et al., 2013) and shown to have favorable performance compared to the RSS and MSP algorithms through numerical simulations. For robust 1-bit CS in presence of noise, Yan et al. (2012) also proposed the Adaptive Outlier Pursuit (AOP) algorithm. In each iteration of the AOP , first the sparse signal is estimated similar to BIHT with the difference that the potentially corrupted measurements are excluded. Then with the new signal estimate fixed, the algorithm updates the list of likely corrupted measurements. The AOP is shown to improve on performance of BIHT through numerical simulations.

Plan and Vershynin (2011) proposed a linear program to solve the 1-bit CS problems in a noise-free scenario. The algorithm is proved to provide accurate solutions, albeit using a sub-optimal number of measurements. Furthermore, in (Plan and Vershynin, 2013) a convex program is proposed that is robust to noise in 1-bit measurements and achieves the optimal number of measurements.

4.2 Problem Formulation

We cast the 1-bit CS problem in the framework of statistical parametric estimation which is also considered in (Zymnis et al., 2010). In 1-bit CS, binary measurements $y \in \{\pm 1\}$ of a signal $\mathbf{x}^\star \in \mathbb{R}^n$ are collected based on the model

$$y = \operatorname{sgn}\left(\langle \mathbf{a}, \mathbf{x}^\star \rangle + e\right), \tag{4.1}$$

where $\mathbf{a}$ is a measurement vector and e denotes an additive noise with distribution $\mathcal{N}\left(0,\sigma^2\right)$. It is straightforward to show the conditional likelihood of y given $\mathbf{a}$ and signal $\mathbf{x}$ can be written as

$$\Pr\left\{y \mid \mathbf{a}; \mathbf{x}\right\} = \Phi\left(y\frac{\langle \mathbf{a}, \mathbf{x}\rangle}{\sigma}\right),$$

with $\Phi(\cdot)$ denoting the standard normal cumulative distribution function (CDF). Then, for measurement pairs $\{(\mathbf{a}_i, y_i)\}_{i=1}^m$the MLE loss function is given by

$$f_{\text{MLE}}(\mathbf{x}) := -\frac{1}{m}\sum_{i=1}^{m} \log\left(\Phi\left(y_i\frac{\langle \mathbf{a}_i, \mathbf{x}\rangle}{\sigma}\right)\right).$$

Note, however, that at high Signal-to-Noise Ratio (SNR) regime this function has erratic behavior. To observe this behavior, rewrite f_{MLE} as

$$f_{\text{MLE}}(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m} g_\eta\left(y_i\left\langle \mathbf{a}_i, \frac{\mathbf{x}}{\|\mathbf{x}\|_2}\right\rangle\right),$$

where $\eta := \frac{\|\mathbf{x}\|_2}{\sigma}$ is the SNR and $g_\omega(t) := -\log \Phi(\omega t)$ for all $\omega \geq 0$. As $\eta \to +\infty$ the function $g_\eta(t)$ tends to

$$g_\infty(t) := \begin{cases} 0 & t > 0 \\ \log 2 & t = 0 \\ +\infty & t < 0 \end{cases}.$$

Therefore, as the SNR increases to infinity $f_{\text{MLE}}(\mathbf{x})$ tends to a sum of discontinuous functions that is difficult to handle in practice. Whether the noise level is too low or the signal too strong relative to the noise, in a high SNR scenario the measurement vectors are likely to become linearly separable with respect to the corresponding binary measurements. In these cases, the minimizer of f_{MLE} would be pushed to infinity resulting in large estimation error.

To avoid the problems mentioned above we consider a modified loss function

$$f_0(\mathbf{x}) := -\frac{1}{m}\sum_{i=1}^{m} \log\left(\Phi\left(y_i \langle \mathbf{a}_i, \mathbf{x}\rangle\right)\right), \tag{4.2}$$

while we merely use an alternative formulation of (4.1) given by

$$y = \operatorname{sgn}\left(\eta \langle \mathbf{a}, \mathbf{x}^\star\rangle + e\right),$$

in which $\eta > 0$ denotes the true SNR, $\mathbf{x}^\star$ is assumed to be unit-norm, and $e \sim \mathcal{N}(0,1)$. The aim is accurate estimation of the unit-norm signal $\mathbf{x}^\star$ which is assumed to be s-sparse. Disregarding computational complexity, the candidate estimator would be

$$\arg\min_{\mathbf{x}} f_0(\mathbf{x}) \quad \text{s.t. } \|\mathbf{x}\|_0 \leq s \text{ and } \|\mathbf{x}\|_2 \leq 1. \tag{4.3}$$

However, finding the exact solution (4.3) may be computationally intractable, thereby we

merely focus on approximate solutions to this optimization problem.

4.3 Algorithm

In this section we introduce a modified version of the GraSP algorithm, outlined in Algorithm 2, for estimation of bounded sparse signals associated with a cost function. While in this chapter the main goal is to study the 1-bit CS problem and in particular the objective function described by (4.2), we state performance guarantees of Algorithm 2 in more general terms. As in GraSP, in each iteration first the $2s$ coordinates at which the gradient of the cost function at the iterate $\mathbf{x}^{(t)}$ has the largest magnitudes are identified. These coordinates, denoted by $\mathcal{Z}$, are then merged with the support set of $\mathbf{x}^{(t)}$ to obtain the set $\mathcal{T}$ in the second step of the iteration. Then, as expressed in line 3 of Algorithm 2, a crude estimate $\mathbf{b}$ is computed by minimizing the cost function over vectors of length no more than r whose supports are subsets of $\mathcal{T}$. Note that this minimization would be a convex program and therefore tractable, provided that the sufficient conditions proposed in Section 4.4 hold. In the final step of the iteration (i.e., line 4) the crude estimate is pruned to its best s-term approximation to obtain the next iterate $\mathbf{x}^{(t+1)}$. By definition we have $\|\mathbf{b}\|_2 \leq r$, thus the new iterate remains in the feasible set (i.e., $\left\|\mathbf{x}^{(t+1)}\right\|_2 \leq r$).

4.4 Accuracy Guarantees

In order to provide accuracy guarantees for Algorithm 2, we rely on the notion of SRH described in Definition 3.1 with a slight modification in its definition. The original definition of SRH basically characterizes the cost functions that have bounded curvature over sparse canonical subspaces, possibly at locations arbitrarily far from the origin. However, we only require the bounded curvature condition to hold at locations that are within a

Algorithm 2: GraSP with Bounded Thresholding

input : s desired sparsity level; r radius of the feasible set; $f(\cdot)$ the cost function

$t \longleftarrow 0$

$\mathbf{x}^{(t)} \longleftarrow \mathbf{0}$

repeat

1. $\mathcal{Z} \longleftarrow \operatorname{supp}\left(\left[\nabla f\left(\mathbf{x}^{(t)}\right)\right]_{2s}\right)$
2. $\mathcal{T} \longleftarrow \operatorname{supp}\left(\mathbf{x}^{(t)}\right) \cup \mathcal{Z}$
3. $\mathbf{b} \longleftarrow \arg\min_{\mathbf{x}} f(\mathbf{x}) \quad \text{s.t. } \mathbf{x}|_{\mathcal{T}^c} = \mathbf{0} \text{ and } \|\mathbf{x}\|_2 \leq r$
4. $\mathbf{x}^{(t+1)} \longleftarrow \mathbf{b}_s$
5. $t \longleftarrow t+1$

until *halting condition holds*

return $\mathbf{x}^{(t)}$

sphere around the origin. More precisely, we redefine the SRH as follows.

Definition 4.1 (Stable Restricted Hessian)**.** Suppose that $f : \mathbb{R}^n \mapsto \mathbb{R}$ is a twice continuously differentiable function and let $k < n$ be a positive integer. Furthermore, let $\alpha_k(\mathbf{x})$ and $\beta_k(\mathbf{x})$ be in turn the largest and smallest real numbers such that

$$\beta_k(\mathbf{x}) \|\mathbf{\Delta}\|_2^2 \leq \mathbf{\Delta}^{\mathrm{T}} \nabla^2 f(\mathbf{x}) \mathbf{\Delta} \leq \alpha_k(\mathbf{x}) \|\mathbf{\Delta}\|_2^2, \tag{4.4}$$

holds for all $\mathbf{\Delta}$ and $\mathbf{x}$ that obey $|\operatorname{supp}(\mathbf{\Delta}) \cup \operatorname{supp}(\mathbf{x})| \leq k$ and $\|\mathbf{x}\|_2 \leq r$. Then f is said to have an Stable Restricted Hessian of order k with constant $\mu_k \geq 1$ in a sphere of radius $r > 0$, or for brevity (μ_k, r)-SRH, if $1 \leq \alpha_k(\mathbf{x}) / \beta_k(\mathbf{x}) \leq \mu_k$ for all k-sparse $\mathbf{x}$ with $\|\mathbf{x}\|_2 \leq r$.

Theorem 4.1. *Let $\overline{\mathbf{x}}$ be a vector such that $\|\overline{\mathbf{x}}\|_0 \leq s$ and $\|\overline{\mathbf{x}}\|_2 \leq r$. If the cost function $f(\mathbf{x})$ have (μ_{4s}, r)-SRH corresponding to the curvature bounds $\alpha_{4s}(\mathbf{x})$ and $\beta_{4s}(\mathbf{x})$ in (4.4), then iterates of Algorithm 2 obey*

$$\left\|\mathbf{x}^{(t+1)} - \overline{\mathbf{x}}\right\|_2 \leq \left(\mu_{4s}^2 - \mu_{4s}\right) \left\|\mathbf{x}^{(t)} - \overline{\mathbf{x}}\right\|_2 + 2\left(\mu_{4s} + 1\right)\epsilon,$$

where ϵ *obeys* $\|[\nabla f(\overline{\mathbf{x}})]_{3s}\|_2 \leq \epsilon\, \beta_{4s}(\mathbf{x})$ *for all* $\mathbf{x}$ *with* $\|\mathbf{x}\|_0 \leq 4s$ *and* $\|\mathbf{x}\|_2 \leq r$.

The immediate implication of this theorem is that if the 1-bit CS loss $f_0(\mathbf{x})$ has $(\mu_{4s}, 1)$-SRH with $\mu_{4s} \leq \frac{1+\sqrt{3}}{2}$ then we have $\left\|\mathbf{x}^{(t)} - \mathbf{x}^\star\right\|_2 \leq 2^{-t}\|\mathbf{x}^\star\|_2 + 2\left(3+\sqrt{3}\right)\epsilon$.

Proof of Theorem 4.1 is almost identical to the proof of Theorem 3.1. For brevity we will provide a proof sketch in Appendix B and elaborate only on the more distinct parts of the proof and borrow the remaining parts from Appendix A.

4.5 Simulations

In our simulations using synthetic data we considered signals of dimensionality $n = 1000$ that are s-sparse with $s = 10, 20,$ or 30. The non-zero entries of the signal constitute a vector randomly drawn from the surface of the unit Euclidean ball in $\mathbb{R}^s$. The $m \times n$ measurement matrix has iid standard Gaussian entries with m varying between 100 and 2000 in steps of size 100. We also considered three different noise variances σ^2 corresponding to input SNR $\eta = 20$dB, 10dB, and 0dB. Figures 4.1–4.5 illustrate the average performance of the considered algorithm over 200 trials versus the sampling ratio (i.e., m/n). In these figures, the results of Algorithm 2 considering f_0 and f_{MLE} as the objective function are demarcated by GraSP and GrasP-η, respectively. Furthermore, the results corresponding to BIHT algorithm with one-sided ℓ_1 and ℓ_2 objective functions are indicated by BIHT and BIHT-ℓ_2, respectively. We also considered the ℓ_0-constrained optimization proposed by Plan and Vershynin (2013) which we refer to as PV-ℓ_0. While Plan and Vershynin (2013) mostly focused on studying the convex relaxation of this method using ℓ_1-norm, as shown in Appendix B the solution to PV-ℓ_0 can be derived explicitly in terms of the one-bit measurements, the measurement matrix, and the sparsity level. We do not evaluate the convex solver proposed in (Plan and Vershynin, 2013) because we did not have access to an efficient implementation of this method. Furthermore, this convex solver is expected to be

inferior to PV-ℓ_0 in terms of accuracy because it operates on a feasible set with larger *mean width* (see Plan and Vershynin, 2013, Theorem 1.1). With the exception of the non-iterative PV-ℓ_0, the other four algorithms considered in our simulations are iterative; they are configured to halt when they produce an estimate whose 1-bit measurements and the real 1-bit measurements have a Hamming distance smaller than an η-dependent threshold.

Figure 4.1 illustrates performance of the considered algorithms in terms of the angular error between the normalized estimate $\widehat{\mathbf{x}}$ and the true signal $\mathbf{x}^\star$ defined as $\mathrm{AE}\left(\widehat{\mathbf{x}}\right) := \frac{1}{\pi}\cos^{-1}\left\langle \widehat{\mathbf{x}}, \mathbf{x}^\star \right\rangle$. As can be seen from the figure, with higher input SNR (i.e., η) and less sparse target signals the algorithms incur larger angular error. While there is no significant difference in performance of GaSP, GraSP-η, and BIHT-ℓ_2 for the examined values of η and s, the BIHT algorithm appears to be sensitive to η. At $\eta = 20$dB and low sampling ratios BIHT outperforms the other methods by a noticeable margin. However, for more noisy measurements BIHT loses its advantage and at $\eta = 0$dB it performs even poorer than the PV-ℓ_0. PV-ℓ_0 never outperforms the two variants of GraSP or the BIHT-ℓ_2, but the gap between their achieved angular error decreases as the measurements become more noisy.

The reconstruction SNR of the estimates produced by the algorithms are compared in Figure 4.2. The reconstruction SNR conveys the same information as the angular error as it can be calculated through the formula

$$\begin{aligned}\text{R-SNR}\left(\widehat{\mathbf{x}}\right) &:= -20\log_{10}\left\|\widehat{\mathbf{x}} - \mathbf{x}^\star\right\|_2 \\ &= -10\log_{10}\left(2 - 2\cos \mathrm{AE}\left(\widehat{\mathbf{x}}\right)\right).\end{aligned}$$

However, it magnifies small differences between the algorithms that were difficult to trace using the angular error. For example, it can be seen in Figure 4.2 that at $\eta = 20$dB and $s = 10$, GraSP-η has an advantage (of up to 2dB) in reconstruction SNR.

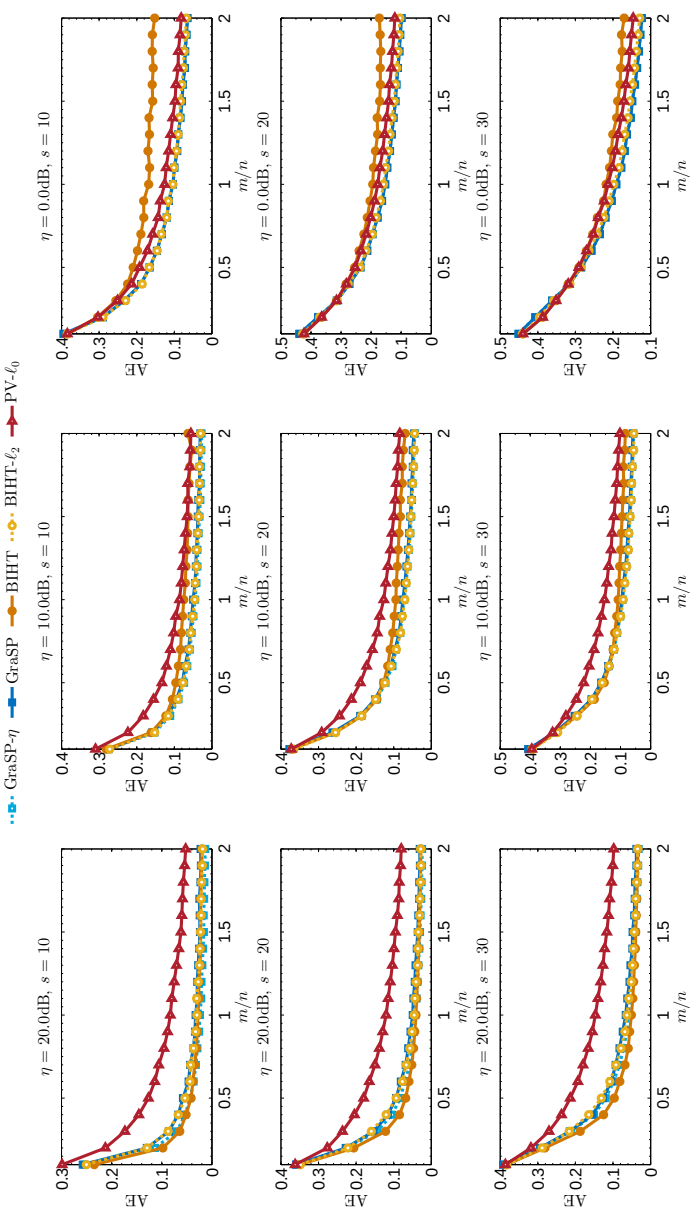


Figure 4.1: Angular error (AE) vs. the sampling ratio (m/n) at different values of input SNR (η) and sparsity (s)

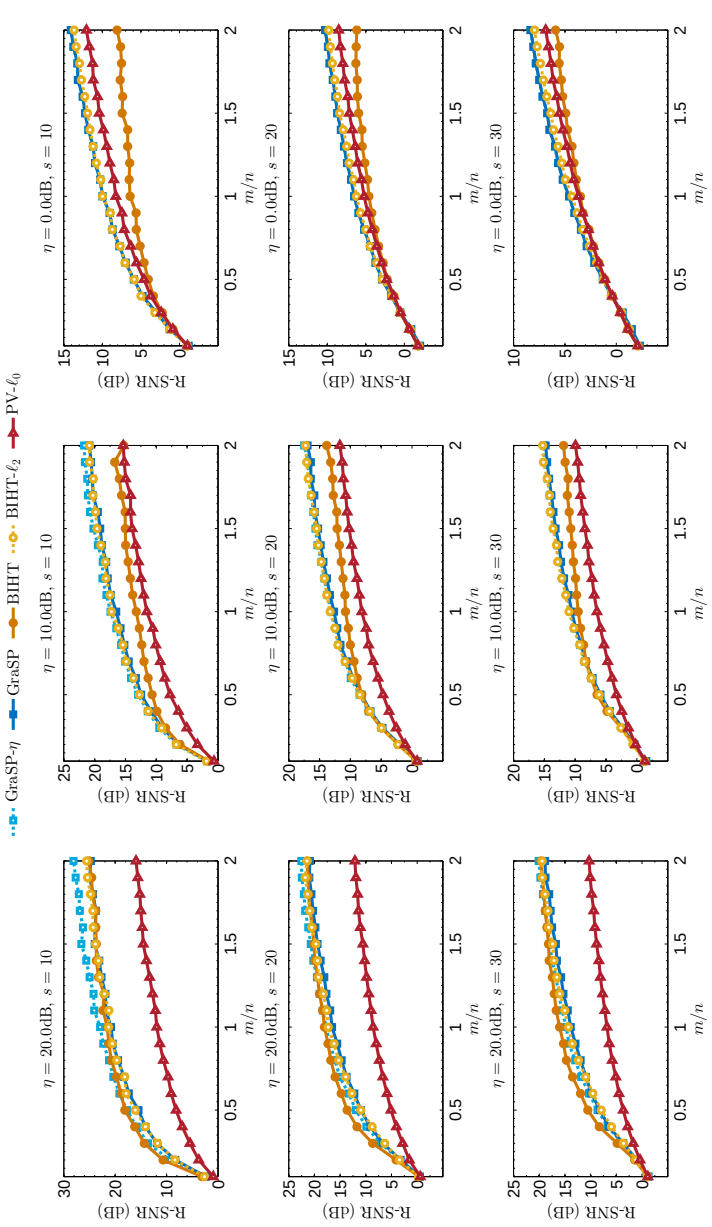


Figure 4.2: Reconstruction SNR on the unit ball (R-SNR) vs. the sampling ratio (m/n) at different values of input SNR (η) and sparsity (s)

Furthermore, we evaluated performance of the algorithms in terms of identifying the correct support set of the target sparse signal by are comparing their achieved False Negative Rate

$$\mathrm{FNR} = \frac{|\mathrm{supp}\,(\mathbf{x}^\star) \setminus \mathrm{supp}\,(\widehat{\mathbf{x}})|}{|\mathrm{supp}\,(\mathbf{x}^\star)|}$$

and False Positive Rate

$$\mathrm{FPR} = \frac{|\mathrm{supp}\,(\widehat{\mathbf{x}}) \setminus \mathrm{supp}\,(\mathbf{x}^\star)|}{n - |\mathrm{supp}\,(\mathbf{x}^\star)|}.$$

Figures 4.3 and 4.4 illustrate these rates for the studied algorithms. It can be seen in Figure 4.3 that at $\eta = 20$dB, BIHT achieves a FNR slightly lower than that of the variants of GraSP, whereas PV-ℓ_0 and BIHT-ℓ_2 rank first and second, respectively, in the highest FNR at a distant from the other algorithms. However, as η decreases the FNR of BIHT deteriorates relative to the other algorithms while BIHT-ℓ_2 shows improved FNR. The GraSP variants exhibit better performance overall at smaller values of η especially with $s = 10$, but for $\eta = 10$dB and at low sampling ratios BIHT attains a slightly better FNR. The relative performance of the algorithms in terms of FPR, illustrated in Figure 4.4, is similar.

We also compared the algorithms in terms of their average execution time (T) measured in seconds. The simulation was ran on a PC with an AMD Phenom™II X6 2.60GHz processor and 8.00GB of RAM. The average execution time of the algorithms, all of which are implemented in MATLAB®, is illustrated in 4.5 in log scale. It can be observed from the figure that PV-ℓ_0 is the fastest algorithm which can be attributed to its non-iterative procedure. Furthermore, in general BIHT-ℓ_2 requires significantly longer time compared to the other algorithms. The BIHT, however, appears to be the fastest among the iterative algorithms at low sampling ratio or at large values of η. The GraSP variants generally run at similar speed, while they are faster than BIHT at low values of η and high sampling

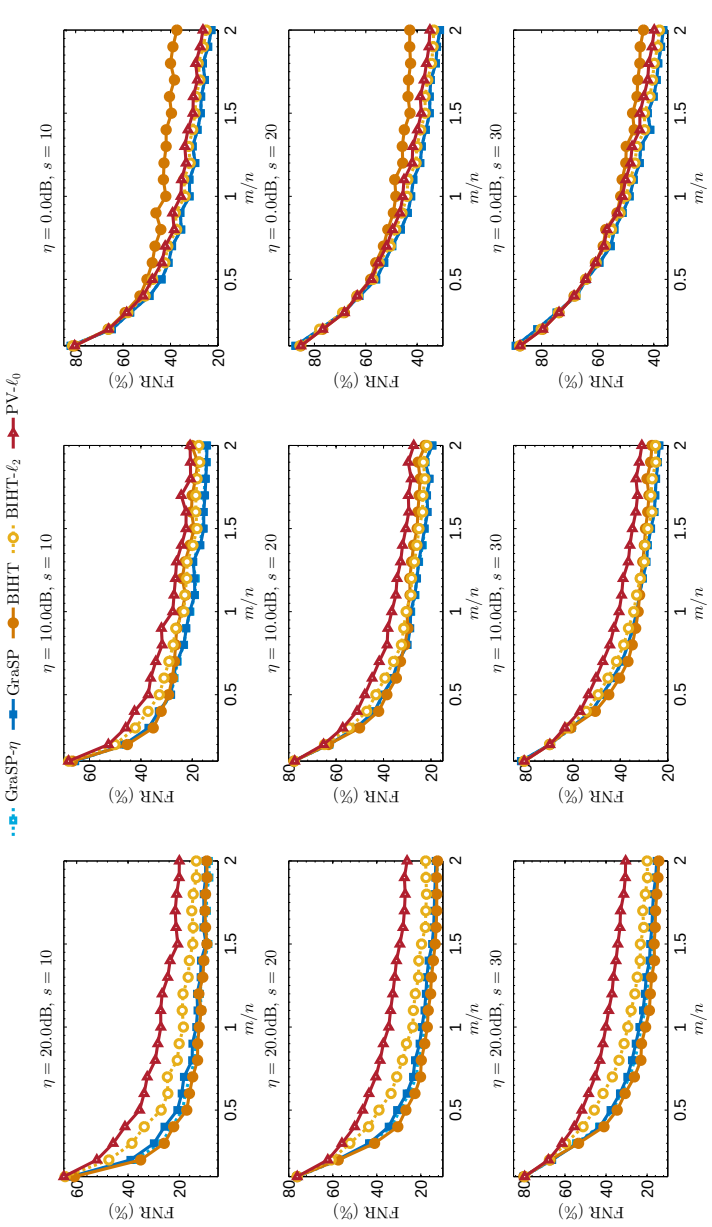


Figure 4.3: False Negative Rate (FNR) vs. the sampling ratio (m/n) at different values of input SNR (η) and sparsity (s)

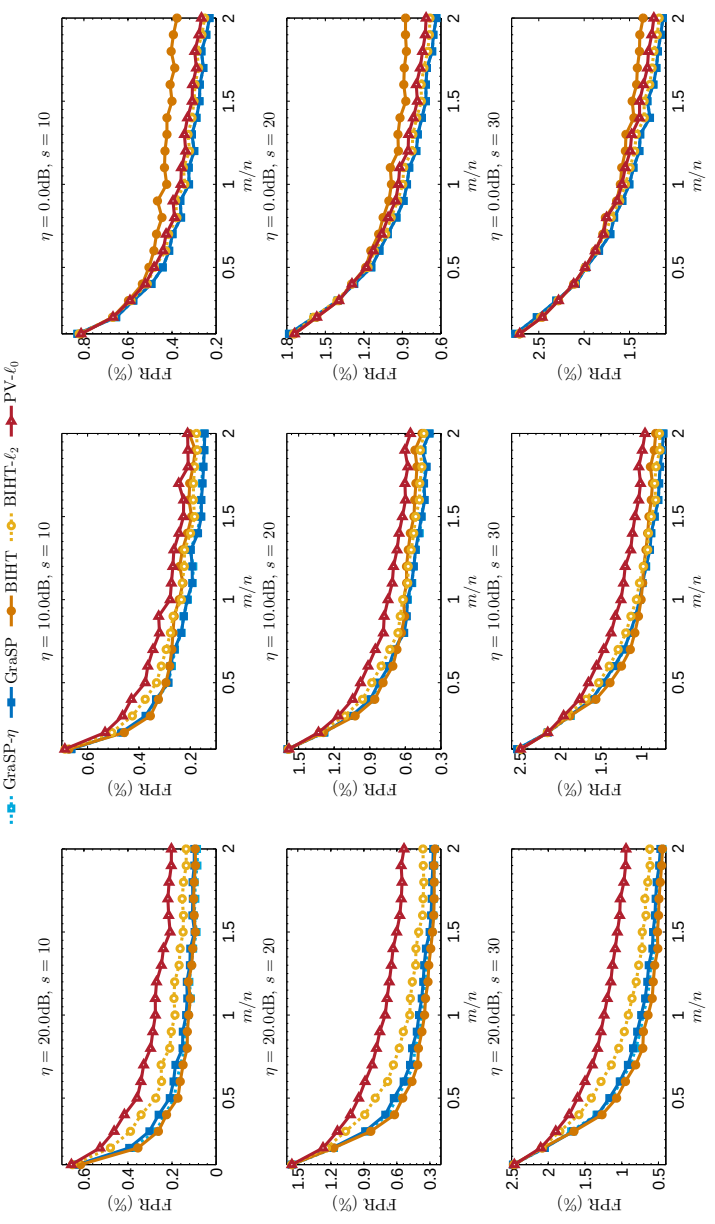


Figure 4.4: False Positive Rate (FPR) vs. the sampling ratio (m/n) at different values of input SNR (η) and sparsity (s)

ratios.

4.6 Summary

In this chapter we revisited a formulation of the 1-bit CS problem and applied a variant of the GraSP algorithm to this problem. We showed through numerical simulations that the proposed algorithms have robust performance in presence of noise. While at high levels of input SNR these algorithms are outperformed by a narrow margin by the competing algorithms, in low input SNR regime our algorithms show a solid performance at reasonable computational cost.

