Greedy Sparsity-Constrained Optimization

Sohail Bahmani*[1], Petros Boufounos†[2], and Bhiksha Raj*‡[3]
[1]sbahmani@andrew.cmu.edu [2]petrosb@merl.com [3]bhiksha@cs.cmu.edu
*Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213
†Mitsubishi Electric Research Laboratories, 201 Broadway, Cambridge, MA 02139
‡Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA 15213

***Abstract*—Finding optimal sparse solutions to estimation problems, particularly in underdetermined regimes has recently gained much attention. Most existing literature study linear models in which the squared error is used as the measure of discrepancy to be minimized. However, in many applications discrepancy is measured in more general forms such as log-likelihood. Regularization by ℓ_1-norm has been shown to induce sparse solutions, but their sparsity level can be merely suboptimal. In this paper we present a greedy algorithm, dubbed Gradient Support Pursuit (GraSP), for sparsity-constrained optimization. Quantifiable guarantees are provided for GraSP when cost functions have the "Stable Hessian Property".**

I. Introduction

Sparsity has emerged as a central topic of study in variety of fields that require high-dimensional data analysis. Sparsity of the parameters of interest allows techniques such as robust regression and hypothesis testing, model reduction and variable selection, and compressive signal acquisition to be feasible in underdetermined settings. Estimation of underlying sparse parameters in these techniques is often cast as an optimization problem. In particular, these optimization problems are studied thoroughly for the case of sparse linear regression in the field of Compressive Sensing (CS).

The majority of the CS reconstruction algorithms rely on the Restricted Isometry Property (RIP), a sufficient condition to guarantee the solution accuracy. A matrix $\mathbf{A}$ satisfies the δ_s-RIP of order s if

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

holds for all s-sparse vectors $\mathbf{x}$ [1]. Given an s-sparse parameter vector, $\mathbf{x}^\star$, measurements $\mathbf{y}=\mathbf{A}\mathbf{x}^\star+\mathbf{e}$, and error $\|\mathbf{e}\|_2 \leq \varepsilon$ then, if $\delta_{2s} < \sqrt{2}-1$, the solution to the convex program

$$\arg\min_{\widehat{\mathbf{x}}} \ \|\widehat{\mathbf{x}}\|_1 \quad \text{s.t} \ \|\mathbf{y}-\mathbf{A}\widehat{\mathbf{x}}\|_2 \leq \varepsilon, \tag{1}$$

known as Basis Pursuit Denoising (BPDN) [2], satisfies $\|\widehat{\mathbf{x}}-\mathbf{x}^\star\|_2 \leq C\varepsilon$ for some constant C [3]. Alternative algorithms collectively known as *Greedy Pursuits* such as Iterative Hard Thresholding (IHT) [4], Compressive Sampling Matching Pursuit (CoSaMP) [5], and Subspace Pursuit [6] also provide similar guarantees based on the RIP. These greedy algorithms attempt to approximate a solution to the sparsity-constrained least-squares optimization

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

by iterating between detection of the location of non-zero elements and estimation of their values.

While least-squares (quadratic) cost functions are encountered often in signal processing, they are not always the appropriate choice in a variety of fields and applications. For example, in statistics and machine learning, Generalized Linear Models (GLMs) are associated with some important non-quadratic cost functions, such as the logistic loss [7]. We therefore desire to extend the theory and algorithms for sparse optimization to include a wider range of cost functions.

In addition to established results for quadratic cost functions, it has been shown that ℓ_1-regularization also induces sparse minimizers for a variety of other cost functions, including those which belong to GLMs [8, 9] or the exponential family [10]. These work demonstrate that, under certain conditions, ℓ_1-regularization provide a solution that accurately estimates the minimizer of the risk, assuming this minimizer is indeed sparse. The main advantage of ℓ_1-regularized problems is that they are convex and tractable, as opposed to the ideal sparsity inducing formulation using ℓ_0-regularization that leads to an NP-hard problem.

Unfortunately, for non-quadratic cost functions, the conditions under which the equivalence of ℓ_0 and ℓ_1 regularization can be guaranteed are mostly unknown. Furthermore, in applications where precise control of the sparsity is critical, imposing explicit sparsity constraints is preferable to ℓ_1-regularization. Thus, several algorithms independent of ℓ_1-regularization have also been proposed in the context of sparsity-constrained optimization. For example, the algorithms in [11] can sparsify a dense minimizer while increasing the cost only slightly. In [12] a non-linear generalization of CS is examined, where the observations are the image of the parameters under a nonlinear operator corrupted by additive noise. Using the squared error to measure discrepancy, a generalization of the IHT algorithm is then shown to find accurate sparse solutions under the smoothness and strong

S.B. was partially supported by the Bertucci Graduate Fellowship.

P.B. is fully supported by Mitsubishi Electric Research Laboratories.