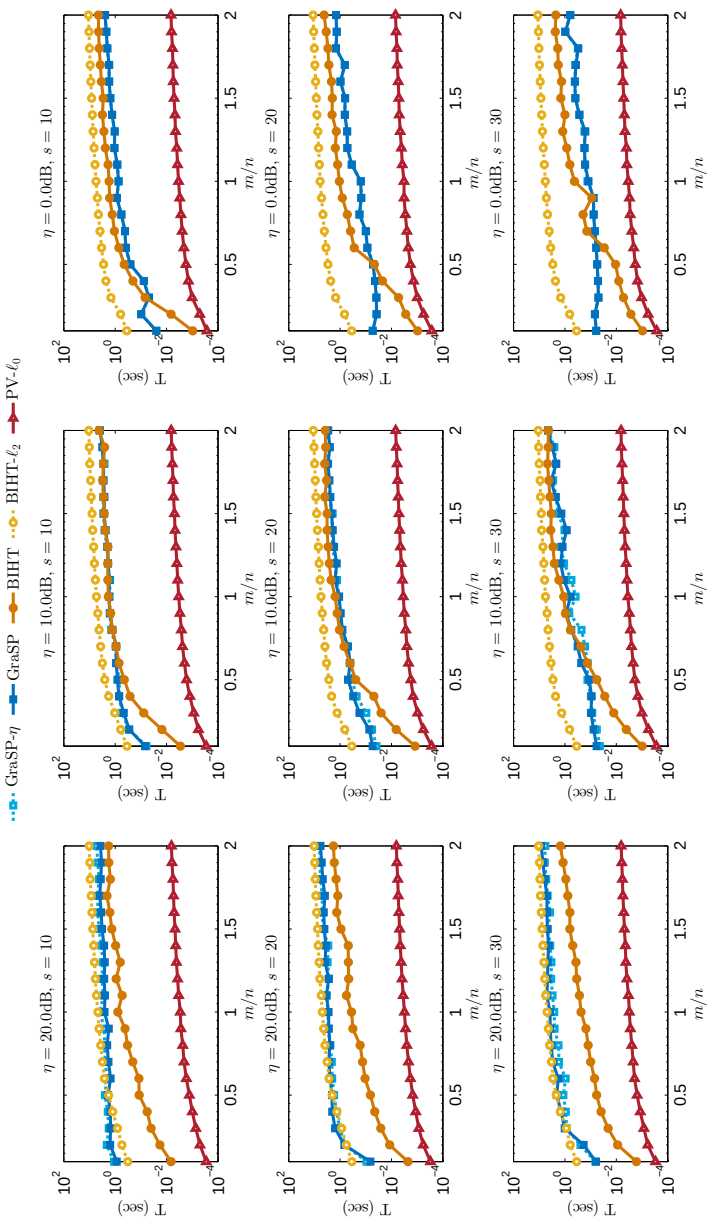


Figure 4.5: Average execution time (T) vs. the sampling ratio (m/n) at different values of input SNR (η) and sparsity (s)

Chapter 5

Estimation Under Model-Based Sparsity

5.1 Background

Beyond the ordinary, extensively studied, plain sparsity model, a variety of structured sparsity models have been proposed in the literature (Bach, 2008; Roth and Fischer, 2008; Jacob et al., 2009; Baraniuk et al., 2010; Bach, 2010; Bach et al., 2012; Chandrasekaran et al., 2012; Kyrillidis and Cevher, 2012a). These sparsity models are designed to capture the interdependence of the locations of the non-zero components that is known *a priori* in certain applications. For instance, the wavelet transform of natural images are often (nearly) sparse and the dependence among the dominant wavelet coefficients can be represented by a rooted and connected tree. Furthermore, in applications such as array processing or sensor networks, while different sensors may take different measurements, the support set of the observed signal is identical across the sensors. Therefore, to model this property of the system, we can compose an enlarged signal with *jointly-sparse* or *block-*

sparse support set, whose non-zero coefficients occur as contiguous blocks.

The models proposed for structured sparsity can be divided into two types. Models of the first type have a combinatorial construction and explicitly enforce the permitted "non-zero patterns" (Baraniuk et al., 2010; Kyrillidis and Cevher, 2012a,b). Greedy algorithms have been proposed for the least squares regression with true parameters belonging to such combinatorial sparsity models (Baraniuk et al., 2010; Kyrillidis and Cevher, 2012b). Models of the second type capture sparsity patterns induced by the convex penalty functions tailored for specific estimation problems. For example, consistency of linear regression with mixed ℓ_1/ℓ_2-norm regularization in estimation of group sparse signals having non-overlapping groups is studied in (Bach, 2008). Furthermore, a different convex penalty to induce group sparsity with overlapping groups is proposed in (Jacob et al., 2009). In (Bach, 2010), using submodular functions and their Lovàsz extension, a more general framework for design of convex penalties that induce given sparsity patterns is proposed. In (Chandrasekaran et al., 2012) a convex signal model is proposed that is generated by a set of base signals called "atoms". The model can describe not only plain and structured sparsity, but also low-rank matrices and several other low-dimensional models. We refer readers to (Duarte and Eldar, 2011; Bach et al., 2012) for extensive reviews on the estimation of signals with structured sparsity.

In addition to linear regression problems under structured sparsity assumptions, nonlinear statistical models have been studied in the convex optimization framework (Roth and Fischer, 2008; Bach, 2008; Jenatton et al., 2011; Tewari et al., 2011). For example, using the signal model introduced in (Chandrasekaran et al., 2012), minimization of a convex function obeying a restricted smoothness property is studied in (Tewari et al., 2011) where a coordinate-descent type of algorithm is shown to converge to the minimizer at a sublinear rate. In this formulation and other similar methods that rely on convex relaxation one

needs to choose a regularization parameter to guarantee the desired statistical accuracy. However, choosing the appropriate value of this parameter may be intractable. Furthermore, the convex signal models usually provide an approximation of the ideal structures the estimates should have, while in certain tasks such as variable selection solutions are required to exhibit the exact structure considered. Therefore, in such tasks, convex optimization techniques may yield estimates that do not satisfy the desired structural properties, albeit accurately approximating the true parameter. These shortcomings motivate application of combinatorial sparsity structures in nonlinear statistical models, extending prior results such as Baraniuk et al. (2010); Kyrillidis and Cevher (2012b) that have focused exclusively on linear models.

Among the non-convex greedy algorithms, a generalization of CS is considered in (Blumensath, 2010) where the measurement operator is a nonlinear map and the union of subspaces is assumed as the signal model. As mentioned in Chapter 3 this formulation admits only a limited class of objective functions that are described using a norm. Furthermore, in (Lozano et al., 2011) proposed a generalization of the Orthogonal Matching Pursuit algorithm (Pati et al., 1993) that is specifically designed for estimation of group sparse parameters in GLMs.

In this chapter we study the *Projected Gradient Descent* method to approximate the minimizer of a cost function subject to a model-based sparsity constraint. The sparsity model considered in this chapter is similar to the models in (Baraniuk et al., 2010; Kyrillidis and Cevher, 2012b) with minor differences in the definitions. To guarantee the accuracy of the algorithm our analysis requires the cost function to have a Stable Model-Restricted Hessian (SMRH) as defined in Section 5.3. Using this property we show that for any given reference point in the considered model, each iteration shrinks the distance to the reference point up to an approximation error. As an example, Section 5.3 considers the cost func-

tions that arise in GLMs and discusses how the proposed sufficient condition (i.e., SMRH) can be verified and how large the approximation error of the algorithm is. To make precise statements on the SMRH and on the size of the approximation error we assume some extra properties on the cost function and/or the data distribution. Finally, we discuss and conclude in Section 5.5.

Notation. To proceed, first we introduce a few more notations used specifically in this chapter and Appendix C. For two non-empty families of sets $\mathcal{F}_1$ and $\mathcal{F}_2$ we write $\mathcal{F}_1 \Cup \mathcal{F}_2$ to denote another family of sets given by $\{\mathcal{X}_1 \cup \mathcal{X}_2 \mid \mathcal{X}_1 \in \mathcal{F}_1 \text{ and } \mathcal{X}_2 \in \mathcal{F}_2\}$. Moreover, for any non-empty family of sets $\mathcal{F}$ for conciseness we set $\mathcal{F}^j = \mathcal{F} \Cup \ldots \Cup \mathcal{F}$ where the operation $\Cup$ is performed $j-1$ times. For generality, in this chapter we assume the objective functions are defined over a finite-dimensional Hilbert space $\mathcal{H}$. The inner product associated with this Hilbert space is written as $\langle \cdot, \cdot \rangle$. The norm induced by this inner product is denoted by $\|\cdot\|$.

5.2 Problem Statement and Algorithm

To formulate the problem of minimizing a cost function subject to structured sparsity constraints, first we provide a definition of the sparsity model. This definition is an alternative way of describing the *Combinatorial Sparse Models* in (Kyrillidis and Cevher, 2012a). In comparison, our definition merely emphasizes the role of a family of index sets as a *generator* of the sparsity model.

Definition 5.1. Suppose that n and k are two positive integers with $k \ll n$. Furthermore, denote by $\mathcal{C}_k$ a family of some non-empty subsets of $[n]$ that have cardinality at most k. The set $\bigcup_{\mathcal{S} \in \mathcal{C}_k} 2^{\mathcal{S}}$ is called a sparsity model of order k generated by $\mathcal{C}_k$ and denoted by $\mathcal{M}(\mathcal{C}_k)$.

Remark 5.1. Note that if a set $\mathcal{S} \in \mathcal{C}_k$ is a subset of another set in $\mathcal{C}_k$, then the same sparsity model can still be generated after removing $\mathcal{S}$ from $\mathcal{C}_k$ (i.e., $\mathcal{M}(\mathcal{C}_k) = \mathcal{M}(\mathcal{C}_k \backslash \{\mathcal{S}\})$). Thus, we can assume that there is no pair of distinct sets in $\mathcal{C}_k$ that one is a subset of the other.

In this chapter we aim to approximate the solution to the optimization problem

$$\arg\min_{\mathbf{x}\in\mathcal{H}} f(\mathbf{x}) \quad \text{s.t. } \operatorname{supp}(\mathbf{x}) \in \mathcal{M}(\mathcal{C}_k), \tag{5.1}$$

where $f : \mathcal{H} \mapsto \mathbb{R}$ is a cost function with $\mathcal{H}$ being a n-dimensional real Hilbert space, and $\mathcal{M}(\mathcal{C}_k)$ a given sparsity model described by Definition 5.1. To approximate a solution $\widehat{\mathbf{x}}$ to (5.1) we use a Projected Gradient Descent (PGD) method. PGD is one of the elementary tools in convex optimization for constrained minimization. For a differentiable convex objective function $f(\cdot)$, a convex set $\mathcal{Q}$, and a projection operator $\mathrm{P}_{\mathcal{Q}}(\cdot)$ defined by

$$\mathrm{P}_{\mathcal{Q}}(\mathbf{x}_0) = \arg\min_{\mathbf{x}} \|\mathbf{x} - \mathbf{x}_0\| \quad \text{s.t. } \mathbf{x} \in \mathcal{Q}, \tag{5.2}$$

the PGD algorithm solves the minimization

$$\arg\min_{\mathbf{x}} f(\mathbf{x}) \quad \text{s.t. } \mathbf{x} \in \mathcal{Q}$$

via the iterations outlined in Algorithm 3. To find an approximate solution to (5.1), however, we use a non-convex PGD method with the feasible set $\mathcal{Q} \equiv \mathcal{M}(\mathcal{C}_k) \cap \mathcal{B}_{\mathcal{H}}(r)$, where $\mathcal{B}_{\mathcal{H}}(r) := \{\mathbf{x} \mid \|\mathbf{x}\| \leq r\}$ is the centered ball of radius r with respect to the norm of the Hilbert space $\mathcal{H}$. The corresponding projection operator, denoted by $\mathrm{P}_{\mathcal{C}_k,r}(\cdot)$, is a mapping $\mathrm{P}_{\mathcal{C}_k,r} : \mathcal{H} \mapsto \mathcal{H}$ that at any given point $\mathbf{x}_0 \in \mathcal{H}$ evaluates to a solution to

$$\arg\min_{\mathbf{x}\in\mathcal{H}} \|\mathbf{x} - \mathbf{x}_0\| \quad \text{s.t. } \operatorname{supp}(\mathbf{x}) \in \mathcal{M}(\mathcal{C}_k) \text{ and } \|\mathbf{x}\| \leq r. \tag{5.3}$$

Remark 5.2. In parametric estimation problems, fidelity of the estimate is measured by the cost function $f(\cdot)$ that depends on observations generated by an underlying true param-

Algorithm 3: Projected Gradient Descent

input : Objective function $f(\cdot)$ and an operator $\mathrm{P}_{\mathcal{Q}}(\cdot)$ that performs projection onto the feasible set $\mathcal{Q}$

$t \longleftarrow 0$, $\mathbf{x}^{(t)} \longleftarrow \mathbf{0}$

repeat

1 choose step-size $\eta^{(t)} > 0$

2 $\mathbf{z}^{(t)} \longleftarrow \mathbf{x}^{(t)} - \eta^{(t)} \nabla f\left(\mathbf{x}^{(t)}\right)$

3 $\mathbf{x}^{(t+1)} \longleftarrow \mathrm{P}_{\mathcal{Q}}\left(\mathbf{z}^{(t)}\right)$

4 $t \longleftarrow t+1$

until *halting condition holds*

return $\mathbf{x}^{(t)}$

eter $\mathbf{x}^{\star}$. As mentioned in Remark 3.8, it is more desired in these problems to estimate $\mathbf{x}^{\star}$ rather than the solution $\widehat{\mathbf{x}}$ of (5.1), as it describes the data. Our analysis allows evaluating the approximation error of the Algorithm 3 with respect to any parameter vector in the considered sparsity model including $\widehat{\mathbf{x}}$ and $\mathbf{x}^{\star}$. However, the approximation error with respect to the statistical truth $\mathbf{x}^{\star}$ can be simplified and interpreted to a greater extent. We elaborate more on this in Section 5.3.

Remark 5.3. Assuming that for every $\mathcal{S} \in \mathcal{C}_k$ the cost function has a unique minimum over the set $\{\mathbf{x} \mid \mathrm{supp}(\mathbf{x}) \subseteq \mathcal{S} \text{ and } \|\mathbf{x}\| \leq r\}$, the operator $\mathrm{P}_{\mathcal{C}_k,r}(\cdot)$ can be defined without invoking *the axiom of choice* because there are only a finite number of choices for the set $\mathcal{S}$. Furthermore, the constraint $\|\mathbf{x}\| \leq r$ in (5.3) is necessary to validate SMRH as explained in 3.2. Finally, the exact projection onto the sparsity model $\mathcal{M}(\mathcal{C}_k)$ might not be tractable. One may desire to show that accuracy can be guaranteed even using an inexact projection operator, at the cost of an extra error term. Existence and complexity of algorithms that find the desired exact or approximate projections, disregarding the length constraint in (5.3) (i.e., $\mathrm{P}_{\mathcal{C}_k,+\infty}(\cdot)$), are studied in (Kyrillidis and Cevher, 2012a,b) for several interesting structured sparsity models. Also, in the general case where $r < +\infty$ the projection $\mathrm{P}_{\mathcal{C}_k,r}(\mathbf{x})$ can be derived from $\mathrm{P}_{\mathcal{C}_k,+\infty}(\mathbf{x})$ (see Lemma C.2 in Appendix C). Furthermore,

it is straightforward to generalize the guarantees in this chapter to cases where only approximate projection is tractable. However, we do not attempt it here; our focus is to study the algorithm when the cost function is not necessarily quadratic. Instead, we apply the results to certain statistical estimation problems with non-linear models and we derive bounds on the statistical error of the estimate.

5.3 Theoretical Analysis

5.3.1 Stable Model-Restricted Hessian

In order to demonstrate accuracy of estimates obtained using Algorithm 3 we require a variant of the SRH conditions proposed in Chapters 3 and 4 to hold. In contrast with Definitions 3.1 and 4.1, here we require this condition to hold merely for the signals that belong to the considered model and the curvature bounds are assumed to be global constants. Furthermore, similar to Definition 4.1, we explicitly bound the length of the vectors at which the condition should hold. The condition we rely on, the Stable Model-Restricted Hessian (SMRH), can be formally defined as follows.

Definition 5.2. Let $f : \mathcal{H} \mapsto \mathbb{R}$ be a twice continuously differentiable function. Furthermore, let $\alpha_{\mathcal{C}_k}$ and $\beta_{\mathcal{C}_k}$ be in turn the largest and smallest real numbers such that

$$\beta_{\mathcal{C}_k}\|\boldsymbol{\Delta}\|^2 \leq \left\langle \boldsymbol{\Delta}, \nabla^2 f\left(\mathbf{x}\right)\boldsymbol{\Delta}\right\rangle \leq \alpha_{\mathcal{C}_k}\|\boldsymbol{\Delta}\|^2, \tag{5.4}$$

holds for all $\boldsymbol{\Delta}$ and $\mathbf{x}$ such that $\operatorname{supp}\left(\boldsymbol{\Delta}\right) \cup \operatorname{supp}\left(\mathbf{x}\right) \in \mathcal{M}\left(\mathcal{C}_k\right)$ and $\|\mathbf{x}\| \leq r$. Then f is said to have a Stable Model-Restricted Hessian with respect to the model $\mathcal{M}\left(\mathcal{C}_k\right)$ with constant $\mu_{\mathcal{C}_k} \geq 1$ in a sphere of radius $r > 0$, or in short ($\mu_{\mathcal{C}_k}$,r)-SMRH, if $1 \leq \alpha_{\mathcal{C}_k}/\beta_{\mathcal{C}_k} \leq \mu_{\mathcal{C}_k}$.

Remark 5.4. If the true parameter is unbounded, violating the condition of 5.2, we may incur an estimation bias as quantified in Theorem 5.1.

5.3.2 Accuracy Guarantee

Using the notion of SMRH we can now state the main theorem.

Theorem 5.1. *Consider the sparsity model $\mathcal{M}(\mathcal{C}_k)$ for some $k \in \mathbb{N}$ and a cost function $f : \mathcal{H} \mapsto \mathbb{R}$ that satisfies the $\left(\mu_{\mathcal{C}_k^3}, r\right)$-SMRH condition with parameters $\alpha_{\mathcal{C}_k^3}$ and $\beta_{\mathcal{C}_k^3}$ as in (5.4). If $\eta^\star = 2/\left(\alpha_{\mathcal{C}_k^3} + \beta_{\mathcal{C}_k^3}\right)$ then for any $\overline{\mathbf{x}} \in \mathcal{M}(\mathcal{C}_k)$ with $\|\overline{\mathbf{x}}\| \leq r$ the iterates of Algorithm 3 obey*

$$\left\|\mathbf{x}^{(t+1)} - \overline{\mathbf{x}}\right\| \leq 2\gamma^{(t)}\left\|\mathbf{x}^{(t)} - \overline{\mathbf{x}}\right\| + 2\eta^{(t)}\left\|\nabla f(\overline{\mathbf{x}})|_{\overline{\mathcal{I}}}\right\|, \tag{5.5}$$

where $\gamma^{(t)} = \frac{\eta^{(t)}}{\eta^\star}\frac{\mu_{\mathcal{C}_k^3}-1}{\mu_{\mathcal{C}_k^3}+1} + \left|\frac{\eta^{(t)}}{\eta^\star} - 1\right|$ and $\overline{\mathcal{I}} = \operatorname{supp}\left(\mathrm{P}_{\mathcal{C}_k^2, r}(\nabla f(\overline{\mathbf{x}}))\right)$.

Remark 5.5. One should choose the step size to achieve a contraction factor $2\gamma^{(t)}$ that is as small as possible. Straightforward algebra shows that the constant step-size $\eta^{(t)} = \eta^\star$ is optimal, but this choice may not be practical as the constants $\alpha_{\mathcal{C}_k^3}$ and $\beta_{\mathcal{C}_k^3}$ might not be known. Instead, we can always choose the step-size such that $1/\alpha_{\mathcal{C}_k^3} \leq \eta^{(t)} \leq 1/\beta_{\mathcal{C}_k^3}$ provided that the cost function obeys the SMRH condition. It suffices to set $\eta^{(t)} = 1/\left\langle \boldsymbol{\Delta}, \nabla^2 f(\mathbf{x})\,\boldsymbol{\Delta}\right\rangle$ for some $\boldsymbol{\Delta}$,$\mathbf{x} \in \mathcal{H}$ such that $\operatorname{supp}(\boldsymbol{\Delta}) \cup \operatorname{supp}(\mathbf{x}) \in \mathcal{M}\left(\mathcal{C}_k^3\right)$. For this choice of $\eta^{(t)}$, we have $\gamma^{(t)} \leq \mu_{\mathcal{C}_k^3} - 1$.

Corollary 5.1. *A fixed step-size $\eta > 0$ corresponds to a fixed contraction coefficient $\gamma = \frac{\eta}{\eta^\star}\frac{\mu_{\mathcal{C}_k^3}-1}{\mu_{\mathcal{C}_k^3}+1} + \left|\frac{\eta}{\eta^\star} - 1\right|$. In this case, assuming that $2\gamma \neq 1$, the t-th iterate of Algorithm 3 satisfies*

$$\left\|\mathbf{x}^{(t)} - \overline{\mathbf{x}}\right\| \leq (2\gamma)^t\|\overline{\mathbf{x}}\| + 2\eta\frac{1-(2\gamma)^t}{1-2\gamma}\left\|\nabla f(\overline{\mathbf{x}})|_{\overline{\mathcal{I}}}\right\|. \tag{5.6}$$

In particular,

(i) if $\mu_{\mathcal{C}_k^3} < 3$ and $\eta = \eta^\star = 2/\left(\alpha_{\mathcal{C}_k^3} + \beta_{\mathcal{C}_k^3}\right)$, or

(ii) if $\mu_{\mathcal{C}_k^3} < \frac{3}{2}$ and $\eta \in \left[1/\alpha_{\mathcal{C}_k^3}, 1/\beta_{\mathcal{C}_k^3}\right]$,

the iterates converge to $\overline{\mathbf{x}}$ *up to an approximation error bounded above by* $\frac{2\eta}{1-2\gamma}\left\|\nabla f\left(\overline{\mathbf{x}}\right)|_{\overline{\mathcal{I}}}\right\|$ *with contraction factor* $2\gamma < 1$.

Proof. Applying (5.5) recursively under the assumptions of the corollary and using the identity $\sum_{j=0}^{t-1}(2\gamma)^j = \frac{1-(2\gamma)^t}{1-2\gamma}$ proves (5.6). In the first case, if $\mu_{\mathcal{C}_k^3} < 3$ and $\eta = \eta^\star = 2/\left(\alpha_{\mathcal{C}_k^3} + \beta_{\mathcal{C}_k^3}\right)$ we have $2\gamma < 1$ by definition of γ. In the second case, one can deduce from $\eta \in \left[1/\alpha_{\mathcal{C}_k^3}, 1/\beta_{\mathcal{C}_k^3}\right]$ that $|\eta/\eta^\star - 1| \leq \frac{\mu_{\mathcal{C}_k^3}-1}{2}$ and $\eta/\eta^\star \leq \frac{\mu_{\mathcal{C}_k^3}+1}{2}$ where equalities are attained simultaneously at $\eta = 1/\beta_{\mathcal{C}_k^3}$. Therefore, $\gamma \leq \mu_{\mathcal{C}_k^3} - 1 < 1/2$ and thus $2\gamma < 1$. Finally, in both cases it immediately follows from (5.6) that the approximation error converges to $\frac{2\eta}{1-2\gamma}\left\|\nabla f\left(\overline{\mathbf{x}}\right)|_{\overline{\mathcal{I}}}\right\|$ from below as $t \to +\infty$. ∎

5.4 Example: Generalized Linear Models

In this section we study the SMRH condition for objective functions that arise in Generalized Linear Models (GLMs) as described in Section 2.2.1. Recall from Chapter 2 that these objective functions have the form

$$f\left(\mathbf{x}\right) = \frac{1}{m}\sum_{i=1}^{m}\psi\left(\left\langle \mathbf{a}_i, \mathbf{x}\right\rangle\right) - y_i\left\langle \mathbf{a}_i, \mathbf{x}\right\rangle,$$

where $\psi\left(\cdot\right)$ is called the log-partition function. For linear, logistic, and Poisson models, for instance, we have log-partition functions $\psi_{\text{lin}}\left(t\right) = t^2/2\sigma^2$, $\psi_{\text{log}}\left(t\right) = \log\left(1+\exp\left(t\right)\right)$, and $\psi_{\text{Pois}}\left(t\right) = \exp\left(t\right)$, respectively.

5.4.1 Verifying SMRH for GLMs

Assuming that the log-partition function $\psi(\cdot)$ is twice continuously differentiable, the Hessian of $f(\cdot)$ is equal to

$$\nabla^2 f(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m} \psi''(\langle \mathbf{a}_i, \mathbf{x}\rangle)\, \mathbf{a}_i \mathbf{a}_i^{\mathrm{T}}.$$

Under the assumptions for GLMs, it can be shown that $\psi''(\cdot)$ is non-negative (i.e., $\psi(\cdot)$ is convex). For a given sparsity model generated by $\mathcal{C}_k$ let $\mathcal{S}$ be an arbitrary support set in $\mathcal{C}_k$ and suppose that $\operatorname{supp}(\mathbf{x}) \subseteq \mathcal{S}$ and $\|\mathbf{x}\| \leq r$. Furthermore, define

$$D_{\psi,r}(u) := \max_{t\in[-r,r]} \psi''(tu) \quad \text{and} \quad d_{\psi,r}(u) := \min_{t\in[-r,r]} \psi''(tu).$$

Using the Cauchy-Schwarz inequality we have $|\langle \mathbf{a}_i, \mathbf{x}\rangle| \leq r\|\mathbf{a}_i|_{\mathcal{S}}\|$ which implies

$$\frac{1}{m}\sum_{i=1}^{m} d_{\psi,r}(\|\mathbf{a}_i|_{\mathcal{S}}\|)\, \mathbf{a}_i|_{\mathcal{S}}\, \mathbf{a}_i|_{\mathcal{S}}^{\mathrm{T}} \preccurlyeq \nabla^2_{\mathcal{S}} f(\mathbf{x}) \preccurlyeq \frac{1}{m}\sum_{i=1}^{m} D_{\psi,r}(\|\mathbf{a}_i|_{\mathcal{S}}\|)\, \mathbf{a}_i|_{\mathcal{S}}\, \mathbf{a}_i|_{\mathcal{S}}^{\mathrm{T}}.$$

These matrix inequalities are precursors of (5.4). Imposing further restriction on the distribution of the covariate vectors $\{\mathbf{a}_i\}_{i=1}^m$ allows application of the results from random matrix theory regarding the extreme eigenvalues of random matrices (see e.g., (Tropp, 2012) and (Hsu et al., 2012)).

For example, following the same approach explained in Section 3.4, for the logistic model where $\psi \equiv \psi_{\log}$ we can show that $D_{\psi,r}(u) = \frac{1}{4}$ and $d_{\psi,r}(u) = \frac{1}{4}\operatorname{sech}^2\left(\frac{ru}{2}\right)$. Assuming that the covariate vectors are iid instances of a random vectors whose length almost surely bounded by one, we obtain $d_{\psi,r}(u) \geq \frac{1}{4}\operatorname{sech}^2\left(\frac{r}{2}\right)$. Using the matrix Chernoff inequality (Tropp, 2012) the extreme eigenvalues of $\frac{1}{m}\mathbf{A}_{\mathcal{S}}\mathbf{A}_{\mathcal{S}}^{\mathrm{T}}$ can be bounded with probability $1 - \exp(\log k - Cm)$ for some constant $C > 0$ (see Corollary 3.1 for detailed derivations). Using these results and taking the union bound over all $\mathcal{S} \in \mathcal{C}_k$ we obtain bounds for the extreme eigenvalues of $\nabla^2_{\mathcal{S}} f(\mathbf{x})$ that hold uniformly for all sets $\mathcal{S} \in \mathcal{C}_k$ with probability

$1-\exp\left(\log\left(k\left|\mathcal{C}_k\right|\right)-Cm\right)$. Thus (5.4) may hold if $m=O\left(\log\left(k\left|\mathcal{C}_k\right|\right)\right)$.

5.4.2 Approximation Error for GLMs

Suppose that the approximation error is measured with respect to $\mathbf{x}^{\perp}=\mathrm{P}_{\mathcal{C}_k,r}\left(\mathbf{x}^{\star}\right)$ where $\mathbf{x}^{\star}$ is the statistical truth in the considered GLM. It is desirable to further simplify the approximation error bound provided in Corollary 5.1 which is related to the statistical precision of the estimation problem. The corollary provides an approximation error that is proportional to $\left\|\nabla_{\mathcal{T}} f\left(\mathbf{x}^{\perp}\right)\right\|$ where $\mathcal{T}=\operatorname{supp}\left(\mathrm{P}_{\mathcal{C}_k^2,r}\left(\nabla f\left(\mathbf{x}^{\perp}\right)\right)\right)$. We can write

$$\nabla_{\mathcal{T}} f\left(\mathbf{x}^{\perp}\right)=\frac{1}{m}\sum_{i=1}^{m}\left(\psi'\left(\left\langle\mathbf{a}_i,\mathbf{x}^{\perp}\right\rangle\right)-y_i\right)\mathbf{a}_i|_{\mathcal{T}},$$

which yields $\left\|\nabla_{\mathcal{T}} f\left(\mathbf{x}^{\perp}\right)\right\|=\left\|\mathbf{A}_{\mathcal{T}}\mathbf{z}\right\|$ where $\mathbf{A}=\frac{1}{\sqrt{m}}\begin{bmatrix}\mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_m\end{bmatrix}$ and $\mathbf{z}|_{\{i\}}=z_i=\frac{\psi'\left(\left\langle\mathbf{a}_i,\mathbf{x}^{\perp}\right\rangle\right)-y_i}{\sqrt{m}}$. Therefore,

$$\left\|\nabla_{\mathcal{T}} f\left(\mathbf{x}^{\perp}\right)\right\|^2\leq\left\|\mathbf{A}_{\mathcal{T}}\right\|_{\mathrm{op}}^2\|\mathbf{z}\|^2,$$

where $\|\cdot\|_{\mathrm{op}}$ denotes the operator norm. Again using random matrix theory one can find an upper bound for $\left\|\mathbf{A}_{\mathcal{I}}\right\|_{\mathrm{op}}$ that holds uniformly for any $\mathcal{I}\in\mathcal{C}_k^2$ and in particular for $\mathcal{I}=\mathcal{T}$. Henceforth, $W>0$ is used to denote this upper bound.

The second term in the bound can be written as

$$\|\mathbf{z}\|^2=\frac{1}{m}\sum_{i=1}^{m}\left(\psi'\left(\left\langle\mathbf{a}_i,\mathbf{x}^{\perp}\right\rangle\right)-y_i\right)^2.$$

To further simplify this term we need to make assumptions about the log-partition function $\psi\left(\cdot\right)$ and/or the distribution of the covariate-response pair $(\mathbf{a},y)$. For instance, if $\psi'\left(\cdot\right)$ and the response variable y are bounded, as in the logistic model, then Hoeffding's inequality implies that for some small $\epsilon>0$ we have $\|\mathbf{z}\|^2\leq\mathbb{E}\left[\left(\psi'\left(\left\langle\mathbf{a},\mathbf{x}^{\perp}\right\rangle\right)-y\right)^2\right]+\epsilon$ with prob-

ability at least $1-\exp\left(-O\left(\epsilon^2 m\right)\right)$. Since in GLMs the true parameter $\mathbf{x}^\star$ is the minimizer of the expected loss $\mathbb{E}\left[\psi\left(\langle \mathbf{a}, \mathbf{x}\rangle\right) - y\left\langle \mathbf{a}, \mathbf{x}\right\rangle \mid \mathbf{a}\right]$ we deduce that $\mathbb{E}\left[\psi'\left(\langle \mathbf{a}, \mathbf{x}^\star\rangle\right) - y \mid \mathbf{a}\right] = 0$ and hence $\mathbb{E}\left[\psi'\left(\langle \mathbf{a}, \mathbf{x}^\star\rangle\right) - y\right] = 0$. Therefore,

$$\begin{aligned}
\|\mathbf{z}\|^2 &\leq \mathbb{E}\left[\mathbb{E}\left[\left(\psi'\left(\left\langle \mathbf{a}, \mathbf{x}^\perp\right\rangle\right) - \psi'\left(\langle \mathbf{a}, \mathbf{x}^\star\rangle\right) \;\; + \psi'\left(\langle \mathbf{a}, \mathbf{x}^\star\rangle\right) - y\right)^2 \mid \mathbf{a}\right]\right] + \epsilon \\
&\leq \mathbb{E}\left[\left(\psi'\left(\left\langle \mathbf{a}, \mathbf{x}^\perp\right\rangle\right) - \psi'\left(\langle \mathbf{a}, \mathbf{x}^\star\rangle\right)\right)^2\right] + \mathbb{E}\left[\left(\psi'\left(\langle \mathbf{a}, \mathbf{x}^\star\rangle\right) - y\right)^2\right] + \epsilon. \\
&= \underbrace{\mathbb{E}\left[\left(\psi'\left(\left\langle \mathbf{a}, \mathbf{x}^\perp\right\rangle\right) - \psi'\left(\langle \mathbf{a}, \mathbf{x}^\star\rangle\right)\right)^2\right]}_{\delta_1} + \underbrace{\operatorname{var}\left(\psi'\left(\langle \mathbf{a}, \mathbf{x}^\star\rangle\right) - y\right) + \epsilon}_{\sigma^2_{\text{stat}}}.
\end{aligned}$$

Then it follows from Corollary 5.1 and the fact that $\|\mathbf{A}|_{\mathcal{I}}\|_{\text{op}} \leq W$ that

$$\begin{aligned}
\left\|\mathbf{x}^{(t)} - \mathbf{x}^\star\right\| &\leq \left\|\mathbf{x}^{(t)} - \mathbf{x}^\perp\right\| + \underbrace{\left\|\mathbf{x}^\perp - \mathbf{x}^\star\right\|}_{\delta_2} \\
&\leq (2\gamma)^t \left\|\mathbf{x}^\perp\right\| + \frac{2\eta W}{1-2\gamma}\sigma^2_{\text{stat}} + \frac{2\eta W}{1-2\gamma}\delta_1 + \delta_2.
\end{aligned}$$

The total approximation error is comprised of two parts. The first part is due to statistical error that is given by $\frac{2\eta W}{1-2\gamma}\sigma^2_{\text{stat}}$, and $\frac{2\eta W}{1-2\gamma}\delta_1 + \delta_2$ is the second part of the error due to the bias that occurs because of an infeasible true parameter. The bias vanishes if the true parameter lies in the considered bounded sparsity model (i.e., $\mathbf{x}^\star = \mathrm{P}_{\mathcal{C}_k, r}\left(\mathbf{x}^\star\right)$).

5.5 Summary

We studied the projected gradient descent method for minimization of a real valued cost function defined over a finite-dimensional Hilbert space, under structured sparsity constraints. Using previously known combinatorial sparsity models, we define a sufficient condition for accuracy of the algorithm, the SMRH. Under this condition the algorithm converges to the desired optimum at a linear rate up to an approximation error. Unlike

the previous results on greedy-type methods that merely have focused on linear statistical models, our algorithm applies to a broader family of estimation problems. To provide an example, we examined application of the algorithm in estimation with GLMs. The approximation error can also be bounded by statistical precision and the potential bias. An interesting follow-up problem is to find whether the approximation error can be improved and the derived error is merely a by-product of requiring some form of restricted strong convexity through SMRH. Another problem of interest is to study the properties of the algorithm when the domain of the cost function is not finite-dimensional.

Chapter 6

Projected Gradient Descent for ℓ_p-constrained Least Squares

6.1 Background

As mentioned in Chapter 2, to avoid the combinatorial computational cost of (2.2), often the ℓ_0-norm is substituted by the ℓ_1-norm to reach at a convex program. More generally, one can approximate the ℓ_0-norm by an ℓ_p-norm $\|\mathbf{x}\|_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$ for some $p \in (0, 1]$ that yields the ℓ_p-minimization

$$\arg\min_{\mathbf{x}} \|\mathbf{x}\|_p \quad \text{s.t. } \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2 \leq \varepsilon.$$

Several theoretical and experimental results (see e.g., Chartrand, 2007a; Saab et al., 2008; Saab and Yilmaz, 2010) suggest that ℓ_p-minimization with $p \in (0, 1)$ has the advantage that it requires fewer observations than the ℓ_1-minimization to produce accurate estimates. However, ℓ_p-minimization is a non-convex problem for this range of p and finding the global minimizer is not guaranteed and can be computationally more expensive than the

ℓ_1-minimization.

An alternative approach in the framework of sparse linear regression is to solve the sparsity-constrained least squares problem

$$\arg\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{A}\mathbf{x}-\mathbf{y}\|_2^2 \quad \text{s.t. } \|\mathbf{x}\|_0 \leq s, \tag{6.1}$$

where $s = \|\mathbf{x}^\star\|_0$ is given. Similar to (2.2) solving (6.1) is not tractable and approximate solvers must be sought. Several CS algorithms jointly known as the *greedy pursuits* including Iterative Hard Thresholding (IHT) (Blumensath and Davies, 2009), Subspace Pursuit (SP) (Dai and Milenkovic, 2009), and Compressive Sampling Matching Pursuit (CoSaMP) (Needell and Tropp, 2009) are implicitly approximate solvers of (6.1).

As a relaxation of (6.1) one may also consider the ℓ_p-constrained least squares

$$\arg\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{A}\mathbf{x}-\mathbf{y}\|_2^2 \quad \text{s.t. } \|\mathbf{x}\|_p \leq R^\star, \tag{6.2}$$

given $R^\star = \|\mathbf{x}^\star\|_p$. The Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani, 1996) is a well-known special case of this optimization problem with $p = 1$. The optimization problem of (6.2) typically does not have a closed-form solution, but can be (approximately) solved using iterative PGD described in Algorithm 3. Previous studies of these algorithms, henceforth referred to as ℓ_p-PGD, are limited to the cases of $p = 0$ and $p = 1$. The algorithm corresponding to the case of $p = 0$ is recognized in the literature as the IHT algorithm. The Iterative Soft Thresholding (IST) algorithm (Beck and Teboulle, 2009) is originally proposed as a solver of the Basis Pursuit Denoising (BPDN) (Chen et al., 1998), which is the unconstrained equivalent of the LASSO with the ℓ_1-norm as the regularization term. However, the IST algorithm also naturally describes a PGD solver of (6.2) for $p = 1$ (see for e.g, Agarwal et al., 2010) by considering varying shrinkage in iterations, as described in (Beck and Teboulle, 2009), to enforce the iterates to have sufficiently

small ℓ_1-norm. The main contribution of this chapter is a comprehensive analysis of the performance of ℓ_p-PGD algorithms for the entire regime of $p \in [0, 1]$.

In the extreme case of $p = 0$ we have the ℓ_0-PGD algorithm which is indeed the IHT algorithm. Unlike conventional PGD algorithms, the feasible set—the set of points that satisfy the optimization constraints—for IHT is the non-convex set of s-sparse vectors. Therefore, the standard analysis for PGD algorithms with convex feasible sets that relies on the fact that projection onto convex sets defines a contraction map will no longer apply. However, imposing extra conditions on the matrix $\mathbf{A}$ can be leveraged to provide convergence guarantees (Blumensath and Davies, 2009; Foucart, 2012).

At $p = 1$ where (6.2) is a convex program, the corresponding ℓ_1-PGD algorithm has been studied under the name of IST in different scenarios (see Beck and Teboulle, 2009, and references therein). Ignoring the sparsity of the vector $\mathbf{x}^\star$, it can be shown that the IST algorithm exhibits a sublinear rate of convergence as a convex optimization algorithm (Beck and Teboulle, 2009). In the context of the *sparse* estimation problems, however, faster rates of convergence can be guaranteed for IST. For example, in (Agarwal et al., 2010) PGD algorithms are studied in a broad category of regression problems regularized with "decomposable" norms. In this configuration, which includes sparse linear regression via IST, the PGD algorithms are shown to possess a linear rate of convergence provided the objective function—the squared error in our case—satisfies *Restricted Strong Convexity* (RSC) and *Restricted Smoothness* (RSM) conditions (Agarwal et al., 2010). Although the results provided in (Agarwal et al., 2010) consolidate the analysis of several interesting problems, they do not readily extend to the case of ℓ_p-constrained least squares since the constraint is not defined by a true norm.

In this chapter, by considering ℓ_p-balls of given radii as feasible sets in the general case, we study the ℓ_p-PGD algorithms that render a continuum of sparse reconstruction algo-

rithms, and encompass both the IHT and the IST algorithms. Note that in this chapter we consider the observation model (2.1) with the signal, the measurement matrix, the observations, and the noise having complex valued entries, i.e., $\mathbf{x}^\star \in \mathbb{C}^n$, $\mathbf{A} \in \mathbb{C}^{m\times n}$, $\mathbf{y} \in \mathbb{C}^m$, and $\mathbf{e} \in \mathbb{C}^m$. Our results suggest that as p increases from zero to one the convergence and robustness to noise deteriorates. This conclusion is particularly in agreement with the empirical studies of the *phase transition* of the IST and IHT algorithms provided in (Maleki and Donoho, 2010). Our results for ℓ_0-PGD coincides with the guarantees for IHT derived in (Foucart, 2012). Furthermore, to the best of our knowledge the RIP-based accuracy guarantees we provide for IST, which is the ℓ_1-PGD algorithm, have not been derived before.

6.2 Projected Gradient Descent for ℓ_p-constrained Least Squares

In a broad range of applications where the objective function is the squared error of the form $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{A}\mathbf{x}-\mathbf{y}\|_2^2$, the iterate update equation of the PGD method outlined in Algorithm 3 reduces to

$$\mathbf{x}^{(t+1)} = \mathrm{P}_{\mathcal{Q}}\left(\mathbf{x}^{(t)} - \eta^{(t)}\mathbf{A}^{\mathsf{H}}\left(\mathbf{A}\mathbf{x}^{(t)}-\mathbf{y}\right)\right).$$

In the context of compressed sensing if (2.1) holds and $\mathcal{Q}$ is the ℓ_1-ball of radius $\|\mathbf{x}^\star\|_1$ centered at the origin, Algorithm 3 reduces to the IST algorithm (except perhaps for variable step-size) that solves (6.2) for $p = 1$. By relaxing the convexity restriction imposed on $\mathcal{Q}$ the PGD iterations also describe the IHT algorithm where $\mathcal{Q}$ is the set of vectors whose ℓ_0-norm is not greater than $s = \|\mathbf{x}^\star\|_0$.

Henceforth, we refer to an ℓ_p-ball centered at the origin and aligned with the axes

simply as an ℓ_p-ball for brevity. To proceed let us define the set

$$\mathcal{F}_p(c) = \left\{ \mathbf{x} \in \mathbb{C}^n \mid \sum_{i=1}^{n} |x_i|^p \leq c \right\},$$

for $c \in \mathbb{R}^+$, which describes an ℓ_p-ball. Although c can be considered as the radius of this ℓ_p-ball with respect to the metric $d(\mathbf{a}, \mathbf{b}) = \|\mathbf{a} - \mathbf{b}\|_p^p$, we call c the "p-radius" of the ℓ_p-ball to avoid confusion with the conventional definition of the radius for an ℓ_p-ball, i.e., $\max_{\mathbf{x} \in \mathcal{F}_p(c)} \|\mathbf{x}\|_p$. Furthermore, at $p = 0$ where $\mathcal{F}_p(c)$ describes the same "ℓ_0-ball" for different values of c, we choose the smallest c as the p-radius of the ℓ_p-ball for uniqueness. In this section we will show that to estimate the signal $\mathbf{x}^\star$ that is either sparse or *compressible* in fact the PGD method can be applied in a more general framework where the feasible set is considered to be an ℓ_p-ball of given p-radius. Ideally the p-radius of the feasible set should be $\|\mathbf{x}^\star\|_p^p$, but in practice this information might not be available. In our analysis, we merely assume that the p-radius of the feasible set is not greater than $\|\mathbf{x}^\star\|_p^p$, i.e., the feasible set does not contain $\mathbf{x}^\star$ in its interior.

Note that for the feasible sets $\mathcal{Q} \equiv \mathcal{F}_p(c)$ with $p \in (0, 1]$ the minimum value in (5.2) is always attained because the objective is continuous and the set $\mathcal{Q}$ is compact. Therefore, there is at least one minimizer in $\mathcal{Q}$. However, for $p < 1$ the set $\mathcal{Q}$ is nonconvex and there might be multiple projection points in general. For the purpose of the analysis presented in this chapter, however, any such minimizer is acceptable. Using the axiom of choice, we can assume existence of a choice function that for every $\mathbf{x}$ selects one of the solutions of (5.2). This function indeed determines a projection operator which we denote by $\mathrm{P}_{\mathcal{Q}}(\mathbf{x})$.

Many compressed sensing algorithms such as those of Blumensath and Davies (2009); Dai and Milenkovic (2009); Needell and Tropp (2009); Candès (2008) rely on sufficient conditions expressed in terms of the RIP of the matrix $\mathbf{A}$. We also provide accuracy guarantees of the ℓ_p-PGD algorithm with the assumption that certain RIP conditions hold. The

following definition states the RIP in its asymmetric form. This definition is previously proposed in the literature (Foucart and Lai, 2009), though in a slightly different format.

Definition (RIP)**.** Matrix $\mathbf{A}$ is said to have RIP of order k with restricted isometry constants α_k and β_k if they are in order the smallest and the largest non-negative numbers such that

$$\beta_k \|\mathbf{x}\|_2^2 \leq \|\mathbf{A}\mathbf{x}\|_2^2 \leq \alpha_k \|\mathbf{x}\|_2^2$$

holds for all k-sparse vectors $\mathbf{x}$.

In the literature usually the symmetric form of the RIP is considered in which $\alpha_k = 1+\delta_k$ and $\beta_k = 1-\delta_k$ with $\delta_k \in [0,1]$. For example, in (Foucart, 2012) the ℓ_1-minimization is shown to accurately estimate $\mathbf{x}^\star$ provided $\delta_{2s} < 3/\left(4+\sqrt{6}\right) \approx 0.46515$. Similarly, accuracy of the estimates obtained by IHT, SP, and CoSaMP are guaranteed provided $\delta_{3s} < 1/2$ (Foucart, 2012), $\delta_{3s} < 0.205$ (Dai and Milenkovic, 2009), and $\delta_{4s} < \sqrt{2/\left(5+\sqrt{73}\right)} \approx 0.38427$ (Foucart, 2012), respectively.

As our first contribution, in the following theorem we show that the ℓ_p-PGD accurately solves ℓ_p-constrained least squares provided the matrix $\mathbf{A}$ satisfies a proper RIP criterion. To proceed we define

$$\rho_s = \frac{\alpha_s - \beta_s}{\alpha_s + \beta_s},$$

which can be interpreted as the equivalent of the standard symmetric RIP constant δ_s.

Theorem 6.1. *Let $\mathbf{x}^\star$ be an s-sparse vector whose compressive measurements are observed according to (2.1) using a measurement matrix $\mathbf{A}$ that satisfies RIP of order $3s$. To estimate $\mathbf{x}^\star$ via the ℓ_p-PGD algorithm an ℓ_p-ball $\widehat{\mathcal{B}}$ with p-radius $\widehat{c}$ (i.e., $\widehat{\mathcal{B}} = \mathcal{F}_p(\widehat{c})$) is given as the feasible set for the algorithm such that $\widehat{c} = (1-\epsilon)^p \|\mathbf{x}^\star\|_p^p$ for some[1] $\epsilon \in [0,1)$. Furthermore, suppose that*

[1] At $p = 0$ we have $(1-\epsilon)^0 = 1$ which enforces $\widehat{c} = \|\mathbf{x}^\star\|_0$. In this case ϵ is not unique, but to make a coherent statement we assume that $\epsilon = 0$.

the step-size $\eta^{(t)}$ *of the algorithm can be chosen to obey* $\left|\frac{\eta^{(t)}(\alpha_{3s}+\beta_{3s})}{2}-1\right|\leq\tau$ *for some* $\tau\geq 0$. *If*

$$(1+\tau)\,\rho_{3s}+\tau<\frac{1}{2\left(1+\sqrt{2}\xi\left(p\right)\right)^{2}} \tag{6.3}$$

with $\xi\left(p\right)$ *denoting the function* $\sqrt{p}\left(\frac{2}{2-p}\right)^{1/2-1/p}$, *then* $\mathbf{x}^{(t)}$, *the* t*-th iterate of the algorithm, obeys*

$$\begin{aligned}\left\|\mathbf{x}^{(t)}-\mathbf{x}^{\star}\right\|_{2}\leq\left(2\gamma\right)^{t}\left\|\mathbf{x}^{\star}\right\|_{2}\\ +\frac{2\left(1+\tau\right)}{1-2\gamma}\left(1+\xi\left(p\right)\right)\left(\epsilon\left(1+\rho_{3s}\right)\left\|\mathbf{x}^{\star}\right\|_{2}+\frac{2\sqrt{\alpha_{2s}}}{\alpha_{3s}+\beta_{3s}}\|\mathbf{e}\|_{2}\right)+\epsilon\|\mathbf{x}^{\star}\|_{2},\end{aligned} \tag{6.4}$$

where

$$\gamma=\left(\left(1+\tau\right)\rho_{3s}+\tau\right)\left(1+\sqrt{2}\xi\left(p\right)\right)^{2}. \tag{6.5}$$

Remark 6.1. Note that the parameter ϵ indicates how well the feasible set $\widehat{\mathcal{B}}$ approximates the ideal feasible set $\mathcal{B}^{\star}=\mathcal{F}_{p}\left(\|\mathbf{x}^{\star}\|_{p}^{p}\right)$. The terms in (6.4) that depend on ϵ determine the error caused by the mismatch between $\widehat{\mathcal{B}}$ and $\mathcal{B}^{\star}$. Ideally, one has $\epsilon=0$ and the residual error becomes merely dependent on the noise level $\|\mathbf{e}\|_{2}$.

Remark 6.2. The parameter τ determines the deviation of the step-size $\eta^{(t)}$ from $\frac{2}{\alpha_{3s}+\beta_{3s}}$ which might not be known *a priori*. In this formulation, smaller values of τ are desirable since they impose less restrictive condition on ρ_{3s} and also result in smaller residual error. Furthermore, we can naively choose $\eta^{(t)}=\|\mathbf{A}\mathbf{x}\|_{2}^{2}/\|\mathbf{x}\|_{2}^{2}$ for some $3s$-sparse vector $\mathbf{x}\neq\mathbf{0}$ to ensure $1/\alpha_{3s}\leq\eta^{(t)}\leq 1/\beta_{3s}$ and thus $\left|\eta^{(t)}\frac{\alpha_{3s}+\beta_{3s}}{2}-1\right|\leq\frac{\alpha_{3s}-\beta_{3s}}{2\beta_{3s}}$. Therefore, we can always assume that $\tau\leq\frac{\alpha_{3s}-\beta_{3s}}{2\beta_{3s}}$.

Remark 6.3. Note that the function $\xi\left(p\right)$, depicted in Fig. 6.1, controls the variation of the stringency of the condition (6.3) and the variation of the residual error in (6.4) in terms of p. Straightforward algebra shows that $\xi\left(p\right)$ is an increasing function of p with

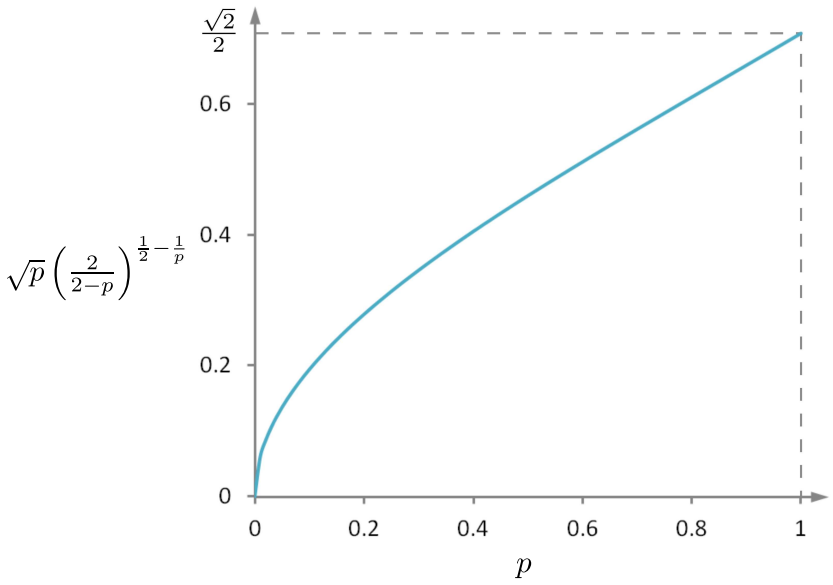


Figure 6.1: Plot of the function $\xi(p) = \sqrt{p}\left(\frac{2}{2-p}\right)^{\frac{1}{2}-\frac{1}{p}}$ which determines the contraction factor and the residual error.

$\xi(0) = 0$. Therefore, as p increases from zero to one, the RHS of (6.3) decreases, which implies the measurement matrix must have a smaller ρ_{3s} to satisfy the sufficient condition (6.3). Similarly, as p increases from zero to one the residual error in (6.4) increases. To contrast this result with the existing guarantees of other iterative algorithms, suppose that $\tau = 0$, $\epsilon = 0$, and we use the symmetric form of RIP (i.e., $\alpha_{3s} = 1 + \delta_{3s}$ and $\beta_{3s} = 1 - \delta_{3s}$) which implies $\rho_{3s} = \delta_{3s}$. At $p = 0$, corresponding to the IHT algorithm, (6.3) reduces to $\delta_{3s} < 1/2$ that is identical to the condition derived in (Foucart, 2012). Furthermore, the required condition at $p = 1$, corresponding to the IST algorithm, would be $\delta_{3s} < 1/8$.

The guarantees stated in Theorem 6.1 can be generalized for nearly sparse or *compressible* signals that can be defined using power laws as described in (Candès and Tao, 2006). The following corollary provides error bounds for a general choice of $\mathbf{x}^\star$.

Corollary 6.1. *Suppose that $\mathbf{x}^\star$ is an arbitrary vector in $\mathbb{C}^n$ and the conditions of Theorem 6.1*

hold for $\mathbf{x}_s^\star$, then the t-th iterate of the ℓ_p-PGD algorithm provides an estimate of $\mathbf{x}_s^\star$ that obeys

$$\left\|\mathbf{x}^{(t)}-\mathbf{x}^\star\right\|_2 \leq (2\gamma)^t \left\|\mathbf{x}_s^\star\right\|_2 + \frac{2\left(1+\tau\right)\left(1+\xi\left(p\right)\right)}{1-2\gamma}\left(\epsilon\left(1+\rho_{3s}\right)\left\|\mathbf{x}_s^\star\right\|_2 + \frac{2\alpha_{2s}}{\alpha_{3s}+\beta_{3s}}\left(\|\mathbf{x}^\star-\mathbf{x}_s^\star\|_2 + \|\mathbf{x}^\star-\mathbf{x}_s^\star\|_1/\sqrt{2s}\right) + \frac{2\sqrt{\alpha_{2s}}}{\alpha_{3s}+\beta_{3s}}\|\mathbf{e}\|_2\right) + \epsilon\|\mathbf{x}_s^\star\|_2 + \|\mathbf{x}^\star-\mathbf{x}_s^\star\|_2.$$

Proof. Let $\widetilde{\mathbf{e}} = \mathbf{A}\left(\mathbf{x}^\star-\mathbf{x}_s^\star\right)+\mathbf{e}$. We can write $\mathbf{y}=\mathbf{A}\mathbf{x}^\star+\mathbf{e}=\mathbf{A}\mathbf{x}_s^\star+\widetilde{\mathbf{e}}$. Thus, we can apply Theorem 6.1 considering $\mathbf{x}_s^\star$ as the signal of interest and $\widetilde{\mathbf{e}}$ as the noise vector and obtain

$$\left\|\mathbf{x}^{(t)}-\mathbf{x}_s^\star\right\|_2 \leq (2\gamma)^t \left\|\mathbf{x}_s^\star\right\|_2 + \frac{2\left(1+\tau\right)}{1-2\gamma}\left(1+\xi\left(p\right)\right)\left(\epsilon\left(1+\rho_{3s}\right)\left\|\mathbf{x}_s^\star\right\|_2 + \frac{2\sqrt{\alpha_{2s}}}{\alpha_{3s}+\beta_{3s}}\|\widetilde{\mathbf{e}}\|_2\right) + \epsilon\|\mathbf{x}_s^\star\|_2. \tag{6.6}$$

Furthermore, we have

$$\begin{aligned}\|\widetilde{\mathbf{e}}\|_2 &= \|\mathbf{A}\left(\mathbf{x}^\star-\mathbf{x}_s^\star\right)+\mathbf{e}\|_2\\ &\leq \|\mathbf{A}\left(\mathbf{x}^\star-\mathbf{x}_s^\star\right)\|_2+\|\mathbf{e}\|_2.\end{aligned}$$

Then applying Proposition 3.5 of (Needell and Tropp, 2009) yields

$$\|\widetilde{\mathbf{e}}\|_2 \leq \sqrt{\alpha_{2s}}\left(\|\mathbf{x}^\star-\mathbf{x}_s^\star\|_2+\frac{1}{\sqrt{2s}}\|\mathbf{x}^\star-\mathbf{x}_s^\star\|_1\right)+\|\mathbf{e}\|_2.$$

Applying this inequality in (6.6) followed by the triangle inequality

$$\left\|\mathbf{x}^{(t)}-\mathbf{x}^\star\right\|_2 \leq \left\|\mathbf{x}^{(t)}-\mathbf{x}_s^\star\right\|_2+\|\mathbf{x}^\star-\mathbf{x}_s^\star\|_2$$

yields the desired inequality. ∎

6.3 Discussion

In this chapter we studied the accuracy of the Projected Gradient Descent algorithm in solving sparse least squares problems where sparsity is dictated by an ℓ_p-norm constraint. Assuming that one has an algorithm that can find a projection of any given point onto ℓ_p-balls with $p \in [0, 1]$, we have shown that the PGD method converges to the true signal, up to the statistical precision, at a linear rate. The convergence guarantees in this chapter are obtained by requiring proper RIP conditions to hold for the measurement matrix. By varying p from zero to one, these sufficient conditions become more stringent while robustness to noise and convergence rate worsen. This behavior suggests that smaller values of p are preferable, and in fact the PGD method at $p = 0$ (i.e., the IHT algorithm) outperforms the PGD method at $p > 0$ in every aspect. These conclusions, however, are not definitive as we have merely presented sufficient conditions for accuracy of the PGD method.