This work is 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.

convexity criteria imposed by the so-called Restricted Strong Convexity Property (RSCP).

In this paper we develop a framework for sparsity-constrained minimization for a broad class of cost functions. In particular we provide a greedy algorithm, the Gradient Support Pursuit (GraSP), which approximates the solution to sparsity-constrained minimization problems, providing explicit control of the solution sparsity. Our algorithm generalizes and is inspired by the CoSaMP algorithm. We also develop a sufficient condition on the function—the Stable Hessian Property (SHP)—used to guarantee accuracy of the GraSP approximation.Our work is similar in spirit and complements [12], but has significant differences. In particular, the form of the cost function we examine in this work is more general. Furthermore, the RSCP upper and lower bounds in [12] are global. Instead the SHP merely requires that only a ratio of upper and lower bounds is globally bounded, related to the conditioning of the (reduced) Hessian of the cost function being restricted to canonical sparse supspaces. Finally, our algorithm generalizes CoSaMP instead of the IHT. The former is a more sophisticated algorithm, that has demonstrated significantly better performance in simulations [13].

The next section introduces notation and the problem formulation. Section III presents GraSP, the SHP, and the implied performance guarantees. Section IV discusses our results and concludes. The proofs of our results are relegated to the appendix.

II. Problem Formulation

Notation: In the remainder of this paper we use boldface letters to denote matrices and vectors. For a positive integer m we use $[m]$ as a shorthand for the set $\{1, 2, \cdots, m\}$. Suppose that $\mathbf{M}$ is an $a \times b$ matrix, $\mathbf{v}$ is a b-dimensional vector, and $\mathcal{J}$ is a subset of $[b]$ for arbitrary positive integers a and b. The set of nonzero entries, the support set, of $\mathbf{v}$ is denoted by supp($\mathbf{v}$). We use $\mathbf{M}^{\mathrm{T}}$ and $\mathbf{M}^{\dagger}$ to denote the transpose and pseudo-inverse of $\mathbf{M}$, respectively. The largest and the smallest eigenvalues of $\mathbf{M}$ are denoted by $\lambda_{\max}(\mathbf{M})$ and $\lambda_{\min}(\mathbf{M})$, respectively. Furthermore, $\mathbf{M}_{\mathcal{J}}$ denotes the restriction of $\mathbf{M}$ to the columns enumerated by $\mathcal{J}$. Similarly, $\mathbf{v}|_{\mathcal{J}}$ is the restriction of $\mathbf{v}$ to the rows indicated by $\mathcal{J}$. The best r-term approximation of $\mathbf{v}$ is denoted by $\mathbf{v}_r$. The support set of $\mathbf{v}$, i.e., indices of the non-zero entries, is denoted by $\mathrm{supp}\,(\mathbf{v})$. Finally, $\mathbf{P}_{\mathcal{J}}$ denotes restriction of the identity matrix (i.e., $\mathbf{I}_{b\times b}$) to the rows indicated by $\mathcal{J}$ (i.e., $\mathbf{P}_{\mathcal{J}} = \mathbf{I}_{\mathcal{J}}^{\mathrm{T}}$).

Sparsity-constrained minimization: We generalize (2) with a generic cost function replacing the squared error. Using $f(\mathbf{x})$ to denote the cost function, we attempt to approximate a solution to

$$\arg\min_{\widehat{\mathbf{x}}} f(\widehat{\mathbf{x}}) \qquad \text{s.t. } \|\widehat{\mathbf{x}}\|_0 \leq s. \tag{3}$$

To perform this minimization, we provide an algorithm, the Gradient Support Pursuit (GraSP), which is inspired by and generalizes the CoSaMP algorithm. Of course, even for a simple quadratic objective, (3) can have combinatorial complexity and become NP-hard. Thus, we also provide a sufficient condition, the Stable Hessian Property (SHP) that enables accurate and tractable approximation. The SHP is analogous to the RIP in the sense that in linear regression problems with squared error as the cost function, the SHP basically reduces to RIP.

III. Gradient Support Pursuit (GraSP) Algorithm

Algorithm Description: 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. In each iteration, first the gradient of the cost function is evaluated at the current estimate to obtain $\mathbf{z} = \nabla f(\widehat{\mathbf{x}})$. Then indices of $2s$ entries of $\mathbf{z}$ with largest magnitudes are collected in the set Ω to indicate the coordinates in which estimation error is dominant. In the next step Ω is merged with the support of the current estimate to obtain $\mathcal{T}$ which contains at most $3s$ coordinates. The function f is then minimized over the vectors supported on $\mathcal{T}$ to produce a crude estimate $\mathbf{b}$. As will be seen later, by imposing the SHP this inner optimization step becomes a convex program which can be solved efficiently. Finally, the estimate $\widehat{\mathbf{x}}$ is updated to the best s-term approximation of $\mathbf{b}$. The iterations continue until a terminating condition, e.g., on the change of the cost function or the change of the estimated minimum from the previous iteration holds.