Unfortunately and surprisingly, for $p \in (0, 1)$ the algorithm for projection onto ℓ_p-balls is not as simple as the cases of $p = 0$ and $p = 1$, leaving practicality of the algorithm unclear for the intermediate values p. We have shown in the Appendix D that a projection $\mathbf{x}^\perp$ of point $\mathbf{x} \in \mathbb{C}^n$ has the following properties

(i) $\left|x_i^\perp\right| \leq |x_i|$ for all $i \in [n]$ while there is at most one $i \in [n]$ such that $\left|x_i^\perp\right| < \frac{1-p}{2-p}\,|x_i|$,

(ii) $\text{Arg}\,(x_i) = \text{Arg}\left(x_i^\perp\right)$ for $i \in [n]$,

(iii) if $|x_i| > |x_j|$ for some $i, j \in [n]$ then $\left|x_i^\perp\right| \geq \left|x_j^\perp\right|$, and

(iv) there exist $\lambda \geq 0$ such that for all $i \in \text{supp}\left(\mathbf{x}^\perp\right)$ we have $\left|x_i^\perp\right|^{1-p}\left(|x_i| - \left|x_i^\perp\right|\right) = p\lambda$.

However, these properties are not sufficient for full characterization of a projection. One may ask that if the PGD method performs the best at $p = 0$ then why is it important at all to design a projection algorithm for $p > 0$? We believe that developing an efficient algo-

rithm for projection onto ℓ_p-balls with $p \in (0, 1)$ is an interesting problem that can provide a building block for other methods of sparse signal estimation involving the ℓ_p-norm. Furthermore, studying this problem may help to find an insight on how the complexity of these algorithms vary in terms of p.

In future work, we would like to examine the performance of more sophisticated first-order methods such as the Nesterov's optimal gradient methods (Nesterov, 2004) for ℓ_p-constrained least squares problems. Finding a computationally efficient way to solve the non-convex projection could also help to further understand non-convex CS algorithms and their performance. Furthermore, it could be possible to extend the provided framework further to analyze ℓ_p-constrained minimization with objective functions other than the squared error. This generalized framework can be used in problems such as regression with GLMs that arise in statistics and machine learning.

Chapter 7

Conclusion and Future Work

In this thesis, we studied sparsity-constrained optimization problems and proposed a number of greedy algorithms as approximate solvers for these problems. Unlike the existing convex programming methods, the proposed greedy methods do not require the objective to be convex everywhere and produce a solution that is exactly sparse. We showed that if the objective function has well-behaved second order variations, namely if it obeys the SRH or the SRL conditions, then our proposed algorithms provide accurate solutions. Some of these algorithms are also examined through simulations for the 1-bit CS problem and sparse logistic regression. In our work the minimization of functions subject to structured sparsity is also addressed. Assuming the objective function obeys a variant of the SRH condition tailored for model-based sparsity, we showed that a non-convex PGD method can produce and accurate estimate of the underlying parameter.

In high-dimensional estimation problems one of the important challenges is the computational complexity of the algorithms. One solution to this problem is to introduce randomization in the algorithm in order to reduce the cost of evaluating the function or its derivatives. It is also possible to reformulate the algorithm in stochastic optimization

framework to not only simplify the iterations, but also address scenarios with streaming data. In future work, it would be interesting to study these aspects in our proposed algorithms. Furthermore, it would be interesting to prove accuracy guarantees of the algorithms based on sufficient conditions that are less stringent that SRH or SRL. For example, it may be possible to measure accuracy in metrics other than the ℓ_2-error and thus one might require conditions similar to SRH or SRL, but with bounds defined using another appropriately chosen metric.

We also studied the problem of ℓ_p-constrained least squares under the RIP assumption. In particular, we showed that if one can perform projection onto a given ℓ_p-ball efficiently, then PGD method provides an accurate solution to the non-convex ℓ_p-constrained least squares. Our results suggest that the corresponding algorithm at $p = 0$ outperforms the algorithm for any other choice of $p \in (0, 1]$. Nevertheless, study of this algorithm reveals an interesting problem: while there are computationally tractable algorithms for projection onto "ℓ_0-ball" and ℓ_1-ball, computational complexity of projection onto an ℓ_p-ball is still unknown. We derived the necessary conditions for a point to be the projection of any given point on an ℓ_p-ball. Furthermore, based on limited numerical observations we conjecture that the desired projection is indeed tractable. Proving this open problem is an interesting topic for future work as it can help to better understand the computational complexity of the other non-convex CS algorithms that involve the ℓ_p-norms.

Appendix A

Proofs of Chapter 3

A.1 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 A.1 and A.2 are used to prove Lemmas A.1 and A.2. These Lemmas then are used to prove Lemma A.3 that provides an iteration invariant which in turn yields the main result.

Proposition A.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

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

and

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

which imply the desired result. ∎

Proposition A.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*

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

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

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

and thereby

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

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

$$\left\|\widetilde{\mathbf{M}}(t)\right\| \leq \left\|\mathbf{M}(t) - \frac{A(t)+B(t)}{2}\mathbf{I}\right\| \leq \frac{A(t)-B(t)}{2}. \tag{A.1}$$

Finally, it follows from the convexity of the operator norm, Jensen's inequality, and (A.1) that

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

as desired. ∎

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 (3.2) and (3.3), respectively.

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

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

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

$$\left\|\mathbf{z}|_{\mathcal{R}\setminus\mathcal{Z}}\right\|_2 \leq \left\|\mathbf{z}|_{\mathcal{Z}\setminus\mathcal{R}}\right\|_2. \tag{A.2}$$

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

$$\begin{aligned} \left\|\mathbf{z}|_{\mathcal{R}\setminus\mathcal{Z}}\right\|_2 &\geq \left\|\nabla f(\widehat{\mathbf{x}})|_{\mathcal{R}\setminus\mathcal{Z}} - \nabla f(\mathbf{x}^\star)|_{\mathcal{R}\setminus\mathcal{Z}}\right\|_2 - \left\|\nabla f(\mathbf{x}^\star)|_{\mathcal{R}\setminus\mathcal{Z}}\right\|_2 \\ &= \left\| \left(\int_0^1 \mathbf{P}_{\mathcal{R}\setminus\mathcal{Z}}^{\mathrm{T}} \nabla^2 f(t\widehat{\mathbf{x}} + (1-t)\,\mathbf{x}^\star)\,\mathrm{d}t \right) (\widehat{\mathbf{x}} - \mathbf{x}^\star) \right\|_2 - \left\|\nabla f(\mathbf{x}^\star)|_{\mathcal{R}\setminus\mathcal{Z}}\right\|_2 \end{aligned}$$

$$\geq \left\| \left(\int_0^1 \mathbf{P}_{\mathcal{R}\setminus\mathcal{Z}}^{\mathrm{T}} \nabla^2 f\left(t\widehat{\mathbf{x}} + (1-t)\,\mathbf{x}^\star\right) \mathbf{P}_{\mathcal{R}\setminus\mathcal{Z}} \mathrm{d}t \right) \left(\widehat{\mathbf{x}} - \mathbf{x}^\star\right)|_{\mathcal{R}\setminus\mathcal{Z}} \right\|_2 - \left\|\nabla f\left(\mathbf{x}^\star\right)|_{\mathcal{R}\setminus\mathcal{Z}}\right\|_2$$
$$- \left\| \left(\int_0^1 \mathbf{P}_{\mathcal{R}\setminus\mathcal{Z}}^{\mathrm{T}} \nabla^2 f\left(t\widehat{\mathbf{x}} + (1-t)\,\mathbf{x}^\star\right) \mathbf{P}_{\mathcal{Z}\cap\mathcal{R}} \mathrm{d}t \right) \left(\widehat{\mathbf{x}} - \mathbf{x}^\star\right)|_{\mathcal{Z}\cap\mathcal{R}} \right\|_2,$$

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

$$\begin{aligned}\left\|\mathbf{z}|_{\mathcal{R}\setminus\mathcal{Z}}\right\|_2 &\geq \beta_{2s}\left(\widehat{\mathbf{x}},\mathbf{x}^\star\right)\left\|\left(\widehat{\mathbf{x}}-\mathbf{x}^\star\right)|_{\mathcal{R}\setminus\mathcal{Z}}\right\|_2 - \frac{\gamma_{2s}\left(\widehat{\mathbf{x}},\mathbf{x}^\star\right)}{2}\left\|\left(\widehat{\mathbf{x}}-\mathbf{x}^\star\right)|_{\mathcal{Z}\cap\mathcal{R}}\right\|_2 - \left\|\nabla f\left(\mathbf{x}^\star\right)|_{\mathcal{R}\setminus\mathcal{Z}}\right\|_2 \\ &\geq \beta_{2s}\left(\widehat{\mathbf{x}},\mathbf{x}^\star\right)\left\|\left(\widehat{\mathbf{x}}-\mathbf{x}^\star\right)|_{\mathcal{R}\setminus\mathcal{Z}}\right\|_2 - \frac{\gamma_{2s}\left(\widehat{\mathbf{x}},\mathbf{x}^\star\right)}{2}\left\|\widehat{\mathbf{x}}-\mathbf{x}^\star\right\|_2 - \left\|\nabla f\left(\mathbf{x}^\star\right)|_{\mathcal{R}\setminus\mathcal{Z}}\right\|_2. \end{aligned} \tag{A.3}$$

Similarly, we have

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

Combining (A.2), (A.3), and (A.4) we obtain

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

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

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

which proves the claim. ∎

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

$$\mathbf{b}=\arg\min f\left(\mathbf{x}\right)\ \text{s.t. }\mathbf{x}|_{\mathcal{T}^c}=0 \tag{A.5}$$

satisfies

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

Proof. We have

$$\nabla f\left(\mathbf{x}^{\star}\right)-\nabla f\left(\mathbf{b}\right)=\int_0^1\nabla^2 f\left(t\mathbf{x}^{\star}+(1-t)\,\mathbf{b}\right)\mathrm{d}t\ \left(\mathbf{x}^{\star}-\mathbf{b}\right).$$

Furthermore, since $\mathbf{b}$ is the solution to (A.5) we must have $\nabla f\left(\mathbf{b}\right)|_{\mathcal{T}}=0$. Therefore,

$$\begin{aligned}\nabla f\left(\mathbf{x}^{\star}\right)|_{\mathcal{T}}&=\left(\int_0^1\mathbf{P}_{\mathcal{T}}^{\mathrm{T}}\nabla^2 f\left(t\mathbf{x}^{\star}+(1-t)\,\mathbf{b}\right)\mathrm{d}t\right)\left(\mathbf{x}^{\star}-\mathbf{b}\right)\\&=\left(\int_0^1\mathbf{P}_{\mathcal{T}}^{\mathrm{T}}\nabla^2 f\left(t\mathbf{x}^{\star}+(1-t)\,\mathbf{b}\right)\mathbf{P}_{\mathcal{T}}\mathrm{d}t\right)\left(\mathbf{x}^{\star}-\mathbf{b}\right)|_{\mathcal{T}}\\&+\left(\int_0^1\mathbf{P}_{\mathcal{T}}^{\mathrm{T}}\nabla^2 f(t\mathbf{x}^{\star}+(1-t)\,\mathbf{b})\,\mathbf{P}_{\mathcal{T}^c}\mathrm{d}t\right)\left(\mathbf{x}^{\star}-\mathbf{b}\right)|_{\mathcal{T}^c}.\end{aligned}\tag{A.6}$$

Since f has μ_{4s}-SRH and $|\mathcal{T}\cup\operatorname{supp}\left(t\mathbf{x}^{\star}+(1-t)\,\mathbf{b}\right)|\leq 4s$ for all $t\in[0,1]$, functions $A_{4s}\left(\cdot\right)$ and $B_{4s}\left(\cdot\right)$, defined using (3.2) and (3.3), exist such that we have

$$B_{4s}\left(t\mathbf{x}^{\star}+(1-t)\,\mathbf{b}\right)\leq\lambda_{\min}\left(\mathbf{P}_{\mathcal{T}}^{\mathrm{T}}\nabla^2 f\left(t\mathbf{x}^{\star}+(1-t)\,\mathbf{b}\right)\mathbf{P}_{\mathcal{T}}\right)$$

and

$$A_{4s}\left(t\mathbf{x}^{\star}+\left(1-t\right)\mathbf{b}\right)\geq\lambda_{\max}\left(\mathbf{P}_{\mathcal{T}}^{\mathrm{T}}\nabla^{2}f\left(t\mathbf{x}^{\star}+\left(1-t\right)\mathbf{b}\right)\mathbf{P}_{\mathcal{T}}\right).$$

Thus, from Proposition A.1 we obtain

$$\beta_{4s}\left(\mathbf{b},\mathbf{x}^{\star}\right)\leq\lambda_{\min}\left(\int_{0}^{1}\mathbf{P}_{\mathcal{T}}^{\mathrm{T}}\nabla^{2}f\left(t\mathbf{x}^{\star}+\left(1-t\right)\mathbf{b}\right)\mathbf{P}_{\mathcal{T}}\mathrm{d}t\right)$$

and

$$\alpha_{4s}\left(\mathbf{b},\mathbf{x}^{\star}\right)\geq\lambda_{\max}\left(\int_{0}^{1}\mathbf{P}_{\mathcal{T}}^{\mathrm{T}}\nabla^{2}f\left(t\mathbf{x}^{\star}+\left(1-t\right)\mathbf{b}\right)\mathbf{P}_{\mathcal{T}}\mathrm{d}t\right).$$

This result implies that the matrix $\int_0^1 \mathbf{P}_{\mathcal{T}}^{\mathrm{T}}\nabla^{2}f\left(t\mathbf{x}^{\star}+\left(1-t\right)\mathbf{b}\right)\mathbf{P}_{\mathcal{T}}\mathrm{d}t$, henceforth denoted by $\mathbf{W}$, is invertible and

$$\frac{1}{\alpha_{4s}\left(\mathbf{b},\mathbf{x}^{\star}\right)}\leq\lambda_{\min}\left(\mathbf{W}^{-1}\right)\leq\lambda_{\max}\left(\mathbf{W}^{-1}\right)\leq\frac{1}{\beta_{4s}\left(\mathbf{b},\mathbf{x}^{\star}\right)},\tag{A.7}$$

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 (A.6) by $\mathbf{W}^{-1}$ obtain

$$\mathbf{W}^{-1}\nabla f\left(\mathbf{x}^{\star}\right)|_{\mathcal{T}}=\left(\mathbf{x}^{\star}-\mathbf{b}\right)|_{\mathcal{T}}+\mathbf{W}^{-1}\left(\int_{0}^{1}\mathbf{P}_{\mathcal{T}}^{\mathrm{T}}\nabla^{2}f(t\mathbf{x}^{\star}+(1-t)\,\mathbf{b})\,\mathbf{P}_{\mathcal{T}^{c}}\mathrm{d}t\right)\mathbf{x}^{\star}|_{\mathcal{T}^{c}},$$

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, (A.7), and Proposition A.2 then we obtain

$$\begin{aligned}\left\|\mathbf{x}^{\star}|_{\mathcal{T}}-\mathbf{b}\right\|_{2}&=\left\|\left(\mathbf{x}^{\star}-\mathbf{b}\right)|_{\mathcal{T}}\right\|_{2}\\&\leq\left\|\mathbf{W}^{-1}\left(\int_{0}^{1}\mathbf{P}_{\mathcal{T}}^{\mathrm{T}}\nabla^{2}f(t\mathbf{x}^{\star}+(1-t)\,\mathbf{b})\,\mathbf{P}_{\mathcal{T}^{c}\cap\mathcal{S}^{\star}}\mathrm{d}t\right)\mathbf{x}^{\star}|_{\mathcal{T}^{c}\cap\mathcal{S}^{\star}}\right\|_{2}+\left\|\mathbf{W}^{-1}\nabla f(\mathbf{x}^{\star})\,|_{\mathcal{T}}\right\|_{2}\end{aligned}$$

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

as desired. ∎

Lemma A.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:*

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

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

$$\|\mathbf{x}^{\star}|_{\mathcal{T}^c}\|_2 \leq \frac{\gamma_{4s}(\widehat{\mathbf{x}},\mathbf{x}^{\star})+\gamma_{2s}(\widehat{\mathbf{x}},\mathbf{x}^{\star})}{2\beta_{2s}(\widehat{\mathbf{x}},\mathbf{x}^{\star})}\|\widehat{\mathbf{x}}-\mathbf{x}^{\star}\|_2+\frac{\left\|\nabla f(\mathbf{x}^{\star})|_{\mathcal{R}\setminus\mathcal{Z}}\right\|_2+\left\|\nabla f(\mathbf{x}^{\star})|_{\mathcal{Z}\setminus\mathcal{R}}\right\|_2}{\beta_{2s}(\widehat{\mathbf{x}},\mathbf{x}^{\star})}. \tag{A.8}$$

Furthermore,

$$\begin{aligned}\|\mathbf{b}_s-\mathbf{x}^{\star}\|_2 &\leq \|\mathbf{b}_s-\mathbf{x}^{\star}|_{\mathcal{T}}\|_2+\|\mathbf{x}^{\star}|_{\mathcal{T}^c}\|_2 \\ &\leq \|\mathbf{x}^{\star}|_{\mathcal{T}}-\mathbf{b}\|_2+\|\mathbf{b}_s-\mathbf{b}\|_2+\|\mathbf{x}^{\star}|_{\mathcal{T}^c}\|_2 \leq 2\|\mathbf{x}^{\star}|_{\mathcal{T}}-\mathbf{b}\|_2+\|\mathbf{x}^{\star}|_{\mathcal{T}^c}\|_2,\end{aligned} \tag{A.9}$$

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

$$\|\mathbf{b}_s-\mathbf{x}^{\star}\|_2 \leq \frac{2}{\beta_{4s}\left(\mathbf{b},\mathbf{x}^{\star}\right)}\|\nabla f\left(\mathbf{x}^{\star}\right)|_{\mathcal{T}}\|_2+\left(1+\frac{\gamma_{4s}\left(\mathbf{b},\mathbf{x}^{\star}\right)}{\beta_{4s}\left(\mathbf{b},\mathbf{x}^{\star}\right)}\right)\|\mathbf{x}^{\star}|_{\mathcal{T}^c}\|_2. \tag{A.10}$$

Combining (A.8) and (A.10) we obtain

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

as the lemma stated. ∎

Using the results above, we can now prove Theorem 3.1.

Proof of Theorem 3.1. Using definition 3.1 it is easy to verify that for $k \leq k'$ and any vector $\mathbf{u}$ we have $A_k(\mathbf{u}) \leq A_{k'}(\mathbf{u})$ and $B_k(\mathbf{u}) \geq B_{k'}(\mathbf{u})$. Consequently, for $k \leq k'$ and any pair of vectors $\mathbf{p}$ and $\mathbf{q}$ we have $\alpha_k(\mathbf{p},\mathbf{q}) \leq \alpha_{k'}(\mathbf{p},\mathbf{q})$, $\beta_k(\mathbf{p},\mathbf{q}) \geq \beta_{k'}(\mathbf{p},\mathbf{q})$, and $\mu_k \leq \mu_{k'}$. Furthermore, for any function that satisfies μ_k-SRH we can write

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

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

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

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

$$\left\|\widehat{\mathbf{x}}^{(i)}-\mathbf{x}^\star\right\|_2 \leq \frac{1}{2}\left\|\widehat{\mathbf{x}}^{(i-1)}-\mathbf{x}^\star\right\|_2 + \left(3+\sqrt{3}\right)\epsilon.$$

The theorem follows using this inequality recursively. ∎

A.2 Iteration Analysis For Non-Smooth Cost Functions

In this part we provide analysis of GraSP for non-smooth functions. Definition 3.3 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

$$\beta_k(\mathbf{x})\left\|\boldsymbol{\Delta}\right\|_2^2 \leq \mathrm{B}_f(\mathbf{x}+\boldsymbol{\Delta} \parallel \mathbf{x}) \leq \alpha_k(\mathbf{x})\left\|\boldsymbol{\Delta}\right\|_2^2 \tag{A.11}$$

holds for all vectors $\boldsymbol{\Delta} \in \mathbb{R}^n$ that satisfy $|\mathrm{supp}(\mathbf{x}) \cup \mathrm{supp}(\boldsymbol{\Delta})| \leq k$. By interchanging $\mathbf{x}$ and $\mathbf{x}+\boldsymbol{\Delta}$ in (A.11) and using the fact that

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

one can easily deduce

$$[\beta_k(\mathbf{x}+\boldsymbol{\Delta})+\beta_k(\mathbf{x})]\|\boldsymbol{\Delta}\|_2^2 \leq \langle \nabla_f(\mathbf{x}+\boldsymbol{\Delta}) - \nabla_f(\mathbf{x}), \boldsymbol{\Delta} \rangle \leq [\alpha_k(\mathbf{x}+\boldsymbol{\Delta})+\alpha_k(\mathbf{x})]\|\boldsymbol{\Delta}\|_2^2 \tag{A.12}$$

Propositions A.3, A.4, and A.5 establish some basic inequalities regarding the restricted Bregman divergence under SRL assumption. Using these inequalities we prove Lemmas A.4 and A.5. These two Lemmas are then used to prove an iteration invariant result in Lemma A.6 which in turn is used to prove Theorem 3.2.

Note In Propositions A.3, A.4, and A.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 $\boldsymbol{\Delta} = \mathbf{x}_1 - \mathbf{x}_2$ and denote $\mathrm{supp}(\boldsymbol{\Delta})$ by $\mathcal{R}$. We also denote $\nabla_f(\mathbf{x}_1) - \nabla_f(\mathbf{x}_2)$ by $\boldsymbol{\Delta}'$. 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 A.3. *Let $\mathcal{R}'$ be a subset of $\mathcal{R}$. Then the following inequalities hold.*

$$\begin{aligned} \left|\overline{\alpha}_r\|\boldsymbol{\Delta}|_{\mathcal{R}'}\|_2^2 - \langle \boldsymbol{\Delta}', \boldsymbol{\Delta}|_{\mathcal{R}'} \rangle\right| &\leq \overline{\gamma}_r\|\boldsymbol{\Delta}|_{\mathcal{R}'}\|_2\|\boldsymbol{\Delta}\|_2 \\ \left|\overline{\beta}_r\|\boldsymbol{\Delta}|_{\mathcal{R}'}\|_2^2 - \langle \boldsymbol{\Delta}', \boldsymbol{\Delta}|_{\mathcal{R}'} \rangle\right| &\leq \overline{\gamma}_r\|\boldsymbol{\Delta}|_{\mathcal{R}'}\|_2\|\boldsymbol{\Delta}\|_2 \end{aligned} \tag{A.13}$$

Proof. Using (A.11) we can write

$$\beta_r(\mathbf{x}_1)\,\|\boldsymbol{\Delta}|_{\mathcal{R}'}\|_2^2 t^2 \le \mathrm{B}_f(\mathbf{x}_1 - t\,\boldsymbol{\Delta}|_{\mathcal{R}'} \parallel \mathbf{x}_1) \le \alpha_r(\mathbf{x}_1)\,\|\boldsymbol{\Delta}|_{\mathcal{R}'}\|_2^2 t^2 \tag{A.14}$$

$$\beta_r(\mathbf{x}_2)\,\|\boldsymbol{\Delta}|_{\mathcal{R}'}\|_2^2 t^2 \le \mathrm{B}_f(\mathbf{x}_2 - t\,\boldsymbol{\Delta}|_{\mathcal{R}'} \parallel \mathbf{x}_2) \le \alpha_r(\mathbf{x}_2)\,\|\boldsymbol{\Delta}|_{\mathcal{R}'}\|_2^2 t^2 \tag{A.15}$$

and

$$\beta_r(\mathbf{x}_1)\,\|\boldsymbol{\Delta} - t\,\boldsymbol{\Delta}|_{\mathcal{R}'}\|_2^2 \le \mathrm{B}_f(\mathbf{x}_2 + t\,\boldsymbol{\Delta}|_{\mathcal{R}'} \parallel \mathbf{x}_1) \le \alpha_r(\mathbf{x}_1)\,\|\boldsymbol{\Delta} - t\,\boldsymbol{\Delta}|_{\mathcal{R}'}\|_2^2 \tag{A.16}$$

$$\beta_r(\mathbf{x}_2)\,\|\boldsymbol{\Delta} - t\,\boldsymbol{\Delta}|_{\mathcal{R}'}\|_2^2 \le \mathrm{B}_f(\mathbf{x}_1 - t\,\boldsymbol{\Delta}|_{\mathcal{R}'} \parallel \mathbf{x}_2) \le \alpha_r(\mathbf{x}_2)\,\|\boldsymbol{\Delta} - t\,\boldsymbol{\Delta}|_{\mathcal{R}'}\|_2^2, \tag{A.17}$$

where t is an arbitrary real number. Using the definition of the Bregman divergence we can add (A.14) and (A.15) to obtain

$$\begin{aligned}\overline{\beta}_r\|\boldsymbol{\Delta}|_{\mathcal{R}'}\|_2^2 t^2 &\le f(\mathbf{x}_1 - t\,\boldsymbol{\Delta}|_{\mathcal{R}'}) - f(\mathbf{x}_1) + f(\mathbf{x}_2 + t\,\boldsymbol{\Delta}|_{\mathcal{R}'}) - f(\mathbf{x}_2) + \left\langle \boldsymbol{\Delta}', \boldsymbol{\Delta}|_{\mathcal{R}'}\right\rangle t \\ &\le \overline{\alpha}_r\|\boldsymbol{\Delta}|_{\mathcal{R}'}\|_2^2 t^2.\end{aligned} \tag{A.18}$$

Similarly, (A.16) and (A.17) yield

$$\begin{aligned}\overline{\beta}_r\|\boldsymbol{\Delta} - t\,\boldsymbol{\Delta}|_{\mathcal{R}'}\|_2^2 &\le f(\mathbf{x}_1 - t\,\boldsymbol{\Delta}|_{\mathcal{R}'}) - f(\mathbf{x}_1) + f(\mathbf{x}_2 + t\,\boldsymbol{\Delta}|_{\mathcal{R}'}) - f(\mathbf{x}_2) + \left\langle \boldsymbol{\Delta}', \boldsymbol{\Delta} - t\,\boldsymbol{\Delta}|_{\mathcal{R}'}\right\rangle \\ &\le \overline{\alpha}_r\|\boldsymbol{\Delta} - t\,\boldsymbol{\Delta}|_{\mathcal{R}'}\|_2^2.\end{aligned} \tag{A.19}$$

Expanding the quadratic bounds of (A.19) and using (A.18) then we obtain

$$0 \le \overline{\gamma}_r\|\boldsymbol{\Delta}|_{\mathcal{R}'}\|_2^2 t^2 + 2\left(\overline{\beta}_r\|\boldsymbol{\Delta}|_{\mathcal{R}'}\|_2^2 - \langle \boldsymbol{\Delta}, \boldsymbol{\Delta}|_{\mathcal{R}'}\rangle\right) t - \overline{\beta}_r\|\boldsymbol{\Delta}\|_2^2 + \left\langle \boldsymbol{\Delta}', \boldsymbol{\Delta}\right\rangle \tag{A.20}$$

$$0 \le \overline{\gamma}_r\|\boldsymbol{\Delta}|_{\mathcal{R}'}\|_2^2 t^2 - 2\left(\overline{\alpha}_r\|\boldsymbol{\Delta}|_{\mathcal{R}'}\|_2^2 - \langle \boldsymbol{\Delta}, \boldsymbol{\Delta}|_{\mathcal{R}'}\rangle\right) t + \overline{\alpha}_r\|\boldsymbol{\Delta}\|_2^2 - \left\langle \boldsymbol{\Delta}', \boldsymbol{\Delta}\right\rangle. \tag{A.21}$$

It follows from (A.12), (A.20), and (A.21) that

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

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

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

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

which immediately yields the desired result. ∎

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

$$\left| \| \boldsymbol{\Delta}'|_{\mathcal{R}'}\|_2^2 - \overline{\alpha}_r \langle \boldsymbol{\Delta}', \boldsymbol{\Delta}|_{\mathcal{R}'}\rangle \right| \leq \overline{\gamma}_r \| \boldsymbol{\Delta}|_{\mathcal{R}'}\|_2 \|\boldsymbol{\Delta}\|_2 \tag{A.22}$$
$$\left| \| \boldsymbol{\Delta}'|_{\mathcal{R}'}\|_2^2 - \overline{\beta}_r \langle \boldsymbol{\Delta}', \boldsymbol{\Delta}|_{\mathcal{R}'}\rangle \right| \leq \overline{\gamma}_r \| \boldsymbol{\Delta}|_{\mathcal{R}'}\|_2 \|\boldsymbol{\Delta}\|_2$$

Proof. From (A.11) we have

$$\beta_r(\mathbf{x}_1) \| \boldsymbol{\Delta}'|_{\mathcal{R}'}\|_2^2 t^2 \leq \mathrm{B}_f\left(\mathbf{x}_1 - t\, \boldsymbol{\Delta}'|_{\mathcal{R}'} \parallel \mathbf{x}_1\right) \leq \alpha_r(\mathbf{x}_1) \| \boldsymbol{\Delta}'|_{\mathcal{R}'}\|_2^2 t^2 \tag{A.23}$$
$$\beta_r(\mathbf{x}_2) \| \boldsymbol{\Delta}'|_{\mathcal{R}'}\|_2^2 t^2 \leq \mathrm{B}_f\left(\mathbf{x}_2 + t\, \boldsymbol{\Delta}'|_{\mathcal{R}'} \parallel \mathbf{x}_2\right) \leq \alpha_r(\mathbf{x}_2) \| \boldsymbol{\Delta}'|_{\mathcal{R}'}\|_2^2 t^2 \tag{A.24}$$

and

$$\beta_r(\mathbf{x}_1) \|\boldsymbol{\Delta} - t\, \boldsymbol{\Delta}'|_{\mathcal{R}'}\|_2^2 \leq \mathrm{B}_f\left(\mathbf{x}_2 + t\, \boldsymbol{\Delta}'|_{\mathcal{R}'} \parallel \mathbf{x}_1\right) \leq \alpha_r(\mathbf{x}_1) \|\boldsymbol{\Delta} - t\, \boldsymbol{\Delta}'|_{\mathcal{R}'}\|_2^2 \tag{A.25}$$
$$\beta_r(\mathbf{x}_2) \|\boldsymbol{\Delta} - t\, \boldsymbol{\Delta}'|_{\mathcal{R}'}\|_2^2 \leq \mathrm{B}_f\left(\mathbf{x}_1 - t\, \boldsymbol{\Delta}'|_{\mathcal{R}'} \parallel \mathbf{x}_2\right) \leq \alpha_r(\mathbf{x}_2) \|\boldsymbol{\Delta} - t\, \boldsymbol{\Delta}'|_{\mathcal{R}'}\|_2^2, \tag{A.26}$$

for any $t \in \mathbb{R}$. By subtracting the sum of (A.25) and (A.26) from that of (A.23) and (A.24) we obtain

$$\begin{aligned}\overline{\beta}_r \| \boldsymbol{\Delta}'|_{\mathcal{R}'}\|_2^2 t^2 - \overline{\alpha}_r \|\boldsymbol{\Delta} - t\, \boldsymbol{\Delta}'|_{\mathcal{R}'}\|_2^2 &\leq 2\langle \boldsymbol{\Delta}', \boldsymbol{\Delta}'|_{\mathcal{R}'}\rangle t - \langle \boldsymbol{\Delta}', \boldsymbol{\Delta}\rangle \\ &\leq \overline{\alpha}_r \| \boldsymbol{\Delta}'|_{\mathcal{R}'}\|_2^2 t^2 - \overline{\beta}_r \|\boldsymbol{\Delta} - t\, \boldsymbol{\Delta}'|_{\mathcal{R}'}\|_2^2. \end{aligned} \tag{A.27}$$

Expanding the bounds of (A.27) then yields

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

Note that $\langle \boldsymbol{\Delta}', \boldsymbol{\Delta}'|_{\mathcal{R}'} \rangle = \| \boldsymbol{\Delta}'|_{\mathcal{R}'} \|_2^2$ and $\langle \boldsymbol{\Delta}, \boldsymbol{\Delta}'|_{\mathcal{R}'} \rangle = \langle \boldsymbol{\Delta}|_{\mathcal{R}'}, \boldsymbol{\Delta}' \rangle$. Therefore, using (A.12) we obtain

$$0 \leq \overline{\gamma}_r \| \boldsymbol{\Delta}'|_{\mathcal{R}'} \|_2^2 t^2 + 2 \left(\| \boldsymbol{\Delta}'|_{\mathcal{R}'} \|_2^2 - \overline{\alpha}_r \langle \boldsymbol{\Delta}', \boldsymbol{\Delta}|_{\mathcal{R}'} \rangle \right) t + \overline{\gamma}_r \|\boldsymbol{\Delta}\|_2^2 \tag{A.28}$$
$$0 \leq \overline{\gamma}_r \| \boldsymbol{\Delta}'|_{\mathcal{R}'} \|_2^2 t^2 - 2 \left(\| \boldsymbol{\Delta}'|_{\mathcal{R}'} \|_2^2 - \overline{\beta}_r \langle \boldsymbol{\Delta}', \boldsymbol{\Delta}|_{\mathcal{R}'} \rangle \right) t + \overline{\gamma}_r \|\boldsymbol{\Delta}\|_2^2 . \tag{A.29}$$

Since the right-hand sides of (A.28) and (A.29) are quadratics in t and always non-negative for all values of $t \in \mathbb{R}$, their discriminants cannot be positive. Thus we have

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

which yield the desired result. ∎

Corollary A.1. *The inequality*

$$\| \boldsymbol{\Delta}'|_{\mathcal{R}'} \|_2 \geq \overline{\beta}_r \| \boldsymbol{\Delta}|_{\mathcal{R}'} \|_2 - \overline{\gamma}_r \left\| \boldsymbol{\Delta}|_{\mathcal{R} \setminus \mathcal{R}'} \right\|_2,$$

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

Proof. It follows from (A.22) and (A.13) that

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

Therefore, after straightforward calculations we get

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

which proves the corollary. ∎

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

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

Proof. Using (A.11) for any $t \in \mathbb{R}$ we can write

$$\beta_{k+r}(\mathbf{x}_1) \left\| \boldsymbol{\Delta}'|_{\mathcal{K}} \right\|_2^2 t^2 \leq \mathrm{B}_f\left(\mathbf{x}_1 + t\, \boldsymbol{\Delta}'|_{\mathcal{K}} \parallel \mathbf{x}_1\right) \leq \alpha_{k+r}(\mathbf{x}_1) \left\| \boldsymbol{\Delta}'|_{\mathcal{K}} \right\|_2^2 t^2 \tag{A.30}$$

$$\beta_{k+r}(\mathbf{x}_2) \left\| \boldsymbol{\Delta}'|_{\mathcal{K}} \right\|_2^2 t^2 \leq \mathrm{B}_f\left(\mathbf{x}_2 - t\, \boldsymbol{\Delta}'|_{\mathcal{K}} \parallel \mathbf{x}_2\right) \leq \alpha_{k+r}(\mathbf{x}_2) \left\| \boldsymbol{\Delta}'|_{\mathcal{K}} \right\|_2^2 t^2 \tag{A.31}$$

and similarly

$$\beta_{k+r}(\mathbf{x}_1) \left\| \boldsymbol{\Delta} + t\, \boldsymbol{\Delta}'|_{\mathcal{K}} \right\|_2^2 \leq \mathrm{B}_f\left(\mathbf{x}_2 - t\, \boldsymbol{\Delta}'|_{\mathcal{K}} \parallel \mathbf{x}_1\right) \leq \alpha_{k+r}(\mathbf{x}_1) \left\| \boldsymbol{\Delta} + t\, \boldsymbol{\Delta}'|_{\mathcal{K}} \right\|_2^2 \tag{A.32}$$

$$\beta_{k+r}(\mathbf{x}_2) \left\| \boldsymbol{\Delta} + t\, \boldsymbol{\Delta}'|_{\mathcal{K}} \right\|_2^2 \leq \mathrm{B}_f\left(\mathbf{x}_1 + t\, \boldsymbol{\Delta}'|_{\mathcal{K}} \parallel \mathbf{x}_2\right) \leq \alpha_{k+r}(\mathbf{x}_2) \left\| \boldsymbol{\Delta} + t\, \boldsymbol{\Delta}'|_{\mathcal{K}} \right\|_2^2. \tag{A.33}$$

By subtracting the sum of (A.32) and (A.33) from that of (A.30) and (A.31) we obtain

$$\begin{aligned}\overline{\beta}_{k+r} \left\| \boldsymbol{\Delta}'|_{\mathcal{K}} \right\|_2^2 t^2 - \overline{\alpha}_{k+r} \left\| \boldsymbol{\Delta} + t\, \boldsymbol{\Delta}'|_{\mathcal{K}} \right\|_2^2 &\leq -2t \left\langle \boldsymbol{\Delta}', \boldsymbol{\Delta}'|_{\mathcal{K}} \right\rangle - \left\langle \boldsymbol{\Delta}', \boldsymbol{\Delta} \right\rangle \\ &\leq \overline{\alpha}_{k+r} \left\| \boldsymbol{\Delta}'|_{\mathcal{K}} \right\|_2^2 t^2 - \overline{\beta}_{k+r} \left\| \boldsymbol{\Delta} + t\, \boldsymbol{\Delta}'|_{\mathcal{K}} \right\|_2^2. \end{aligned} \tag{A.34}$$

Note that $\left\langle \boldsymbol{\Delta}', \boldsymbol{\Delta}'|_{\mathcal{K}} \right\rangle = \left\| \boldsymbol{\Delta}'|_{\mathcal{K}} \right\|_2^2$ and $\left\langle \boldsymbol{\Delta}, \boldsymbol{\Delta}'|_{\mathcal{K}} \right\rangle = 0$. Therefore, (A.12) and (A.34) imply

$$0 \leq \overline{\gamma}_{k+r} \left\| \boldsymbol{\Delta}'|_{\mathcal{K}} \right\|_2^2 t^2 \pm 2 \left\| \boldsymbol{\Delta}'|_{\mathcal{K}} \right\|_2^2 t + \overline{\gamma}_{k+r} \|\boldsymbol{\Delta}\|_2^2 \tag{A.35}$$

hold for all $t \in \mathbb{R}$. Hence, as quadratic functions of t, the right-hand side of (A.35) cannot have a positive discriminant. Thus we must have

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

which yields the desired result. ∎

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

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

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

$$\left\|\mathbf{z}|_{\mathcal{R}\setminus\mathcal{Z}}\right\|_2 \leq \left\|\mathbf{z}|_{\mathcal{Z}\setminus\mathcal{R}}\right\|_2. \tag{A.36}$$

Furthermore, using Corollary A.1 we can write

$$\begin{aligned}
\left\|\mathbf{z}|_{\mathcal{R}\setminus\mathcal{Z}}\right\|_2 &= \left\|\nabla_f(\widehat{\mathbf{x}})|_{\mathcal{R}\setminus\mathcal{Z}}\right\|_2 \\
&\geq \left\|(\nabla_f(\widehat{\mathbf{x}}) - \nabla_f(\mathbf{x}^\star))|_{\mathcal{R}\setminus\mathcal{Z}}\right\|_2 - \left\|\nabla_f(\mathbf{x}^\star)|_{\mathcal{R}\setminus\mathcal{Z}}\right\|_2 \\
&\geq \overline{\beta}_{2s}(\widehat{\mathbf{x}},\mathbf{x}^\star)\left\|(\widehat{\mathbf{x}}-\mathbf{x}^\star)|_{\mathcal{R}\setminus\mathcal{Z}}\right\|_2 - \overline{\gamma}_{2s}(\widehat{\mathbf{x}},\mathbf{x}^\star)\,\|(\widehat{\mathbf{x}}-\mathbf{x}^\star)|_{\mathcal{R}\cap\mathcal{Z}}\|_2 - \left\|\nabla_f(\mathbf{x}^\star)|_{\mathcal{R}\setminus\mathcal{Z}}\right\|_2 \\
&\geq \overline{\beta}_{2s}(\widehat{\mathbf{x}},\mathbf{x}^\star)\left\|(\widehat{\mathbf{x}}-\mathbf{x}^\star)|_{\mathcal{R}\setminus\mathcal{Z}}\right\|_2 - \overline{\gamma}_{2s}(\widehat{\mathbf{x}},\mathbf{x}^\star)\,\|\widehat{\mathbf{x}}-\mathbf{x}^\star\|_2 - \left\|\nabla_f(\mathbf{x}^\star)|_{\mathcal{R}\setminus\mathcal{Z}}\right\|_2.
\end{aligned} \tag{A.37}$$

Similarly, using Proposition A.5 we have

$$\begin{aligned}
\left\|\mathbf{z}|_{\mathcal{Z}\setminus\mathcal{R}}\right\|_2 = \left\|\nabla_f(\widehat{\mathbf{x}})|_{\mathcal{Z}\setminus\mathcal{R}}\right\|_2 &\leq \left\|(\nabla_f(\widehat{\mathbf{x}}) - \nabla_f(\mathbf{x}^\star))|_{\mathcal{Z}\setminus\mathcal{R}}\right\|_2 + \left\|\nabla_f(\mathbf{x}^\star)|_{\mathcal{Z}\setminus\mathcal{R}}\right\|_2 \\
&\leq \overline{\gamma}_{4s}(\widehat{\mathbf{x}},\mathbf{x}^\star)\,\|\widehat{\mathbf{x}}-\mathbf{x}^\star\|_2 + \left\|\nabla_f(\mathbf{x}^\star)|_{\mathcal{Z}\setminus\mathcal{R}}\right\|_2.
\end{aligned} \tag{A.38}$$

Combining (A.36), (A.37), and (A.38) then yields

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

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

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

as desired. ∎

Lemma A.5. *The vector* $\mathbf{b}$ *given by*

$$\mathbf{b}=\arg\min_{\mathbf{x}}\ f\left(\mathbf{x}\right)\quad\text{s.t. }\mathbf{x}|_{\mathcal{T}^c}=\mathbf{0} \tag{A.39}$$

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

Proof. Since $\mathbf{b}$ satisfies (A.39) we must have $\nabla f\left(\mathbf{b}\right)|_{\mathcal{T}}=\mathbf{0}$. Then it follows from Corollary A.1 that

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

which proves the lemma. ∎

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

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

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

$$\|\mathbf{x}^\star|_{\mathcal{T}^c}\|_2 \leq \frac{\overline{\gamma}_{2s}(\widehat{\mathbf{x}},\mathbf{x}^\star)+\overline{\gamma}_{4s}(\widehat{\mathbf{x}},\mathbf{x}^\star)}{\overline{\beta}_{2s}(\widehat{\mathbf{x}},\mathbf{x}^\star)}\|\widehat{\mathbf{x}}-\mathbf{x}^\star\|_2+\frac{\left\|\nabla f(\mathbf{x}^\star)|_{\mathcal{R}\setminus\mathcal{Z}}\right\|_2+\left\|\nabla f(\mathbf{x}^\star)|_{\mathcal{Z}\setminus\mathcal{R}}\right\|_2}{\overline{\beta}_{2s}(\widehat{\mathbf{x}},\mathbf{x}^\star)}. \tag{A.40}$$

Furthermore, as showed by (A.9) during the proof of Lemma A.3, we again have

$$\|\mathbf{b}_s-\mathbf{x}^\star\|_2 \leq 2\|\mathbf{x}^\star|_{\mathcal{T}}-\mathbf{b}\|_2+\|\mathbf{x}^\star|_{\mathcal{T}^c}\|_2.$$

Hence, it follows from Lemma A.5 that

$$\|\mathbf{b}_s-\mathbf{x}^\star\|_2 \leq \frac{2\left\|\nabla_f(\mathbf{x}^\star)|_{\mathcal{T}}\right\|_2}{\overline{\beta}_{4s}(\mathbf{x}^\star,\mathbf{b})}+\left(1+\frac{2\overline{\gamma}_{4s}(\mathbf{x}^\star,\mathbf{b})}{\overline{\beta}_{4s}(\mathbf{x}^\star,\mathbf{b})}\right)\|\mathbf{x}^\star|_{\mathcal{T}^c}\|_2. \tag{A.41}$$

Combining (A.40) and (A.41) yields

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

as desired. ∎

Proof of Theorem 3.2. 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

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

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 A.6 that

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

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

$$\left\|\widehat{\mathbf{x}}-\mathbf{x}^{\star}\right\|_2 \leq 2^{-i}\left\|\mathbf{x}^{\star}\right\|_2+\left(6+2\sqrt{3}\right)\epsilon$$

which is the desired result. ∎

Appendix B

Proofs of Chapter 4

To prove Theorem 4.1 we use the following two lemmas. We omit the proofs since they can be easily adapted from Appendix A Lemmas A.1 and A.2 using straightforward changes. It suffices to notice that

1. the proof in Appendix A still holds if the estimation errors are measured with respect to the true sparse minimizer or any other feasible (i.e., s-sparse) point, rather than the statistical true parameter, and

2. the iterates and the crude estimates will always remain in the sphere of radius r centered at the origin where the SRH applies.

In what follows $\int_0^1 \alpha_k\left(\tau\mathbf{x}+(1-\tau)\,\overline{\mathbf{x}}\right)\mathrm{d}\tau$ and $\int_0^1 \beta_k\left(\tau\mathbf{x}+(1-\tau)\,\overline{\mathbf{x}}\right)\mathrm{d}\tau$ are denoted by $\widetilde{\alpha}_k\left(\mathbf{x}\right)$ and $\widetilde{\beta}_k\left(\mathbf{x}\right)$, respectively. We also define $\widetilde{\gamma}_k\left(\mathbf{x}\right) := \widetilde{\alpha}_k\left(\mathbf{x}\right) - \widetilde{\beta}_k\left(\mathbf{x}\right)$.

Lemma B.1. *Let $\mathcal{Z}$ be the index set defined in Algorithm 2 and $\mathcal{R}$ denote the set* $\operatorname{supp}\left(\mathbf{x}^{(t)}-\overline{\mathbf{x}}\right)$.

Then the iterate $\mathbf{x}^{(t)}$ *obeys*

$$\left\| \left(\mathbf{x}^{(t)} - \overline{\mathbf{x}} \right) \Big|_{\mathcal{Z}^{\mathrm{c}}} \right\|_2 \leq \frac{\widetilde{\gamma}_{4s}\left(\mathbf{x}^{(t)}\right) + \widetilde{\gamma}_{2s}\left(\mathbf{x}^{(t)}\right)}{\widetilde{\beta}_{2s}\left(\mathbf{x}^{(t)}\right)} \left\| \mathbf{x}^{(t)} - \overline{\mathbf{x}} \right\|_2 + \frac{\left\| \nabla_{\mathcal{R}\setminus\mathcal{Z}} f\left(\overline{\mathbf{x}}\right) \right\|_2 + \left\| \nabla_{\mathcal{Z}\setminus\mathcal{R}} f\left(\overline{\mathbf{x}}\right) \right\|_2}{\widetilde{\beta}_{2s}\left(\mathbf{x}^{(t)}\right)}.$$

Lemma B.2. *The vector* $\mathbf{b}$ *defined at line 3 of Algorithm 2 obeys*

$$\left\| \overline{\mathbf{x}}|_{\mathcal{T}} - \mathbf{b} \right\|_2 \leq \frac{\left\| \nabla_{\mathcal{T}} f\left(\overline{\mathbf{x}}\right) \right\|_2}{\widetilde{\beta}_{4s}\left(\mathbf{b}\right)} + \frac{\widetilde{\gamma}_{4s}\left(\mathbf{b}\right)}{2\widetilde{\beta}_{4s}\left(\mathbf{b}\right)} \left\| \overline{\mathbf{x}}|_{\mathcal{T}^{\mathrm{c}}} \right\|_2.$$

Proof of Theorem 4.1. Since $\mathcal{Z} \subseteq \mathcal{T}$ we have $\mathcal{T}^{\mathrm{c}} \subseteq \mathcal{Z}^{\mathrm{c}}$ and thus

$$\begin{aligned} \left\| \left(\mathbf{x}^{(t)} - \overline{\mathbf{x}} \right) \Big|_{\mathcal{Z}^{\mathrm{c}}} \right\|_2 &\geq \left\| \left(\mathbf{x}^{(t)} - \overline{\mathbf{x}} \right) \Big|_{\mathcal{T}^{\mathrm{c}}} \right\|_2 \\ &= \left\| \overline{\mathbf{x}}|_{\mathcal{T}^{\mathrm{c}}} \right\|_2. \end{aligned}$$

Then it follows from Lemma B.1 that

$$\begin{aligned} \left\| \overline{\mathbf{x}}|_{\mathcal{T}^{\mathrm{c}}} \right\|_2 &\leq \frac{\widetilde{\gamma}_{4s}\left(\mathbf{x}^{(t)}\right)}{\widetilde{\beta}_{4s}\left(\mathbf{x}^{(t)}\right)} \left\| \mathbf{x}^{(t)} - \overline{\mathbf{x}} \right\|_2 \\ &\quad + \frac{\left\| \nabla_{\mathcal{R}\setminus\mathcal{Z}} f\left(\overline{\mathbf{x}}\right) \right\|_2 + \left\| \nabla_{\mathcal{Z}\setminus\mathcal{R}} f\left(\overline{\mathbf{x}}\right) \right\|_2}{\beta_{4s}} \\ &\leq \left(\mu_{4s} - 1\right) \left\| \mathbf{x}^{(t)} - \overline{\mathbf{x}} \right\|_2 + 2\epsilon, \end{aligned} \tag{B.1}$$

where we used the fact that $\alpha_{4s} \geq \alpha_{2s}$ and $\beta_{4s} \leq \beta_{2s}$ to simplify the expressions. Furthermore, we have

$$\begin{aligned} \left\| \mathbf{x}^{(t+1)} - \overline{\mathbf{x}} \right\|_2 &= \left\| \mathbf{b}_s - \overline{\mathbf{x}} \right\|_2 \\ &\leq \left\| \mathbf{b}_s - \overline{\mathbf{x}}|_{\mathcal{T}} \right\|_2 + \left\| \overline{\mathbf{x}}|_{\mathcal{T}^{\mathrm{c}}} \right\|_2 \\ &\leq \left\| \mathbf{b}_s - \mathbf{b} \right\|_2 + \left\| \mathbf{b} - \overline{\mathbf{x}}|_{\mathcal{T}} \right\|_2 + \left\| \overline{\mathbf{x}}|_{\mathcal{T}^{\mathrm{c}}} \right\|_2 \\ &\leq 2\left\| \mathbf{b} - \overline{\mathbf{x}}|_{\mathcal{T}} \right\|_2 + \left\| \overline{\mathbf{x}}|_{\mathcal{T}^{\mathrm{c}}} \right\|_2, \end{aligned}$$

where the last inequality holds because $\mathbf{b}_s$ is the best s-term approximation of $\mathbf{b}$. Hence, it follows from Lemma B.2 that

$$\begin{aligned}\left\|\mathbf{x}^{(t+1)}-\overline{\mathbf{x}}\right\|_2 &\leq 2\frac{\|\nabla_{\mathcal{T}} f\left(\overline{\mathbf{x}}\right)\|_2}{\widetilde{\beta}_{4s}\left(\mathbf{b}\right)}+\frac{\widetilde{\alpha}_{4s}\left(\mathbf{b}\right)}{\widetilde{\beta}_{4s}\left(\mathbf{b}\right)}\|\overline{\mathbf{x}}|_{\mathcal{T}^c}\|_2\\ &\leq 2\epsilon+\mu_{4s}\|\overline{\mathbf{x}}|_{\mathcal{T}^c}\|_2.\end{aligned}$$

Then applying (B.1) and simplifying the resulting inequality yield

$$\begin{aligned}\left\|\mathbf{x}^{(t+1)}-\overline{\mathbf{x}}\right\|_2 &\leq 2\epsilon+\mu_{4s}\left(\left(\mu_{4s}-1\right)\left\|\mathbf{x}^{(t)}-\overline{\mathbf{x}}\right\|_2+2\epsilon\right)\\ &\leq \left(\mu_{4s}^2-\mu_{4s}\right)\left\|\mathbf{x}^{(t)}-\overline{\mathbf{x}}\right\|_2+2\left(\mu_{4s}+1\right)\epsilon,\end{aligned}$$

which is the desired result. ∎

Lemma B.3 (Bounded Sparse Projection)**.** *For any* $\mathbf{x}\in\mathbb{R}^n$ *the vector* $\max\left\{1,\frac{r}{\|\mathbf{x}_s\|_2}\right\}\mathbf{x}_s$ *is a solution to the minimization*

$$\arg\min_{\mathbf{w}}\ \|\mathbf{x}-\mathbf{w}\|_2 \quad \textit{s.t. } \|\mathbf{w}\|_2\leq r \textit{ and } \|\mathbf{w}\|_0\leq s. \tag{B.2}$$

Proof. Given an index set $\mathcal{S}\subseteq[n]$ we can write $\|\mathbf{x}-\mathbf{w}\|_2{}^2=\|\mathbf{x}-\mathbf{w}|_{\mathcal{S}}\|_2{}^2+\|\mathbf{x}|_{\mathcal{S}^c}\|_2{}^2$ for vectors $\mathbf{w}$ with $\operatorname{supp}(\mathbf{w})\subseteq\mathcal{S}$. Therefore, the solution to

$$\arg\min_{\mathbf{w}}\ \|\mathbf{x}-\mathbf{w}\|_2 \quad \text{s.t. } \|\mathbf{w}\|_2\leq r \text{ and } \operatorname{supp}(\mathbf{w})\subseteq\mathcal{S}$$

is simply obtained by projection of $\mathbf{x}|_{\mathcal{S}}$ onto the sphere of radius r, i.e.,

$$\mathrm{P}_{\mathcal{S}}(\mathbf{x})=\max\left\{1,\frac{r}{\|\mathbf{x}|_{\mathcal{S}}\|_2}\right\}\mathbf{x}|_{\mathcal{S}}.$$

Therefore, to find a solution to (B.2) it suffices to find the index set $\mathcal{S}$ with $|\mathcal{S}|=s$ and thus the corresponding $\mathrm{P}_{\mathcal{S}}(\mathbf{x})$ that minimize $\|\mathbf{x}-\mathrm{P}_{\mathcal{S}}(\mathbf{x})\|_2$. Note that we have

$$\|\mathbf{x}-\mathrm{P}_{\mathcal{S}}(\mathbf{x})\|_2{}^2=\|\mathbf{x}|_{\mathcal{S}}-\mathrm{P}_{\mathcal{S}}(\mathbf{x})\|_2{}^2+\|\mathbf{x}|_{\mathcal{S}^c}\|_2{}^2$$

$$
\begin{aligned}
&= \left(\|\mathbf{x}|_{\mathcal{S}}\|_2 - r\right)_+^2 + \|\mathbf{x}|_{\mathcal{S}^c}\|_2^2 \\
&= \begin{cases} \|\mathbf{x}\|_2^2 - \|\mathbf{x}|_{\mathcal{S}}\|_2^2 & , \|\mathbf{x}|_{\mathcal{S}}\|_2 < r \\ \|\mathbf{x}\|_2^2 + r^2 - 2r\|\mathbf{x}|_{\mathcal{S}}\|_2 & , \|\mathbf{x}|_{\mathcal{S}}\|_2 \geq r \end{cases}.
\end{aligned}
$$

For all valid $\mathcal{S}$ with $\|\mathbf{x}|_{\mathcal{S}}\|_2 < r$ we have $\|\mathbf{x}\|_2^2 - \|\mathbf{x}|_{\mathcal{S}}\|_2^2 > \|\mathbf{x}\|_2^2 - r^2$. Similarly, for all valid $\mathcal{S}$ with $\|\mathbf{x}|_{\mathcal{S}}\|_2 < r$ we have $\|\mathbf{x}\|_2^2 + r^2 - 2r\|\mathbf{x}|_{\mathcal{S}}\|_2 \leq \|\mathbf{x}\|_2^2 - r^2$. Furthermore, both $\|\mathbf{x}\|_2^2 - \|\mathbf{x}|_{\mathcal{S}}\|_2^2$ and $\|\mathbf{x}\|_2^2 + r^2 - 2r\|\mathbf{x}|_{\mathcal{S}}\|_2$ are decreasing functions of $\|\mathbf{x}|_{\mathcal{S}}\|_2$. Therefore, $\|\mathbf{x} - \mathrm{P}_{\mathcal{S}}(\mathbf{x})\|_2^2$ is a decreasing function of $\|\mathbf{x}|_{\mathcal{S}}\|_2$. Hence, $\|\mathbf{x} - \mathrm{P}_{\mathcal{S}}(\mathbf{x})\|_2$ attains its minimum at $\mathcal{S} = \mathrm{supp}(\mathbf{x}_s)$. ∎

On non-convex formulation of Plan and Vershynin (2013)

Plan and Vershynin (2013) derived accuracy guarantees for

$$\arg\max_{\mathbf{x}} \ \langle \mathbf{y}, \mathbf{A}\mathbf{x} \rangle \quad \text{s.t. } \mathbf{x} \in \mathcal{K}$$

as a solver for the 1-bit CS problem, where $\mathcal{K}$ is a subset of the unit Euclidean ball. While their result (Plan and Vershynin, 2013, Theorem 1.1) applies to both convex and non-convex sets $\mathcal{K}$, the focus of their work has been on the set $\mathcal{K}$ that is the intersection of a centered ℓ_1-ball and the unit Euclidean ball. Our goal, however, is to examine the other interesting choice of $\mathcal{K}$, namely the intersection of canonical sparse subspaces and the unit Euclidean ball. The estimator in this case can be written as

$$\arg\max_{\mathbf{x}} \ \langle \mathbf{y}, \mathbf{A}\mathbf{x} \rangle \quad \text{s.t. } \|\mathbf{x}\|_0 \leq s \text{ and } \|\mathbf{x}\|_2 \leq 1. \tag{B.3}$$

We show that a solution to the optimization above can be obtained explicitly.

Lemma B.4. *A solution to (B.3) is* $\widehat{\mathbf{x}} = \left(\mathbf{A}^{\mathrm{T}}\mathbf{y}\right)_s / \left\|\left(\mathbf{A}^{\mathrm{T}}\mathbf{y}\right)_s\right\|_2$.

Proof. For $\mathcal{I} \subseteq [n]$ define

$$\widehat{\mathbf{x}}(\mathcal{I}) := \arg\max_{\mathbf{x}} \ \langle \mathbf{y}, \mathbf{A}\mathbf{x} \rangle \quad \text{s.t. } \mathbf{x}|_{\mathcal{I}^c} = 0 \text{ and } \|\mathbf{x}\|_2 \leq 1.$$

Furthermore, choose

$$\widehat{\mathcal{I}} \in \arg\max_{\mathcal{I}} \ \langle \mathbf{y}, \mathbf{A}\widehat{\mathbf{x}}(\mathcal{I}) \rangle \quad \text{s.t. } \mathcal{I} \subseteq [n] \text{ and } |\mathcal{I}| \leq s.$$

Then $\widehat{\mathbf{x}}\left(\widehat{\mathcal{I}}\right)$ would be a solution to (B.3). Using the fact that $\langle \mathbf{y}, \mathbf{A}\mathbf{x} \rangle = \left\langle \mathbf{A}^{\mathrm{T}}\mathbf{y}, \mathbf{x} \right\rangle$, straightforward application of the Cauchy-Schwarz inequality shows that $\widehat{\mathbf{x}}(\mathcal{I}) = \left(\mathbf{A}^{\mathrm{T}}\mathbf{y}\right)\big|_{\mathcal{I}} / \left\| \left(\mathbf{A}^{\mathrm{T}}\mathbf{y}\right)\big|_{\mathcal{I}} \right\|_2$ for which we have

$$\langle \mathbf{y}, \mathbf{A}\widehat{\mathbf{x}}(\mathcal{I}) \rangle = \left\| \left(\mathbf{A}^{\mathrm{T}}\mathbf{y}\right)\big|_{\mathcal{I}} \right\|_2.$$

Thus, we obtain $\widehat{\mathcal{I}} = \operatorname{supp}\left(\left(\mathbf{A}^{\mathrm{T}}\mathbf{y}\right)_s\right)$ and thereby $\widehat{\mathbf{x}}\left(\widehat{\mathcal{I}}\right) = \widehat{\mathbf{x}}$, which proves the claim. ∎

Appendix C

Proofs of Chapter 5

Lemma C.1. *Suppose that f is a twice differentiable function that satisfies (5.4) for a given $\mathbf{x}$ and all $\boldsymbol{\Delta}$ such that $\operatorname{supp}(\boldsymbol{\Delta}) \cup \operatorname{supp}(\mathbf{x}) \in \mathcal{M}(\mathcal{C}_k)$. Then we have*

$$\left|\langle \mathbf{u}, \mathbf{v} \rangle - \eta \left\langle \mathbf{u}, \nabla^2 f(\mathbf{x})\, \mathbf{v} \right\rangle\right| \leq \left(\eta \frac{\alpha_{\mathcal{C}_k} - \beta_{\mathcal{C}_k}}{2} + \left| \eta \frac{\alpha_{\mathcal{C}_k} + \beta_{\mathcal{C}_k}}{2} - 1 \right| \right) \|\mathbf{u}\| \|\mathbf{v}\|,$$

for all $\eta > 0$ and $\mathbf{u}, \mathbf{v} \in \mathcal{H}$ such that $\operatorname{supp}(\mathbf{u} \pm \mathbf{v}) \cup \operatorname{supp}(\mathbf{x}) \in \mathcal{M}(\mathcal{C}_k)$.

Proof. We first the prove the lemma for unit-norm vectors $\mathbf{u}$ and $\mathbf{v}$. Since $\operatorname{supp}(\mathbf{u} \pm \mathbf{v}) \cup \operatorname{supp}(\mathbf{x}) \in \mathcal{M}(\mathcal{C}_k)$ we can use (5.4) for $\boldsymbol{\Delta} = \mathbf{u} \pm \mathbf{v}$ to obtain

$$\beta_{\mathcal{C}_k} \|\mathbf{u} \pm \mathbf{v}\|^2 \leq \left\langle \mathbf{u} \pm \mathbf{v}, \nabla^2 f(\mathbf{x})\, (\mathbf{u} \pm \mathbf{v}) \right\rangle \leq \alpha_{\mathcal{C}_k} \|\mathbf{u} \pm \mathbf{v}\|^2.$$

These inequalities and the assumption $\|\mathbf{u}\| = \|\mathbf{v}\| = 1$ then yield

$$\frac{\beta_{\mathcal{C}_k} - \alpha_{\mathcal{C}_k}}{2} + \frac{\alpha_{\mathcal{C}_k} + \beta_{\mathcal{C}_k}}{2} \langle \mathbf{u}, \mathbf{v} \rangle \leq \left\langle \mathbf{u}, \nabla^2 f(\mathbf{x})\, \mathbf{v} \right\rangle \leq \frac{\alpha_{\mathcal{C}_k} - \beta_{\mathcal{C}_k}}{2} + \frac{\alpha_{\mathcal{C}_k} + \beta_{\mathcal{C}_k}}{2} \langle \mathbf{u}, \mathbf{v} \rangle,$$

where we used the fact that $\nabla^2 f(\mathbf{x})$ is symmetric since f is twice continuously differen-

tiable. Multiplying all sides by η and rearranging the terms then imply

$$\begin{aligned}\eta\frac{\alpha_{\mathcal{C}_k}-\beta_{\mathcal{C}_k}}{2} &\geq \left|\left(\eta\frac{\alpha_{\mathcal{C}_k}+\beta_{\mathcal{C}_k}}{2}-1\right)\langle\mathbf{u},\mathbf{v}\rangle+\langle\mathbf{u},\mathbf{v}\rangle-\eta\left\langle\mathbf{u},\nabla^2 f(\mathbf{x})\,\mathbf{v}\right\rangle\right| \\ &\geq \left|\langle\mathbf{u},\mathbf{v}\rangle-\eta\left\langle\mathbf{u},\nabla^2 f(\mathbf{x})\,\mathbf{v}\right\rangle\right|-\left|\left(\eta\frac{\alpha_{\mathcal{C}_k}+\beta_{\mathcal{C}_k}}{2}-1\right)\langle\mathbf{u},\mathbf{v}\rangle\right| \\ &\geq \left|\langle\mathbf{u},\mathbf{v}\rangle-\eta\left\langle\mathbf{u},\nabla^2 f(\mathbf{x})\,\mathbf{v}\right\rangle\right|-\left|\eta\frac{\alpha_{\mathcal{C}_k}+\beta_{\mathcal{C}_k}}{2}-1\right|, \end{aligned} \tag{C.1}$$

which is equivalent to result for unit-norm $\mathbf{u}$ and $\mathbf{v}$ as desired. For the general case one can write $\mathbf{u}=\|\mathbf{u}\|\mathbf{u}'$ and $\mathbf{v}=\|\mathbf{v}\|\mathbf{v}'$ such that $\mathbf{u}'$ and $\mathbf{v}'$ are both unit-norm. It is straightforward to verify that using (C.1) for $\mathbf{u}'$ and $\mathbf{v}'$ as the unit-norm vectors and multiplying both sides of the resulting inequality by $\|\mathbf{u}\|\|\mathbf{v}\|$ yields the desired general case. ∎

***Proof of Theorem** 5.1.* Using optimality of $\mathbf{x}^{(t+1)}$ and feasibility of $\overline{\mathbf{x}}$ one can deduce

$$\left\|\mathbf{x}^{(t+1)}-\mathbf{z}^{(t)}\right\|^2 \leq \left\|\overline{\mathbf{x}}-\mathbf{z}^{(t)}\right\|^2,$$

with $\mathbf{z}^{(t)}$ as in line 2 of Algorithm 3. Expanding the squared norms using the inner product of $\mathcal{H}$ then shows $0\leq\left\langle\mathbf{x}^{(t+1)}-\overline{\mathbf{x}},2\mathbf{z}^{(t)}-\mathbf{x}^{(t+1)}-\overline{\mathbf{x}}\right\rangle$ or equivalently

$$0\leq\left\langle\boldsymbol{\Delta}^{(t+1)},2\mathbf{x}^{(t)}-2\eta^{(t)}\nabla f\left(\overline{\mathbf{x}}+\boldsymbol{\Delta}^{(t)}\right)-\boldsymbol{\Delta}^{(t+1)}\right\rangle,$$

where $\boldsymbol{\Delta}^{(t)}=\mathbf{x}^{(t)}-\overline{\mathbf{x}}$ and $\boldsymbol{\Delta}^{(t+1)}=\mathbf{x}^{(t+1)}-\overline{\mathbf{x}}$. Adding and subtracting $2\eta^{(t)}\left\langle\boldsymbol{\Delta}^{(t+1)},\nabla f(\overline{\mathbf{x}})\right\rangle$ and rearranging yields

$$\begin{aligned}\left\|\boldsymbol{\Delta}^{(t+1)}\right\|^2 \leq\ & 2\left\langle\boldsymbol{\Delta}^{(t+1)},\mathbf{x}^{(t)}\right\rangle-2\eta^{(t)}\left\langle\boldsymbol{\Delta}^{(t+1)},\nabla f\left(\overline{\mathbf{x}}+\boldsymbol{\Delta}^{(t)}\right)-\nabla f(\overline{\mathbf{x}})\right\rangle \\ & -2\eta^{(t)}\left\langle\boldsymbol{\Delta}^{(t+1)},\nabla f(\overline{\mathbf{x}})\right\rangle \end{aligned} \tag{C.2}$$

Since f is twice continuously differentiable by assumption, it follows form the mean-value theorem that $\left\langle\boldsymbol{\Delta}^{(t+1)},\nabla f\left(\overline{\mathbf{x}}+\boldsymbol{\Delta}^{(t)}\right)-\nabla f(\overline{\mathbf{x}})\right\rangle = \left\langle\boldsymbol{\Delta}^{(t+1)},\nabla^2 f\left(\overline{\mathbf{x}}+\tau\boldsymbol{\Delta}^{(t)}\right)\boldsymbol{\Delta}^{(t)}\right\rangle$, for some $\tau\in(0,1)$. Furthermore, because $\overline{\mathbf{x}}$, $\mathbf{x}^{(t)}$, $\mathbf{x}^{(t+1)}$ all belong to the model set $\mathcal{M}(\mathcal{C}_k)$ we

have $\operatorname{supp}\left(\overline{\mathbf{x}}+\tau\boldsymbol{\Delta}^{(t)}\right)\in\mathcal{M}\left(\mathcal{C}_k^2\right)$ and thereby $\operatorname{supp}\left(\boldsymbol{\Delta}^{(t+1)}\right)\cup\operatorname{supp}\left(\overline{\mathbf{x}}+\tau\boldsymbol{\Delta}^{(t)}\right)\in\mathcal{M}\left(\mathcal{C}_k^3\right)$. Invoking the $\left(\mu_{\mathcal{C}_k^3},r\right)$-SMRH condition of the cost function and applying Lemma C.1 with the sparsity model $\mathcal{M}\left(\mathcal{C}_k^3\right)$, $\mathbf{x}=\overline{\mathbf{x}}+\tau\boldsymbol{\Delta}^{(t)}$, and $\eta=\eta^{(t)}$ then yields

$$\left|\left\langle\boldsymbol{\Delta}^{(t+1)},\boldsymbol{\Delta}^{(t)}\right\rangle-\eta^{(t)}\left\langle\boldsymbol{\Delta}^{(t+1)},\nabla f\left(\overline{\mathbf{x}}+\boldsymbol{\Delta}^{(t)}\right)-\nabla f\left(\overline{\mathbf{x}}\right)\right\rangle\right|\leq\gamma^{(t)}\left\|\boldsymbol{\Delta}^{(t+1)}\right\|\left\|\boldsymbol{\Delta}^{(t)}\right\|.$$

Using the Cauchy-Schwarz inequality and the fact that $\left\|\nabla f\left(\overline{\mathbf{x}}\right)|_{\operatorname{supp}\left(\boldsymbol{\Delta}^{(t+1)}\right)}\right\|\leq\left\|\nabla f\left(\overline{\mathbf{x}}\right)|_{\overline{\mathcal{I}}}\right\|$ by the definition of $\overline{\mathcal{I}}$, (C.2) implies that

$$\left\|\boldsymbol{\Delta}^{(t+1)}\right\|^2\leq2\gamma^{(t)}\left\|\boldsymbol{\Delta}^{(t+1)}\right\|\left\|\boldsymbol{\Delta}^{(t)}\right\|+2\eta^{(t)}\left\|\boldsymbol{\Delta}^{(t+1)}\right\|\left\|\nabla f\left(\overline{\mathbf{x}}\right)|_{\overline{\mathcal{I}}}\right\|.$$

Canceling $\left\|\boldsymbol{\Delta}^{(t+1)}\right\|$ from both sides proves the theorem. ∎

Lemma C.2 (Bounded Model Projection)**.** *Given an arbitrary $\mathbf{h}_0\in\mathcal{H}$, a positive real number r, and a sparsity model generator $\mathcal{C}_k$, a projection $\mathrm{P}_{\mathcal{C}_k,r}\left(\mathbf{h}_0\right)$ can be obtained as the projection of $\mathrm{P}_{\mathcal{C}_k,+\infty}\left(\mathbf{h}_0\right)$ on to the sphere of radius r.*

Proof. To simplify the notation let $\widehat{\mathbf{h}}=\mathrm{P}_{\mathcal{C}_k,r}\left(\mathbf{h}_0\right)$ and $\widehat{\mathcal{S}}=\operatorname{supp}\left(\widehat{\mathbf{h}}\right)$. For $\mathcal{S}\subseteq[p]$ define

$$\mathbf{h}_0\left(\mathcal{S}\right)=\arg\min_{\mathbf{h}}\ \left\|\mathbf{h}-\mathbf{h}_0\right\|\quad\text{s.t. }\left\|\mathbf{h}\right\|\leq r\text{ and }\operatorname{supp}\left(\mathbf{h}\right)\subseteq\mathcal{S}.$$

It follows from the definition of $\mathrm{P}_{\mathcal{C}_k,r}\left(\mathbf{h}_0\right)$ that $\widehat{\mathcal{S}}\in\arg\min_{\mathcal{S}\in\mathcal{C}_k}\ \left\|\mathbf{h}_0\left(\mathcal{S}\right)-\mathbf{h}_0\right\|$. Using

$$\left\|\mathbf{h}_0\left(\mathcal{S}\right)-\mathbf{h}_0\right\|^2=\left\|\mathbf{h}_0\left(\mathcal{S}\right)-\mathbf{h}_0|_{\mathcal{S}}-\mathbf{h}_0|_{\mathcal{S}^{\mathrm{c}}}\right\|^2=\left\|\mathbf{h}_0\left(\mathcal{S}\right)-\mathbf{h}_0|_{\mathcal{S}}\right\|^2+\left\|\mathbf{h}_0|_{\mathcal{S}^{\mathrm{c}}}\right\|^2,$$

we deduce that $\mathbf{h}_0\left(\mathcal{S}\right)$ is the projection of $\mathbf{h}_0|_{\mathcal{S}}$ onto the sphere of radius r. Therefore, we can write $\mathbf{h}_0\left(\mathcal{S}\right)=\min\left\{1,r/\left\|\mathbf{h}_0|_{\mathcal{S}}\right\|\right\}\mathbf{h}_0|_{\mathcal{S}}$ and from that

$$\begin{aligned}\widehat{\mathcal{S}}&\in\arg\min_{\mathcal{S}\in\mathcal{C}_k}\ \left\|\min\left\{1,r/\left\|\mathbf{h}_0|_{\mathcal{S}}\right\|\right\}\mathbf{h}_0|_{\mathcal{S}}-\mathbf{h}_0\right\|^2\\&=\arg\min_{\mathcal{S}\in\mathcal{C}_k}\ \left\|\min\left\{0,r/\left\|\mathbf{h}_0|_{\mathcal{S}}\right\|-1\right\}\mathbf{h}_0|_{\mathcal{S}}\right\|^2+\left\|\mathbf{h}_0|_{\mathcal{S}^{\mathrm{c}}}\right\|^2\end{aligned}$$

$$= \arg\min_{\mathcal{S}\in\mathcal{C}_k} \left((1 - r/\|\mathbf{h}_0|_{\mathcal{S}}\|)_+^2 - 1\right) \|\mathbf{h}_0|_{\mathcal{S}}\|^2$$
$$= \arg\max_{\mathcal{S}\in\mathcal{C}_k} q(\mathcal{S}) := \|\mathbf{h}_0|_{\mathcal{S}}\|^2 - (\|\mathbf{h}_0|_{\mathcal{S}}\| - r)_+^2 .$$

Furthermore, let

$$\mathcal{S}_0 = \mathrm{supp}\left(\mathrm{P}_{\mathcal{C}_k,+\infty}(\mathbf{h}_0)\right) = \arg\max_{\mathcal{S}\in\mathcal{C}_k} \|\mathbf{h}_0|_{\mathcal{S}}\|. \tag{C.3}$$

If $\left\|\mathbf{h}_0|_{\mathcal{S}_0}\right\| \le r$ then $q(\mathcal{S}) = \|\mathbf{h}_0|_{\mathcal{S}}\| \le q(\mathcal{S}_0)$ for any $\mathcal{S} \in \mathcal{C}_k$ and thereby $\widehat{\mathcal{S}} = \mathcal{S}_0$. Thus, we focus on cases that $\left\|\mathbf{h}_0|_{\mathcal{S}_0}\right\| > r$ which implies $q(\mathcal{S}_0) = 2\left\|\mathbf{h}_0|_{\mathcal{S}_0}\right\|r - r^2$. For any $\mathcal{S} \in \mathcal{C}_k$ if $\|\mathbf{h}_0|_{\mathcal{S}}\| \le r$ we have $q(\mathcal{S}) = \|\mathbf{h}_0|_{\mathcal{S}}\|^2 \le r^2 < 2\left\|\mathbf{h}_0|_{\mathcal{S}_0}\right\|r - r^2 = q(\mathcal{S}_0)$, and if $\|\mathbf{h}_0|_{\mathcal{S}}\| > r$ we have $q(\mathcal{S}) = 2\|\mathbf{h}_0|_{\mathcal{S}}\|r - r^2 \le 2\left\|\mathbf{h}_0|_{\mathcal{S}_0}\right\|r - r^2 = q(\mathcal{S}_0)$ where (C.3) is applied. Therefore, we have shown that $\widehat{\mathcal{S}} = \mathcal{S}_0$. It is then straightforward to show the desired result that projecting $\mathrm{P}_{\mathcal{C}_k,+\infty}(\mathbf{h}_0)$ onto the centered sphere of radius r yields $\mathrm{P}_{\mathcal{C}_k,r}(\mathbf{h}_0)$. ∎

Appendix D

Proofs of Chapter 6

D.1 Proof of Theorem 6.1

To prove Theorem 6.1 first a series of lemmas should be established. In what follows, $\mathbf{x}_{\perp}^{\star}$ is a projection of the s-sparse vector $\mathbf{x}^{\star}$ onto $\widehat{\mathcal{B}}$ and $\mathbf{x}^{\star} - \mathbf{x}_{\perp}^{\star}$ is denoted by $\mathbf{d}^{\star}$. Furthermore, for $t = 0, 1, 2, \ldots$ we denote $\mathbf{x}^{(t)} - \mathbf{x}_{\perp}^{\star}$ by $\mathbf{d}^{(t)}$ for compactness.

Lemma D.1. *If $\mathbf{x}^{(t)}$ denotes the estimate in the t-th iteration of ℓ_p-PGD, then*

$$\left\|\mathbf{d}^{(t+1)}\right\|_2^2 \leq 2\Re\left[\left\langle \mathbf{d}^{(t)}, \mathbf{d}^{(t+1)}\right\rangle - \eta^{(t)}\left\langle \mathbf{A}\mathbf{d}^{(t)}, \mathbf{A}\mathbf{d}^{(t+1)}\right\rangle\right] + 2\eta^{(t)}\Re\left\langle \mathbf{A}\mathbf{d}^{(t+1)}, \mathbf{A}\mathbf{d}^{\star} + \mathbf{e}\right\rangle.$$

Proof. Note that $\mathbf{x}^{(t+1)}$ is a projection of $\mathbf{x}^{(t)} - \eta^{(t)}\mathbf{A}^{\mathsf{H}}\left(\mathbf{A}\mathbf{x}^{(t)} - \mathbf{y}\right)$ onto $\widehat{\mathcal{B}}$. Since $\mathbf{x}_{\perp}^{\star}$ is also a feasible point (i.e., $\mathbf{x}_{\perp}^{\star} \in \widehat{\mathcal{B}}$) we have

$$\left\|\mathbf{x}^{(t+1)} - \mathbf{x}^{(t)} + \eta^{(t)}\mathbf{A}^{\mathsf{H}}\left(\mathbf{A}\mathbf{x}^{(t)} - \mathbf{y}\right)\right\|_2^2 \leq \left\|\mathbf{x}_{\perp}^{\star} - \mathbf{x}^{(t)} + \eta^{(t)}\mathbf{A}^{\mathsf{H}}\left(\mathbf{A}\mathbf{x}^{(t)} - \mathbf{y}\right)\right\|_2^2.$$

Using (2.1) we obtain

$$\left\|\mathbf{d}^{(t+1)} - \mathbf{d}^{(t)} + \eta^{(t)}\mathbf{A}^{\mathsf{H}}\left(\mathbf{A}\left(\mathbf{d}^{(t)} - \mathbf{d}^{\star}\right) - \mathbf{e}\right)\right\|_2^2 \leq \left\|-\mathbf{d}^{(t)} + \eta^{(t)}\mathbf{A}^{\mathsf{H}}\left(\mathbf{A}\left(\mathbf{d}^{(t)} - \mathbf{d}^{\star}\right) - \mathbf{e}\right)\right\|_2^2.$$

Therefore, we obtain

$$\Re\left\langle \mathbf{d}^{(t+1)}, \mathbf{d}^{(t+1)} - 2\mathbf{d}^{(t)} + 2\eta^{(t)}\mathbf{A}^{\mathsf{H}}\left(\mathbf{A}\mathbf{d}^{(t)} - (\mathbf{A}\mathbf{d}^{\star} + \mathbf{e})\right)\right\rangle \leq 0$$

that yields the the desired result after straightforward algebraic manipulations. ∎

The following lemma is a special case of the generalized shifting inequality proposed in (Foucart, 2012, Theorem 2). Please refer to the reference for the proof.

Lemma D.2 (*Shifting Inequality* (Foucart, 2012))**.** *If* $0 < p < 2$ *and*

$$u_1 \geq u_2 \geq \cdots \geq u_l \geq u_{l+1} \geq \cdots \geq u_r \geq u_{r+1} \geq \cdots \geq u_{r+l} \geq 0,$$

then for $C = \max\left\{r^{\frac{1}{2}-\frac{1}{p}}, \sqrt{\frac{p}{2}}\left(\frac{2}{2-p}l\right)^{\frac{1}{2}-\frac{1}{p}}\right\}$,

$$\left(\sum_{i=l+1}^{l+r} u_i^2\right)^{\frac{1}{2}} \leq C\left(\sum_{i=1}^{r} u_i^p\right)^{\frac{1}{p}}.$$

Lemma D.3. *For* $\mathbf{x}_{\perp}^{\star}$*, a projection of* $\mathbf{x}^{\star}$ *onto* $\widehat{\mathcal{B}}$*, we have* $\operatorname{supp}(\mathbf{x}_{\perp}^{\star}) \subseteq \mathcal{S} = \operatorname{supp}(\mathbf{x}^{\star})$.

Proof. Proof is by contradiction. Suppose that there exists a coordinate i such that $x_i^{\star} = 0$ but $x_{\perp i}^{\star} \neq 0$. Then one can construct vector $\mathbf{x}'$ which is equal to $\mathbf{x}_{\perp}^{\star}$ except at the i-th coordinate where it is zero. Obviously $\mathbf{x}'$ is feasible because $\|\mathbf{x}'\|_p^p < \|\mathbf{x}_{\perp}^{\star}\|_p^p \leq \widehat{c}$. Furthermore,

$$\begin{aligned}
\left\|\mathbf{x}^{\star} - \mathbf{x}'\right\|_2^2 &= \sum_{j=1}^{n} \left|x_j^{\star} - x_j'\right|^2 \\
&= \sum_{\substack{j=1 \\ j\neq i}}^{n} \left|x_j^{\star} - x_{\perp j}^{\star}\right|^2 \\
&< \sum_{j=1}^{n} \left|x_j^{\star} - x_{\perp j}^{\star}\right|^2 \\
&= \|\mathbf{x}^{\star} - \mathbf{x}_{\perp}^{\star}\|_2^2.
\end{aligned}$$

Since by definition

$$\mathbf{x}_{\perp}^{\star} \in \arg\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{x}^{\star} - \mathbf{x}\|_2^2 \quad \text{s.t. } \|\mathbf{x}\|_p^p \leq \widehat{c},$$

we have a contradiction. ∎

To continue, we introduce the following sets which partition the coordinates of vector $\mathbf{d}^{(t)}$ for $t = 0, 1, 2, \ldots$. As defined previously in Lemma D.3, let $\mathcal{S} = \text{supp}\,(\mathbf{x}^{\star})$. Lemma D.3 shows that $\text{supp}\,(\mathbf{x}_{\perp}^{\star}) \subseteq \mathcal{S}$, thus we can assume that $\mathbf{x}_{\perp}^{\star}$ is s-sparse. Let $\mathcal{S}_{t,1}$ be the support of the s largest entries of $\mathbf{d}^{(t)}|_{\mathcal{S}^c}$ in magnitude, and define $\mathcal{T}_t = \mathcal{S} \cup \mathcal{S}_{t,1}$. Furthermore, let $\mathcal{S}_{t,2}$ be the support of the s largest entries of $\mathbf{d}^{(t)}|_{\mathcal{T}_t^c}$, $\mathcal{S}_{t,3}$ be the support of the next s largest entries of $\mathbf{d}^{(t)}|_{\mathcal{T}_t^c}$, and so on. We also set $\mathcal{T}_{t,j} = \mathcal{S}_{t,j} \cup \mathcal{S}_{t,j+1}$ for $j \geq 1$. This partitioning of the vector $\mathbf{d}^{(t)}$ is illustrated in Fig. D.1.

Lemma D.4. *For $t = 0, 1, 2, \ldots$ the vector $\mathbf{d}^{(t)}$ obeys*

$$\sum_{i \geq 2} \left\| \mathbf{d}^{(t)}|_{\mathcal{S}_{t,i}} \right\|_2 \leq \sqrt{2p} \left(\frac{2s}{2-p} \right)^{\frac{1}{2}-\frac{1}{p}} \left\| \mathbf{d}^{(t)}|_{\mathcal{S}^c} \right\|_p.$$

Proof. Since $\mathcal{S}_{t,j}$ and $\mathcal{S}_{t,j+1}$ are disjoint and $\mathcal{T}_{t,j} = \mathcal{S}_{t,j} \cup \mathcal{S}_{t,j+1}$ for $j \geq 1$, we have

$$\left\| \mathbf{d}^{(t)}|_{\mathcal{S}_{t,j}} \right\|_2 + \left\| \mathbf{d}^{(t)}|_{\mathcal{S}_{t,j+1}} \right\|_2 \leq \sqrt{2} \left\| \mathbf{d}^{(t)}|_{\mathcal{T}_{t,j}} \right\|_2.$$

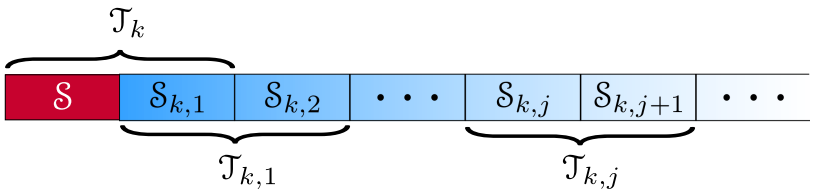


Figure D.1: Partitioning of vector $\mathbf{d}^{(t)} = \mathbf{x}^{(t)} - \mathbf{x}_{\perp}^{\star}$. The color gradient represents decrease of the magnitudes of the corresponding coordinates.

Adding over even j's then we deduce

$$\sum_{j\geq 2}\left\|\mathbf{d}^{(t)}|_{\mathcal{S}_{t,j}}\right\|_2 \leq \sqrt{2}\sum_{i\geq 1}\left\|\mathbf{d}^{(t)}|_{\mathcal{T}_{t,2i}}\right\|_2.$$

Because of the structure of the sets $\mathcal{T}_{t,j}$, Lemma D.2 can be applied to obtain

$$\left\|\mathbf{d}^{(t)}|_{\mathcal{T}_{t,j}}\right\|_2 \leq \sqrt{p}\left(\frac{2s}{2-p}\right)^{\frac{1}{2}-\frac{1}{p}}\left\|\mathbf{d}^{(t)}|_{\mathcal{T}_{t,j-1}}\right\|_p. \tag{D.1}$$

To be precise, based on Lemma D.2 the coefficient on the RHS should be

$$C=\max\left\{(2s)^{\frac{1}{2}-\frac{1}{p}}, \sqrt{\frac{p}{2}}\left(\frac{2s}{2-p}\right)^{\frac{1}{2}-\frac{1}{p}}\right\}.$$

For simplicity, however, we use the upper bound $C \leq \sqrt{p}\left(\frac{2s}{2-p}\right)^{\frac{1}{2}-\frac{1}{p}}$. To verify this upper bound it suffices to show that $(2s)^{\frac{1}{2}-\frac{1}{p}} \leq \sqrt{p}\left(\frac{2s}{2-p}\right)^{\frac{1}{2}-\frac{1}{p}}$ or equivalently $\phi(p) = p\log p + (2-p)\log(2-p) \geq 0$ for $p \in (0,1]$. Since $\phi(\cdot)$ is a deceasing function over $(0,1]$, it attains its minimum at $p=1$ which means that $\phi(p) \geq \phi(1) = 0$ as desired.

Then (D.1) yields

$$\sum_{j\geq 2}\left\|\mathbf{d}^{(t)}|_{\mathcal{S}_{t,j}}\right\|_2 \leq \sqrt{2p}\left(\frac{2s}{2-p}\right)^{\frac{1}{2}-\frac{1}{p}}\sum_{i\geq 1}\left\|\mathbf{d}^{(t)}|_{\mathcal{T}_{t,2i-1}}\right\|_p.$$

Since $\omega_1+\omega_2+\cdots+\omega_l \leq \left(\omega_1^p+\omega_2^p+\cdots+\omega_l^p\right)^{\frac{1}{p}}$ holds for $\omega_1,\cdots,\omega_l \geq 0$ and $p\in(0,1]$, we can write

$$\sum_{i\geq 1}\left\|\mathbf{d}^{(t)}|_{\mathcal{T}_{t,2i-1}}\right\|_p \leq \left(\sum_{i\geq 1}\left\|\mathbf{d}^{(t)}|_{\mathcal{T}_{t,2i-1}}\right\|_p^p\right)^{\frac{1}{p}}.$$

The desired result then follows using the fact that the sets $\mathcal{T}_{t,2i-1}$ are disjoint and $\mathcal{S}^{\mathrm{c}} = \bigcup_{i\geq 1}\mathcal{T}_{t,2i-1}$. ∎

Proof of the following Lemma mostly relies on some common inequalities that have

been used in the compressed sensing literature (see e.g., Chartrand (2007b, Theorem 2.1) and Gribonval and Nielsen (2007, Theorem 2)) .

Lemma D.5. *The error vector* $\mathbf{d}^{(t)}$ *satisfies* $\left\|\mathbf{d}^{(t)}|_{\mathcal{S}^c}\right\|_p \leq s^{\frac{1}{p}-\frac{1}{2}}\left\|\mathbf{d}^{(t)}|_{\mathcal{S}}\right\|_2$ *for all* $t = 0, 1, 2, \cdots$.

Proof. Since $\operatorname{supp}(\mathbf{x}^\star_\perp) \subseteq \mathcal{S} = \operatorname{supp}(\mathbf{x}^\star)$ we have $\mathbf{d}^{(t)}|_{\mathcal{S}^c} = \mathbf{x}^{(t)}|_{\mathcal{S}^c}$. Furthermore, because $\mathbf{x}^{(t)}$ is a feasible point by assumption we have $\left\|\mathbf{x}^{(t)}\right\|_p^p \leq \widehat{c} = \|\mathbf{x}^\star_\perp\|_p^p$ that implies,

$$\begin{aligned}
\left\|\mathbf{d}^{(t)}|_{\mathcal{S}^c}\right\|_p^p &= \left\|\mathbf{x}^{(t)}|_{\mathcal{S}^c}\right\|_p^p \\
&\leq \|\mathbf{x}^\star_\perp\|_p^p - \left\|\mathbf{x}^{(t)}|_{\mathcal{S}}\right\|_p^p \\
&\leq \left\|\mathbf{x}^\star_\perp - \mathbf{x}^{(t)}|_{\mathcal{S}}\right\|_p^p \\
&= \left\|\mathbf{d}^{(t)}|_{\mathcal{S}}\right\|_p^p \\
&\leq s^{1-\frac{p}{2}}\left\|\mathbf{d}^{(t)}|_{\mathcal{S}}\right\|_2^p, \qquad \text{(power means inequality)}
\end{aligned}$$

which yields the desired result. ∎

The next lemma is a straightforward extension of a previously known result (Davenport and Wakin, 2010, Lemma 3.1) to the case of complex vectors and asymmetric RIP.

Lemma D.6. *For* $\mathbf{u},\mathbf{v}\in\mathbb{C}^n$ *suppose that matrix* $\mathbf{A}$ *satisfies RIP of order* $\max\{\|\mathbf{u}+\mathbf{v}\|_0, \|\mathbf{u}-\mathbf{v}\|_0\}$ *with constants* α *and* β*. Then we have*

$$\left|\Re\left[\eta\left\langle\mathbf{A}\mathbf{u},\mathbf{A}\mathbf{v}\right\rangle - \left\langle\mathbf{u},\mathbf{v}\right\rangle\right]\right| \leq \left(\frac{\eta(\alpha-\beta)}{2} + \left|\frac{\eta(\alpha+\beta)}{2} - 1\right|\right)\|\mathbf{u}\|_2\|\mathbf{v}\|_2.$$

Proof. If either of the vectors $\mathbf{u}$ and $\mathbf{v}$ is zero the claim becomes trivial. So without loss of generality we assume that none of these vectors is zero. The RIP condition holds for the vectors $\mathbf{u}\pm\mathbf{v}$ and we have

$$\beta\|\mathbf{u}\pm\mathbf{v}\|_2^2 \leq \|\mathbf{A}(\mathbf{u}\pm\mathbf{v})\|_2^2 \leq \alpha\|\mathbf{u}\pm\mathbf{v}\|_2^2.$$

Therefore, we obtain

$$\begin{aligned}\Re\left\langle \mathbf{A}\mathbf{u}, \mathbf{A}\mathbf{v}\right\rangle &= \frac{1}{4}\left(\left\|\mathbf{A}\left(\mathbf{u}+\mathbf{v}\right)\right\|_2^2 - \left\|\mathbf{A}\left(\mathbf{u}-\mathbf{v}\right)\right\|_2^2\right)\\ &\leq \frac{1}{4}\left(\alpha\|\mathbf{u}+\mathbf{v}\|_2^2 - \beta\|\mathbf{u}-\mathbf{v}\|_2^2\right)\\ &= \frac{\alpha-\beta}{4}\left(\|\mathbf{u}\|_2^2+\|\mathbf{v}\|_2^2\right) + \frac{\alpha+\beta}{2}\Re\left\langle \mathbf{u},\mathbf{v}\right\rangle.\end{aligned}$$

Applying this inequality for vectors $\frac{\mathbf{u}}{\|\mathbf{u}\|_2}$ and $\frac{\mathbf{v}}{\|\mathbf{v}\|_2}$ yields

$$\begin{aligned}\Re\left[\eta\left\langle \mathbf{A}\frac{\mathbf{u}}{\|\mathbf{u}\|_2}, \mathbf{A}\frac{\mathbf{v}}{\|\mathbf{v}\|_2}\right\rangle - \left\langle \frac{\mathbf{u}}{\|\mathbf{u}\|_2}, \frac{\mathbf{v}}{\|\mathbf{v}\|_2}\right\rangle\right] &\leq \frac{\eta\left(\alpha-\beta\right)}{2} + \left(\frac{\eta\left(\alpha+\beta\right)}{2}-1\right)\Re\left\langle \frac{\mathbf{u}}{\|\mathbf{u}\|_2}, \frac{\mathbf{v}}{\|\mathbf{v}\|_2}\right\rangle\\ &\leq \frac{\eta\left(\alpha-\beta\right)}{2} + \left|\frac{\eta\left(\alpha+\beta\right)}{2}-1\right|.\end{aligned}$$

Similarly it can be shown that

$$\Re\left[\eta\left\langle \mathbf{A}\frac{\mathbf{u}}{\|\mathbf{u}\|_2}, \mathbf{A}\frac{\mathbf{v}}{\|\mathbf{v}\|_2}\right\rangle - \left\langle \frac{\mathbf{u}}{\|\mathbf{u}\|_2}, \frac{\mathbf{v}}{\|\mathbf{v}\|_2}\right\rangle\right] \geq -\frac{\eta\left(\alpha-\beta\right)}{2} - \left|\frac{\eta\left(\alpha+\beta\right)}{2}-1\right|.$$

The desired result follows by multiplying the last two inequalities by $\|\mathbf{u}\|_2\|\mathbf{v}\|_2$. ∎

Lemma D.7. *If the step-size of ℓ_p-PGD obeys $\left|\eta^{(t)}\left(\alpha_{3s}+\beta_{3s}\right)/2-1\right| \leq \tau$ for some $\tau \geq 0$, then we have*

$$\begin{aligned}\Re\left[\left\langle \mathbf{d}^{(t)}, \mathbf{d}^{(t+1)}\right\rangle - \eta^{(t)}\left\langle \mathbf{A}\mathbf{d}^{(t)}, \mathbf{A}\mathbf{d}^{(t+1)}\right\rangle\right] &\leq \left(\left(1+\tau\right)\rho_{3s}+\tau\right)\left(1+\sqrt{2p}\left(\frac{2}{2-p}\right)^{\frac{1}{2}-\frac{1}{p}}\right)^2\\ &\quad\times\left\|\mathbf{d}^{(t)}\right\|_2\left\|\mathbf{d}^{(t+1)}\right\|_2.\end{aligned}$$

Proof. Note that

$$\begin{aligned}&\Re\left[\left\langle \mathbf{d}^{(t)}, \mathbf{d}^{(t+1)}\right\rangle - \eta^{(t)}\left\langle \mathbf{A}\mathbf{d}^{(t)}, \mathbf{A}\mathbf{d}^{(t+1)}\right\rangle\right]\\ &= \Re\left[\left\langle \mathbf{d}^{(t)}|_{\mathcal{T}_t}, \mathbf{d}^{(t+1)}|_{\mathcal{T}_{t+1}}\right\rangle - \eta^{(t)}\left\langle \mathbf{A}\mathbf{d}^{(t)}|_{\mathcal{T}_t}, \mathbf{A}\mathbf{d}^{(t+1)}|_{\mathcal{T}_{t+1}}\right\rangle\right]\\ &+ \sum_{i\geq 2}\Re\left[\left\langle \mathbf{d}^{(t)}|_{\mathcal{S}_{t,i}}, \mathbf{d}^{(t+1)}|_{\mathcal{T}_{t+1}}\right\rangle - \eta^{(t)}\left\langle \mathbf{A}\mathbf{d}^{(t)}|_{\mathcal{S}_{t,i}}, \mathbf{A}\mathbf{d}^{(t+1)}|_{\mathcal{T}_{t+1}}\right\rangle\right]\end{aligned}$$

$$+\sum_{j\geq 2}\Re\left[\left\langle \mathbf{d}^{(t)}|_{\mathcal{T}_t},\mathbf{d}^{(t+1)}|_{\mathcal{S}_{t+1,j}}\right\rangle-\eta^{(t)}\left\langle \mathbf{A}\mathbf{d}^{(t)}|_{\mathcal{T}_t},\mathbf{A}\mathbf{d}^{(t+1)}|_{\mathcal{S}_{t+1,j}}\right\rangle\right]$$
$$+\sum_{i,j\geq 2}\Re\left[\left\langle \mathbf{d}^{(t)}|_{\mathcal{S}_{t,i}},\mathbf{d}^{(t+1)}|_{\mathcal{S}_{t+1,j}}\right\rangle-\eta^{(t)}\left\langle \mathbf{A}\mathbf{d}^{(t)}|_{\mathcal{S}_{t,i}},\mathbf{A}\mathbf{d}^{(t+1)}|_{\mathcal{S}_{t+1,j}}\right\rangle\right]. \tag{D.2}$$

Note that $|\mathcal{T}_t\cup\mathcal{T}_{t+1}|\leq 3s$. Furthermore, for $i,j\geq 2$ we have $|\mathcal{T}_t\cup\mathcal{S}_{t+1,j}|\leq 3s$, $|\mathcal{T}_{t+1}\cup\mathcal{S}_{t,i}|\leq 3s$, and $|\mathcal{S}_{t,i}\cup\mathcal{S}_{t+1,j}|\leq 2s$. Therefore, by applying Lemma D.6 for each of the summands in (D.2) and using the fact that

$$\begin{aligned}\rho'_{3s}&:=(1+\tau)\,\rho_{3s}+\tau\\&\geq \eta^{(t)}\left(\alpha_{3s}-\beta_{3s}\right)/2+\left|\eta^{(t)}\left(\alpha_{3s}+\beta_{3s}\right)/2-1\right|\end{aligned}$$

we obtain

$$\begin{aligned}\Re\left[\left\langle \mathbf{d}^{(t)},\mathbf{d}^{(t+1)}\right\rangle-\eta^{(t)}\left\langle \mathbf{A}\mathbf{d}^{(t)},\mathbf{A}\mathbf{d}^{(t+1)}\right\rangle\right]&\leq\rho'_{3s}\left\|\mathbf{d}^{(t)}|_{\mathcal{T}_t}\right\|_2\left\|\mathbf{d}^{(t+1)}|_{\mathcal{T}_{t+1}}\right\|_2\\&+\sum_{i\geq 2}\rho'_{3s}\left\|\mathbf{d}^{(t)}|_{\mathcal{S}_{t,i}}\right\|_2\left\|\mathbf{d}^{(t+1)}|_{\mathcal{T}_{t+1}}\right\|_2\\&+\sum_{j\geq 2}\rho'_{3s}\left\|\mathbf{d}^{(t)}|_{\mathcal{T}_t}\right\|_2\left\|\mathbf{d}^{(t+1)}|_{\mathcal{S}_{t+1,j}}\right\|_2\\&+\sum_{i,j\geq 2}\rho'_{3s}\left\|\mathbf{d}^{(t)}|_{\mathcal{S}_{t,i}}\right\|_2\left\|\mathbf{d}^{(t+1)}|_{\mathcal{S}_{t+1,j}}\right\|_2.\end{aligned}$$

Hence, applying Lemma D.4 yields

$$\begin{aligned}\Re\left[\left\langle \mathbf{d}^{(t)},\mathbf{d}^{(t+1)}\right\rangle-\eta^{(t)}\left\langle \mathbf{A}\mathbf{d}^{(t)},\mathbf{A}\mathbf{d}^{(t+1)}\right\rangle\right]&\leq\rho'_{3s}\left\|\mathbf{d}^{(t)}|_{\mathcal{T}_t}\right\|_2\left\|\mathbf{d}^{(t+1)}|_{\mathcal{T}_{t+1}}\right\|_2\\&+\sqrt{2p}\left(\frac{2s}{2-p}\right)^{\frac{1}{2}-\frac{1}{p}}\rho'_{3s}\left\|\mathbf{d}^{(t)}|_{\mathcal{S}^c}\right\|_p\left\|\mathbf{d}^{(t+1)}|_{\mathcal{T}_{t+1}}\right\|_2\\&+\sqrt{2p}\left(\frac{2s}{2-p}\right)^{\frac{1}{2}-\frac{1}{p}}\rho'_{3s}\left\|\mathbf{d}^{(t)}|_{\mathcal{T}_t}\right\|_2\left\|\mathbf{d}^{(t+1)}|_{\mathcal{S}^c}\right\|_p\\&+2p\left(\frac{2s}{2-p}\right)^{1-\frac{2}{p}}\rho'_{3s}\left\|\mathbf{d}^{(t)}|_{\mathcal{S}^c}\right\|_p\left\|\mathbf{d}^{(t+1)}|_{\mathcal{S}^c}\right\|_p.\end{aligned}$$

Then it follows from Lemma D.5,

$$\begin{aligned}\Re\left[\left\langle \mathbf{d}^{(t)}, \mathbf{d}^{(t+1)}\right\rangle - \eta^{(t)}\left\langle \mathbf{A}\mathbf{d}^{(t)}, \mathbf{A}\mathbf{d}^{(t+1)}\right\rangle\right] &\le \rho'_{3s}\left\|\mathbf{d}^{(t)}|_{\mathcal{T}_t}\right\|_2\left\|\mathbf{d}^{(t+1)}|_{\mathcal{T}_{t+1}}\right\|_2 \\ &+ \sqrt{2p}\left(\frac{2}{2-p}\right)^{\frac{1}{2}-\frac{1}{p}}\rho'_{3s}\left\|\mathbf{d}^{(t)}|_{\mathcal{S}}\right\|_2\left\|\mathbf{d}^{(t+1)}|_{\mathcal{T}_{t+1}}\right\|_2 \\ &+ \sqrt{2p}\left(\frac{2}{2-p}\right)^{\frac{1}{2}-\frac{1}{p}}\rho'_{3s}\left\|\mathbf{d}^{(t)}|_{\mathcal{T}_t}\right\|_2\left\|\mathbf{d}^{(t+1)}|_{\mathcal{S}}\right\|_2 \\ &+ 2p\left(\frac{2}{2-p}\right)^{1-\frac{2}{p}}\rho'_{3s}\left\|\mathbf{d}^{(t)}|_{\mathcal{S}}\right\|_2\left\|\mathbf{d}^{(t+1)}|_{\mathcal{S}}\right\|_2 \\ &\le \rho'_{3s}\left(1+\sqrt{2p}\left(\frac{2}{2-p}\right)^{\frac{1}{2}-\frac{1}{p}}\right)^2\left\|\mathbf{d}^{(t)}\right\|_2\left\|\mathbf{d}^{(t+1)}\right\|_2,\end{aligned}$$

which is the desired result. ∎

Now we are ready to prove the accuracy guarantees for the ℓ_p-PGD algorithm.

Proof of Theorem 6.1. Recall that γ is defined by (6.5). It follows from Lemmas D.1 and D.7 that

$$\begin{aligned}\left\|\mathbf{d}^{(t)}\right\|_2^2 &\le 2\gamma\left\|\mathbf{d}^{(t)}\right\|_2\left\|\mathbf{d}^{(t-1)}\right\|_2 + 2\eta^{(t)}\Re\left\langle \mathbf{A}\mathbf{d}^{(t)}, \mathbf{A}\mathbf{d}^{\star}+\mathbf{e}\right\rangle \\ &\le 2\gamma\left\|\mathbf{d}^{(t)}\right\|_2\left\|\mathbf{d}^{(t-1)}\right\|_2 + 2\eta^{(t)}\left\|\mathbf{A}\mathbf{d}^{(t)}\right\|_2\|\mathbf{A}\mathbf{d}^{\star}+\mathbf{e}\|_2.\end{aligned}$$

Furthermore, using (D.1) and Lemma D.5 we deduce

$$\begin{aligned}\left\|\mathbf{A}\mathbf{d}^{(t)}\right\|_2 &\le \left\|\mathbf{A}\mathbf{d}^{(t)}|_{\mathcal{T}_t}\right\|_2 + \sum_{i\ge 1}\left\|\mathbf{A}\mathbf{d}^{(t)}|_{\mathcal{T}_{t,2i}}\right\|_2 \\ &\le \sqrt{\alpha_{2s}}\left\|\mathbf{d}^{(t)}|_{\mathcal{T}_t}\right\|_2 + \sum_{i\ge 1}\sqrt{\alpha_{2s}}\left\|\mathbf{d}^{(t)}|_{\mathcal{T}_{t,2i}}\right\|_2 \\ &\le \sqrt{\alpha_{2s}}\left\|\mathbf{d}^{(t)}|_{\mathcal{T}_t}\right\|_2 + \sqrt{\alpha_{2s}}\sqrt{p}\left(\frac{2s}{2-p}\right)^{\frac{1}{2}-\frac{1}{p}}\sum_{i\ge 1}\left\|\mathbf{d}^{(t)}|_{\mathcal{T}_{t,2i-1}}\right\|_p \\ &\le \sqrt{\alpha_{2s}}\left\|\mathbf{d}^{(t)}|_{\mathcal{T}_t}\right\|_2 + \sqrt{\alpha_{2s}}\sqrt{p}\left(\frac{2s}{2-p}\right)^{\frac{1}{2}-\frac{1}{p}}\left\|\mathbf{d}^{(t)}|_{\mathcal{S}^c}\right\|_p\end{aligned}$$

$$\leq \sqrt{\alpha_{2s}}\left\|\mathbf{d}^{(t)}|_{\mathcal{T}_t}\right\|_2 + \sqrt{\alpha_{2s}}\sqrt{p}\left(\frac{2}{2-p}\right)^{\frac{1}{2}-\frac{1}{p}}\left\|\mathbf{d}^{(t)}|_{\mathcal{S}}\right\|_2$$
$$\leq \sqrt{\alpha_{2s}}\left(1+\sqrt{p}\left(\frac{2}{2-p}\right)^{\frac{1}{2}-\frac{1}{p}}\right)\left\|\mathbf{d}^{(t)}\right\|_2.$$

Therefore,

$$\left\|\mathbf{d}^{(t)}\right\|_2^2 \leq 2\gamma\left\|\mathbf{d}^{(t)}\right\|_2\left\|\mathbf{d}^{(t-1)}\right\|_2 + 2\eta^{(t)}\sqrt{\alpha_{2s}}\left(1+\sqrt{p}\left(\frac{2}{2-p}\right)^{\frac{1}{2}-\frac{1}{p}}\right)\left\|\mathbf{d}^{(t)}\right\|_2\|\mathbf{A}\mathbf{d}^\star+\mathbf{e}\|_2,$$

which after canceling $\left\|\mathbf{d}^{(t)}\right\|_2$ yields

$$\begin{aligned}\left\|\mathbf{d}^{(t)}\right\|_2 &\leq 2\gamma\left\|\mathbf{d}^{(t-1)}\right\|_2 + 2\eta^{(t)}\sqrt{\alpha_{2s}}\left(1+\sqrt{p}\left(\frac{2}{2-p}\right)^{\frac{1}{2}-\frac{1}{p}}\right)\|\mathbf{A}\mathbf{d}^\star+\mathbf{e}\|_2\\ &= 2\gamma\left\|\mathbf{d}^{(t-1)}\right\|_2 + 2\eta^{(t)}\left(\alpha_{3s}+\beta_{3s}\right)\frac{\sqrt{\alpha_{2s}}}{\alpha_{3s}+\beta_{3s}}\left(1+\sqrt{p}\left(\frac{2}{2-p}\right)^{\frac{1}{2}-\frac{1}{p}}\right)\|\mathbf{A}\mathbf{d}^\star+\mathbf{e}\|_2\\ &\leq 2\gamma\left\|\mathbf{d}^{(t-1)}\right\|_2 + 4\left(1+\tau\right)\frac{\sqrt{\alpha_{2s}}}{\alpha_{3s}+\beta_{3s}}\left(1+\sqrt{p}\left(\frac{2}{2-p}\right)^{\frac{1}{2}-\frac{1}{p}}\right)\left(\|\mathbf{A}\mathbf{d}^\star\|_2+\|\mathbf{e}\|_2\right).\end{aligned}$$

Since $\mathbf{x}_\perp^\star$ is a projection of $\mathbf{x}^\star$ onto the feasible set $\widehat{\mathcal{B}}$ and $\left(\frac{\widehat{c}}{\|\mathbf{x}^\star\|_p^p}\right)^{1/p}\mathbf{x}^\star \in \widehat{\mathcal{B}}$ we have

$$\begin{aligned}\|\mathbf{d}^\star\|_2 &= \|\mathbf{x}_\perp^\star - \mathbf{x}^\star\|_2\\ &\leq \left\|\left(\frac{\widehat{c}}{\|\mathbf{x}^\star\|_p^p}\right)^{1/p}\mathbf{x}^\star - \mathbf{x}^\star\right\|_2 = \epsilon\|\mathbf{x}^\star\|_2.\end{aligned}$$

Furthermore, $\text{supp}\left(\mathbf{d}^\star\right) \subseteq \mathcal{S}$, thereby we can use RIP to obtain

$$\begin{aligned}\|\mathbf{A}\mathbf{d}^\star\|_2 &\leq \sqrt{\alpha_s}\|\mathbf{d}^\star\|_2\\ &\leq \epsilon\sqrt{\alpha_s}\|\mathbf{x}^\star\|_2.\end{aligned}$$

Hence,

$$\left\|\mathbf{d}^{(t)}\right\|_2 \leq 2\gamma\left\|\mathbf{d}^{(t-1)}\right\|_2 + 4\left(1+\tau\right)\frac{\sqrt{\alpha_{2s}}}{\alpha_{3s}+\beta_{3s}}\left(1+\sqrt{p}\left(\frac{2}{2-p}\right)^{\frac{1}{2}-\frac{1}{p}}\right)\left(\epsilon\sqrt{\alpha_s}\|\mathbf{x}^\star\|_2+\|\mathbf{e}\|_2\right)$$

$$
\begin{aligned}
&\leq 2\gamma \left\| \mathbf{d}^{(t-1)} \right\|_2 \\
&+ 2\left(1+\tau\right)\left(1+\sqrt{p}\left(\frac{2}{2-p}\right)^{\frac{1}{2}-\frac{1}{p}}\right)\left(\epsilon\left(1+\rho_{3s}\right)\left\|\mathbf{x}^{\star}\right\|_2 + \frac{2\sqrt{\alpha_{2s}}}{\alpha_{3s}+\beta_{3s}}\|\mathbf{e}\|_2\right).
\end{aligned}
$$

Applying this inequality recursively and using the fact that

$$
\sum_{i=0}^{t-1}\left(2\gamma\right)^i < \sum_{i=0}^{\infty}\left(2\gamma\right)^i = \frac{1}{1-2\gamma},
$$

which holds because of the assumption $\gamma < \frac{1}{2}$, we can finally deduce

$$
\begin{aligned}
\left\|\mathbf{x}^{(t)} - \mathbf{x}^{\star}\right\|_2 &= \left\|\mathbf{d}^{(t)} - \mathbf{d}^{\star}\right\|_2 \\
&\leq \left\|\mathbf{d}^{(t)}\right\|_2 + \|\mathbf{d}^{\star}\|_2 \\
&\leq \left(2\gamma\right)^t \left\|\mathbf{x}_{\perp}^{\star}\right\|_2 + \frac{2\left(1+\tau\right)}{1-2\gamma}\left(1+\xi\left(p\right)\right)\left(\epsilon\left(1+\rho_{3s}\right)\left\|\mathbf{x}^{\star}\right\|_2 + \frac{2\sqrt{\alpha_{2s}}}{\alpha_{3s}+\beta_{3s}}\|\mathbf{e}\|_2\right) \\
&\quad + \|\mathbf{d}^{\star}\|_2 \\
&\leq \left(2\gamma\right)^t \left\|\mathbf{x}^{\star}\right\|_2 + \frac{2\left(1+\tau\right)}{1-2\gamma}\left(1+\xi\left(p\right)\right)\left(\epsilon\left(1+\rho_{3s}\right)\left\|\mathbf{x}^{\star}\right\|_2 + \frac{2\sqrt{\alpha_{2s}}}{\alpha_{3s}+\beta_{3s}}\|\mathbf{e}\|_2\right) \\
&\quad + \epsilon\|\mathbf{x}^{\star}\|_2,
\end{aligned}
$$

where $\xi\left(p\right) = \sqrt{p}\left(\frac{2}{2-p}\right)^{\frac{1}{2}-\frac{1}{p}}$ as defined in the statement of the theorem. ∎

D.2 Lemmas for Characterization of a Projection onto ℓ_p-balls

In what follows we assume that $\mathcal{B}$ is an ℓ_p-ball with p-radius c (i.e., $\mathcal{B} = \mathcal{F}_p\left(c\right)$). For $\mathbf{x} \in \mathbb{C}^n$ we derive some properties of

$$
\mathbf{x}^{\perp} \in \arg\min \ \frac{1}{2}\|\mathbf{x}-\mathbf{u}\|_2^2 \quad \text{s.t. } \mathbf{u} \in \mathcal{B}, \tag{D.3}
$$

a projection of $\mathbf{x}$ onto $\mathcal{B}$.

Lemma D.8. *Let $\mathbf{x}^{\perp}$ be a projection of $\mathbf{x}$ onto $\mathcal{B}$. Then for every $i \in \{1,2,\ldots,n\}$ we have*

$\mathrm{Arg}\,(x_i) = \mathrm{Arg}\left(x_i^{\perp}\right)$ *and* $\left|x_i^{\perp}\right| \leq |x_i|$.

Proof. Proof by contradiction. Suppose that for some i we have $\mathrm{Arg}\,(x_i) \neq \mathrm{Arg}\left(x_i^{\perp}\right)$ or $\left|x_i^{\perp}\right| > |x_i|$. Consider the vector $\mathbf{x}'$ for which $x'_j = x_j^{\perp}$ for $j \neq i$ and

$$x'_i = \min\left\{|x_i|\,, \left|x_i^{\perp}\right|\right\} \exp\left(\imath \mathrm{Arg}\,(x_i)\right),$$

where the character $\imath$ denotes the imaginary unit $\sqrt{-1}$. We have $\|\mathbf{x}'\|_p \leq \left\|\mathbf{x}^{\perp}\right\|_p$ which implies that $\mathbf{x}' \in \mathcal{B}$. Since $|x_i - x'_i| < \left|x_i - x_i^{\perp}\right|$ we have $\|\mathbf{x}' - \mathbf{x}\|_2 < \left\|\mathbf{x}^{\perp} - \mathbf{x}\right\|_2$ which contradicts the choice of $\mathbf{x}^{\perp}$ as a projection. ∎

Assumption. *Lemma D.8 asserts that the projection $\mathbf{x}^{\perp}$ has the same phase components as $\mathbf{x}$. Therefore, without loss of generality and for simplicity in the following lemmas we assume $\mathbf{x}$ has real-valued non-negative entries.*

Lemma D.9. *For any $\mathbf{x}$ in the positive orthant there is a projection $\mathbf{x}^{\perp}$ of $\mathbf{x}$ onto the set $\mathcal{B}$ such that for $i, j \in \{1, 2, \ldots, n\}$ we have $x_i^{\perp} \leq x_j^{\perp}$ iff $x_i \leq x_j$.*

Proof. Note that the set $\mathcal{B}$ is closed under any permutation of coordinates. In particular, by interchanging the i-th and j-th entries of $\mathbf{x}^{\perp}$ we obtain another vector $\mathbf{x}'$ in $\mathcal{B}$. Since $\mathbf{x}^{\perp}$ is a projection of $\mathbf{x}$ onto $\mathcal{B}$ we must have $\left\|\mathbf{x} - \mathbf{x}^{\perp}\right\|_2^2 \leq \|\mathbf{x} - \mathbf{x}'\|_2^2$. Therefore, we have $\left(x_i - x_i^{\perp}\right)^2 + \left(x_j - x_j^{\perp}\right)^2 \leq \left(x_i - x_j^{\perp}\right)^2 + \left(x_j - x_i^{\perp}\right)^2$ and from that $0 \leq (x_i - x_j)\left(x_i^{\perp} - x_j^{\perp}\right)$. For $x_i \neq x_j$ the result follows immediately, and for $x_i = x_j$ without loss of generality we can assume $x_i^{\perp} \leq x_j^{\perp}$. ∎

Lemma D.10. *Let $\mathcal{S}^{\perp}$ be the support set of $\mathbf{x}^{\perp}$. Then there exists a $\lambda \geq 0$ such that*

$$x_i^{\perp(1-p)}\left(x_i - x_i^{\perp}\right) = p\lambda$$

for all $i \in \mathcal{S}^{\perp}$.

Proof. The fact that $\mathbf{x}^{\perp}$ is a solution to the minimization expressed in (D.3) implies that that $\mathbf{x}^{\perp}|_{\mathcal{S}^{\perp}}$ must be a solution to

$$\arg\min_{\mathbf{v}} \frac{1}{2} \left\|\mathbf{x}|_{\mathcal{S}^{\perp}} - \mathbf{v}\right\|_2^2 \quad \text{s.t. } \|\mathbf{v}\|_p^p \leq c.$$

The normal to the feasible set (i.e., the gradient of the constraint function) is uniquely defined at $\mathbf{x}^{\perp}|_{\mathcal{S}^{\perp}}$ since all of its entries are positive by assumption. Consequently, the Lagrangian

$$L(\mathbf{v}, \lambda) = \frac{1}{2} \left\|\mathbf{x}|_{\mathcal{S}^{\perp}} - \mathbf{v}\right\|_2^2 + \lambda \left(\|\mathbf{v}\|_p^p - c\right)$$

has a well-defined partial derivative $\frac{\partial L}{\partial \mathbf{v}}$ at $\mathbf{x}^{\perp}|_{\mathcal{S}^{\perp}}$ which must be equal to zero for an appropriate $\lambda \geq 0$. Hence,

$$\forall i \in \mathcal{S}^{\perp} \; x_i^{\perp} - x_i + p\lambda x_i^{\perp(p-1)} = 0$$

which is equivalent to the desired result. ∎

Lemma D.11. *Let $\lambda \geq 0$ and $p \in [0,1]$ be fixed numbers and set $T_0 = (2-p)\left(p\,(1-p)^{p-1}\,\lambda\right)^{\frac{1}{2-p}}$. Denote the function $t^{1-p}\,(T-t)$ by $h_p(t)$. The following statements hold regarding the roots of $h_p(t) = p\lambda$:*

(i) For $p = 1$ and $T \geq T_0$ the equation $h_1(t) = \lambda$ has a unique solution at $t = T - \lambda \in [0, T]$ which is an increasing function of T.

(ii) For $p \in [0,1)$ and $T \geq T_0$ the equation $h_p(t) = p\lambda$ has two roots t_- and t_+ satisfying $t_- \in \left(0, \frac{1-p}{2-p}T\right]$ and $t_+ \in \left[\frac{1-p}{2-p}T, +\infty\right)$. As a function of T, t_- and t_+ are decreasing and increasing, respectively and they coincide at $T = T_0$.

Proof. Fig. D.2 illustrates $h_p(t)$ for different values of $p \in [0,1]$. To verify part (i) observe that we have $T_0 = \lambda$ thereby $T \geq \lambda$. The claim is then obvious since $h_1(t) - \lambda = T - t - \lambda$ is

zero at $t = T - \lambda$. Part (ii) is more intricate and we divide it into two cases: $p = 0$ and $p \neq 0$. At $p = 0$ we have $T_0 = 0$ and $h_0(t) = t(T - t)$ has two zeros at $t_- = 0$ and $t_+ = T$ that obviously satisfy the claim. So we can now focus on the case $p \in (0, 1)$. It is straightforward to verify that $t_{\max} = \frac{1-p}{2-p}T$ is the location at which $h_p(t)$ peaks. Straightforward algebraic manipulations also show that $T > T_0$ is equivalent to $p\lambda < h_p(t_{\max})$. Furthermore, inspecting the sign of $h_p'(t)$ shows that $h_p(t)$ is strictly increasing over $[0, t_{\max}]$ while it is strictly decreasing over $[t_{\max}, T]$. Then, using the fact that $h_p(0) = h_p(T) = 0 \leq p\lambda < h_p(t_{\max})$, it follows from the *intermediate value theorem* that $h_p(t) = p\lambda$ has exactly two roots, t_- and t_+, that straddle $t_{\max}$ as claimed. Furthermore, taking the derivative of $t_-^{1-p}(T - t_-) = p\lambda$ with respect to T yields

$$(1-p)\, t_-' t_-^{-p}(T - t_-) + t_-^{1-p}\left(1 - t_-'\right) = 0.$$

Hence,

$$\left((1-p)(T - t_-) - t_-\right) t_-' = -t_-$$

which because $t_- \leq t_{\max} = \frac{1-p}{2-p}T$ implies that $t_-' < 0$. Thus t_- is a decreasing function of T. Similarly we can show that t_+ is an increasing function of T using the fact that $t_+ \geq t_{\max}$. Finally, as T decreases to T_0 the peak value $h_p(t_{\max})$ decreases to $p\lambda$ which implies that t_- and t_+ both tend to the same value of $\frac{1-p}{2-p}T_0$. ∎

Lemma D.12. *Suppose that $x_i = x_j > 0$ for some $i \neq j$. If $x_i^\perp = x_j^\perp > 0$ then $x_i^\perp \geq \frac{1-p}{2-p}x_i$.*

Proof. For $p \in \{0, 1\}$ the claim is obvious since at $p = 0$ we have $x_i^\perp = x_i > \frac{1}{2}x_i$ and at $p = 1$ we have $\frac{1-p}{2-p}x_i = 0$. Therefore, without loss of generality we assume $p \in (0, 1)$. The proof is by contradiction. Suppose that $w = \frac{x_i^\perp}{x_i} = \frac{x_j^\perp}{x_j} < \frac{1-p}{2-p}$. Since $\mathbf{x}^\perp$ is a projection it follows

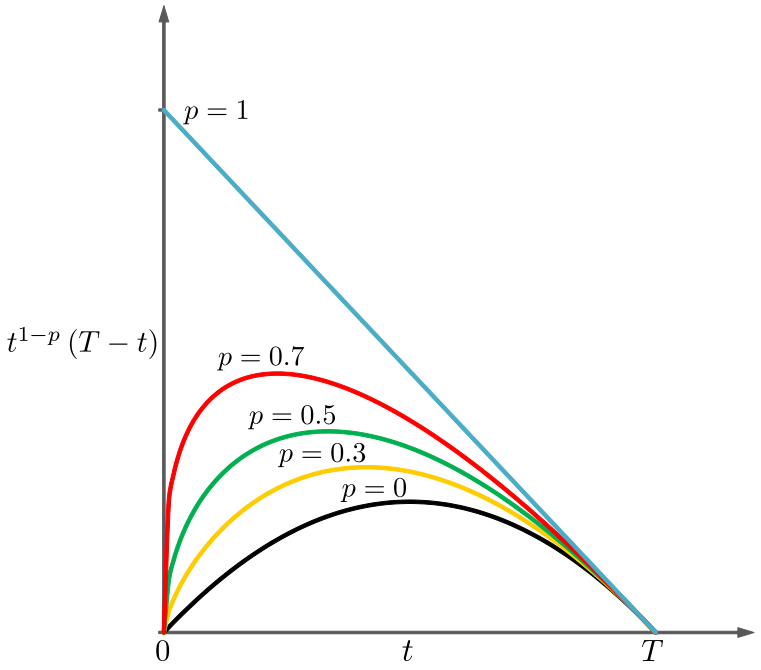


Figure D.2: The function $t^{1-p}(T-t)$ for different values of p

that $a = b = w$ must be the solution to

$$\arg\min_{a,b} \ \psi = \frac{1}{2}\left[(1-a)^2 + (1-b)^2\right] \ \text{ s.t. } a^p + b^p = 2w^p, \ a > 0, \ \text{and } b > 0,$$

otherwise the vector $\mathbf{x}'$ that is identical to $\mathbf{x}^{\perp}$ except for $x'_i = ax_i \neq x_i^{\perp}$ and $x'_j = bx_j \neq x_i^{\perp}$ is also a feasible point (i.e., $\mathbf{x}' \in \mathcal{B}$) that satisfies

$$\begin{aligned}\left\|\mathbf{x}' - \mathbf{x}\right\|_2^2 - \left\|\mathbf{x}^{\perp} - \mathbf{x}\right\|_2^2 &= (1-a)^2 x_i^2 + (1-b)^2 x_j^2 - (1-w)^2 x_i^2 - (1-w)^2 x_j^2 \\ &= \left((1-a)^2 + (1-b)^2 - 2(1-w)^2\right) x_i^2 < 0,\end{aligned}$$

which is absurd. If b is considered as a function of a then ψ can be seen merely as a function of a, i.e., $\psi \equiv \psi(a)$. Taking the derivative of ψ with respect to a yields

$$\psi'(a) = a - 1 + b'(b-1)$$

$$
\begin{aligned}
&= a - 1 - \left(\frac{a}{b}\right)^{p-1}(b-1) \\
&= \left(b^{1-p}(1-b) - a^{1-p}(1-a)\right) a^{p-1} \\
&= (2-p)(b-a)\nu^{-p}\left(\frac{1-p}{2-p} - \nu\right),
\end{aligned}
$$

where the last equation holds by the *mean value theorem* for some $\nu \in (\min\{a,b\}, \max\{a,b\})$. Since $w < \frac{1-p}{2-p}$ we have $r_1 := \min\left\{2^{1/p}w, \frac{1-p}{2-p}\right\} > w$ and $r_0 := (2w^p - r_1^p)^{1/p} < w$. With straightforward algebra one can show that if either a or b belongs to the interval $[r_0, r_1]$, then so does the other one. By varying a in $[r_0, r_1]$ we always have $\nu < r_1 \leq \frac{1-p}{2-p}$, therefore as a increases in this interval the sign of ψ' changes at $a = w$ from positive to negative. Thus, $a = b = w$ is a local maximum of ψ which is a contradiction. ∎

Bibliography

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*, volume 23, pages 37–45. 2010. long version available at arXiv:1104.4824v1 [stat.ML].

F. Bach. Structured sparsity-inducing norms through submodular functions. In *Advances in Neural Information Processing Systems*, volume 23, pages 118–126, Vancouver, BC, Canada, Dec. 2010.

F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Structured sparsity through convex optimization. *Statistical Science*, 27(4):450–468, Nov. 2012.

F. R. Bach. Consistency of the group Lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9:1179–1225, June 2008.

S. Bahmani and B. Raj. A unifying analysis of projected gradient descent for ℓ_p-constrained least squares. *Applied and Computational Harmonic Analysis*, 34(3):366–378, May 2013.

S. Bahmani, P. Boufounos, and B. Raj. Greedy sparsity-constrained optimization. In *Conference Record of the 45th Asilomar Conference on Signals, Systems, and Computers*, pages 1148–1152, Pacific Grove, CA, Nov. 2011.

S. Bahmani, P. T. Boufounos, and B. Raj. Learning model-based sparsity via projected gradient descent. arXiv:1209.1557 [stat.ML], Nov. 2012.

S. Bahmani, B. Raj, and P. T. Boufounos. Greedy sparsity-constrained optimization. *Journal of Machine Learning Research*, 2013. To appear; preprint available at arxiv:1203.5483v1[stat.ML].

R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde. Model-based compressive sensing. *IEEE Transactions on Information Theory*, 56(4):1982–2001, 2010.

A. Beck and Y. C. Eldar. Sparsity constrained nonlinear optimization: Optimality conditions and algorithms. arXiv:1203.4580 [cs.IT], Mar. 2012.

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.

D. Boas, D. Brooks, E. Miller, C. DiMarzio, M. Kilmer, R. Gaudette, and Q. Zhang. Imaging the body with diffuse optical tomography. *IEEE Signal Processing Magazine*, 18(6):57–75, Nov. 2001.

L. Borcea. Electrical impedance tomography. *Inverse Problems*, 18(6):R99–R136, Dec. 2002.

P. Boufounos. Greedy sparse signal reconstruction from sign measurements. In *Conference Record of the Forty-Third Asilomar Conference on Signals, Systems and Computers, 2009*, pages 1305 –1309, Nov. 2009.

P. Boufounos and R. Baraniuk. 1-bit compressive sensing. In *Information Sciences and Systems, 2008. CISS 2008. 42nd Annual Conference on*, pages 16 –21, Mar. 2008.

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 and X. Li. Solving quadratic equations via PhaseLift when there are about as many equations as unknowns. arXiv:1208.6247 [cs.IT], Aug. 2012.

E. J. Candès and T. Tao. Near optimal signal recovery from random projections: universal encoding strategies? *IEEE Transactions on Information Theory*, 52(12):5406–5425, Dec. 2006.

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.

E. J. Candès, T. Strohmer, and V. Voroninski. PhaseLift: exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 2012. doi: 10.1002/cpa.21432.

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.

R. Chartrand. Exact reconstruction of sparse signals via nonconvex minimization. *IEEE Signal Processing Letters*, 14(10):707 –710, Oct. 2007a.

R. Chartrand. Nonconvex compressed sensing and error correction. In *Proceedings of the 32nd IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 3, pages 889–892, Apr. 2007b.

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

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

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

W. Dai, H. V. Pham, and O. Milenkovic. Distortion-rate functions for quantized compressive sensing. In *IEEE Information Theory Workshop on Networking and Information Theory, 2009. ITW 2009*, pages 171 –175, June 2009.

M. Davenport and M. Wakin. Analysis of orthogonal matching pursuit using the restricted isometry property. *IEEE Transactions on Information Theory*, 56(9):4395 –4401, Sept. 2010.

A. J. Dobson and A. Barnett. *An Introduction to Generalized Linear Models*. Chapman and Hall/CRC, Boca Reaton, FL, 3rd edition, May 2008. ISBN 9781584889502.

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.

M. Duarte and Y. Eldar. Structured compressed sensing: From theory to applications. *IEEE Transactions on Signal Processing*, 59(9):4053–4085, Sept. 2011.

M. Figueiredo, R. Nowak, and S. Wright. Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. *IEEE Journal of Selected Topics in Signal Processing*, 1(4):586 –597, dec. 2007. ISSN 1932-4553. doi: 10.1109/JSTSP.2007.910281.

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, San Antonio, TX, 2012. Springer New York.

S. Foucart and M.-J. Lai. Sparsest solutions of underdetermined linear systems via ℓ_q-minimization for $0 < q \leq 1$. *Applied and Computational Harmonic Analysis*, 26(3):395–407, 2009.

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, Feb. 2010. Software available online at http://www-stat.stanford.edu/~tibs/glmnet-matlab/.

R. Gribonval and M. Nielsen. Highly sparse representations from dictionaries are unique and independent of the sparseness measure. *Applied and Computational Harmonic Analysis*, 22(3):335–355, 2007.

I. Guyon, S. Gunn, A. Ben Hur, and G. Dror. Result analysis of the NIPS 2003 feature selection challenge. In *Advances in Neural Information Processing Systems 17*, pages 545–552. 2004. URL http://clopinet.com/isabelle/Projects/NIPS2003/ggad-nips04.pdf.

E. Hale, W. Yin, and Y. Zhang. Fixed-point continuation for ℓ_1-minimization: methodology and convergence. *SIAM Journal on Optimization*, 19(3):1107–1130, 2008.

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

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

D. Hsu, S. Kakade, and T. Zhang. Tail inequalities for sums of random matrices that depend on the intrinsic dimension. *Electron. Commun. Probab.*, 17(14):1–13, 2012.

L. Jacob, G. Obozinski, and J. Vert. Group Lasso with overlap and graph Lasso. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 433–440, New York, NY, USA, 2009.

L. Jacques, D. Hammond, and J. Fadili. Dequantizing compressed sensing: When oversampling and non-gaussian constraints combine. *IEEE Transactions on Information Theory*, 57(1):559 –571, Jan. 2011.

L. Jacques, J. Laska, P. Boufounos, and R. Baraniuk. Robust 1-bit compressive sensing via binary stable embeddings of sparse vectors. *IEEE Transactions on Information Theory*, PP, 2013. doi: 10.1109/TIT.2012.2234823. arXiv:1104.3160 [cs.IT].

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*, volume 24, pages 1935–1943. 2011.

R. Jenatton, J. Audibert, and F. Bach. Structured variable selection with sparsity-inducing norms. *Journal of Machine Learning Research*, 12:2777–2824, Oct. 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 Workshop and Conference Proceedings*, pages 381–388, Sardinia, Italy, 2010.

V. Kolehmainen, M. Vauhkonen, J. Kaipio, and S. Arridge. Recovery of piecewise constant coefficients in optical diffusion tomography. *Optics Express*, 7(13):468–480, Dec. 2000.

A. Kyrillidis and V. Cevher. Combinatorial selection and least absolute shrinkage via the CLASH algorithm. arXiv:1203.2936 [cs.IT], Mar. 2012a.

A. Kyrillidis and V. Cevher. Sublinear time, approximate model-based sparse recovery for all. arXiv:1203.4746 [cs.IT], Mar. 2012b.

J. Laska, Z. Wen, W. Yin, and R. Baraniuk. Trust, but verify: Fast and accurate signal recovery from 1-bit compressive measurements. *IEEE Transactions on Signal Processing*, 59(11):5289 –5301, Nov. 2011a.

J. N. Laska, P. T. Boufounos, and R. G. Baraniuk. Finite range scalar quantization for compressive sensing. In *Proceedings of International Conferenece on Sampling Theory and Applications (SampTA)*, Toulouse, France, May 18-22 2009.

J. N. Laska, P. T. Boufounos, M. A. Davenport, and R. G. Baraniuk. Democracy in action: Quantization, saturation, and compressive sensing. *Applied and Computational Harmonic Analysis*, 31(3):429–443, Nov. 2011b.

C. Lazar, J. Taminau, S. Meganck, D. Steenhoff, A. Coletta, C. Molter, V. de Schaetzen, R. Duque, H. Bersini, and A. Nowe. A survey on filter techniques for feature selection in gene expression microarray analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9(4):1106 –1119, Aug. 2012.

X. Li and V. Voroninski. Sparse signal recovery from quadratic measurements via convex programming. arXiv:1209.4785 [cs.IT], Sept. 2012.

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, *the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15, pages 452–460, Ft. Lauderdale, FL, USA, 2011. JMLR W&CP.

A. Maleki and D. Donoho. Optimally tuned iterative reconstruction algorithms for compressed sensing. *Selected Topics in Signal Processing, IEEE Journal of*, 4(2):330 –341, Apr. 2010.

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*, volume 22, pages 1348–1356. 2009. long version available at arXiv:1010.2731v1 [math.ST].

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

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 27th Asilomar Conference on Signals, Systems and Computers*, volume 1, pages 40–44, Pacific Grove, CA, Nov. 1993.

Y. Plan and R. Vershynin. One-bit compressed sensing by linear programming. arXiv:1109.4299, Sept. 2011.

Y. Plan and R. Vershynin. Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *IEEE Transactions on Information Theory*, 59(1):482–494, Jan. 2013.

V. Roth and B. Fischer. The Group-Lasso for generalized linear models: Uniqueness of solutions and efficient algorithms. In *Proceedings of the 25th International Conference on Machine learning*, ICML '08, pages 848–855, New York, NY, USA, 2008.

R. Saab and Ö. Yilmaz. Sparse recovery by non-convex optimization - instance optimality. *Applied and Computational Harmonic Analysis*, 29(1):30–48, 2010.

R. Saab, R. Chartrand, and Ö. Yilmaz. Stable sparse approximations via nonconvex optimization. In *Proceedings of the 33rd IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3885–3888, Las Vegas, NV, Apr. 2008.

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.

Y. Shechtman, Y. C. Eldar, A. Szameit, and M. Segev. Sparsity based sub-wavelength imaging with partially incoherent light via quadratic compressed sensing. *Optics Express*, 19 (16):14807–14822, July 2011a.

Y. Shechtman, A. Szameit, E. Osherovic, E. Bullkich, H. Dana, S. Gazit, S. Shoham, M. Zibulevsky, I. Yavneh, E. B. Kley, Y. C. Eldar, O. Cohen, and M. Segev. Sparsity-based single-shot sub-wavelength coherent diffractive imaging. In *Frontiers in Optics*, OSA Technical Digest, page PDPA3. Optical Society of America, Oct. 2011b.

J. Sun and V. Goyal. Optimal quantization of random measurements in compressed sensing. In *IEEE International Symposium on Information Theory, 2009. ISIT 2009*, pages 6 –10, July 2009.

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*, volume 24, pages 882–890. 2011.

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

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. ISBN 978-0-471-03003-4.

Z. Wen, W. Yin, D. Goldfarb, and Y. Zhang. A fast algorithm for sparse reconstruction based on shrinkage, subspace optimization, and continuation. *SIAM Journal on Scientific Computing*, 32(4):1832–1857, 2010.

M. Yan, Y. Yang, and S. Osher. Robust 1-bit compressive sensing using adaptive outlier pursuit. *IEEE Transactions on Signal Processing*, 60(7):3868 –3875, July 2012.

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

A. Zymnis, S. Boyd, and E. Candès. Compressed sensing with quantized measurements. *IEEE Signal Processing Letters*, 17(2):149 –152, Feb. 2010.