Using the quadratic cost $f(\widehat{\mathbf{x}}) = \frac{1}{2}\|\mathbf{y} - \mathbf{A}\widehat{\mathbf{x}}\|_2^2$, GraSP reduces to CoSaMP. Specifically, the gradient step and the support-constrained minimization reduce to the proxy step $\mathbf{z} = \mathbf{A}^{\mathrm{T}}(\mathbf{y} - \mathbf{A}\hat{\mathbf{x}})$ and the constrained pseudoinverse step $\mathbf{b}|_{\mathcal{T}} = \mathbf{A}_{\mathcal{T}}^{\dagger}\mathbf{y}, \mathbf{b}|_{\mathcal{T}^c} = \mathbf{0}$ in CoSaMP, respectively.

```
Algorithm 1: The GraSP algorithm
input : $f(\cdot)$ and $s$
output: $\widehat{\mathbf{x}}$
initialize:                 $\widehat{\mathbf{x}} = \mathbf{0}$
repeat
    compute local gradient:  $\mathbf{z} = \nabla f(\widehat{\mathbf{x}})$
    identify directions:     $\Omega = \mathrm{supp}\,(\mathbf{z}_{2s})$
    merge supports:          $\mathcal{T} = \Omega \cup \mathrm{supp}\,(\widehat{\mathbf{x}})$
    minimize over
    support:    $\mathbf{b} = \arg\min f(\mathbf{x})$    s.t. $\mathbf{x}|_{\mathcal{T}^c} = \mathbf{0}$
    prune estimate:          $\widehat{\mathbf{x}} = \mathbf{b}_s$
until terminating condition holds
```

Sparse Reconstruction Conditions: To characterize and provide performance guarantees for the algorithm, we first introduce the Stable Hessian Property (SHP), a sufficient condition on the set of functions that GraSP can minimize.

Definition 1. Suppose that f is a twice continuously differentiable function whose Hessian is denoted by $\mathbf{H}_f(\cdot)$. Furthermore, for a given positive integer k let

$$A_k(\mathbf{u}) = \sup_{\substack{|\mathrm{supp}(\mathbf{u})\cup\mathrm{supp}(\mathbf{v})|\leq k\\ \|\mathbf{v}\|_2=1}} \mathbf{v}^{\mathrm{T}}\mathbf{H}_f(\mathbf{u})\,\mathbf{v} \tag{4}$$

and

$$B_k\left(\mathbf{u}\right)=\inf_{\substack{|\mathrm{supp}(\mathbf{u})\cup\mathrm{supp}(\mathbf{v})|\leq k\\ \|\mathbf{v}\|_2=1}}\mathbf{v}^{\mathrm{T}}\mathbf{H}_f\left(\mathbf{u}\right)\mathbf{v},\tag{5}$$

for all k-sparse vectors $\mathbf{u}$. Then f is said to have the Stable Hessian Property (SHP) with constant μ_k, or in short μ_k-SHP, if $\frac{A_k(\mathbf{u})}{B_k(\mathbf{u})}\leq\mu_k$.

Remark. Note that the SHP only requires that symmetrically selected submatrices of the Hessian to be well-conditioned and in general the Hessian does not have to be positive-semidefinite. Furthermore, there is no global bound on $A_k\left(\mathbf{u}\right)$ and $B_k\left(\mathbf{u}\right)$, only on their ratio. Thus they can be arbitrarily large or small, as long as their ratio is controlled. For the special case of quadratic cost functions as in (2), we can write $\mathbf{H}_f\left(\mathbf{u}\right)=\mathbf{A}^{\mathrm{T}}\mathbf{A}$ which is constant. The SHP condition then implies $B_k\|\mathbf{v}\|_2^2\leq\|\mathbf{A}\mathbf{v}\|_2^2\leq A_k\|\mathbf{v}\|_2^2$ for all k-sparse vectors $\mathbf{v}$ and some positive constants A_k and B_k. Therefore, in this case a matrix with small RIP constant will also have small SHP constant and the conditions become essentially equivalent.

Performance Guarantees: The following theorem shows that if f satisfies the μ_{4s}-SHP with $\mu_{4s}\leq\sqrt{3/2}$ then GraSP finds an accurate estimate of

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

Theorem 1. *Suppose that f has μ_{4s}-SHP with $\mu_{4s}\leq\sqrt{3/2}$. Furthermore, suppose that for some $\epsilon>0$ we have $\epsilon<B_{4s}\left(\mathbf{u}\right)$ for all $4s$-sparse $\mathbf{u}$. Then $\widehat{\mathbf{x}}^{(i)}$, the estimate at the i-th iteration, satisfies*

$$\|\widehat{\mathbf{x}}^{(i)}-\mathbf{x}^\star\|_2\leq2^{-i}\|\mathbf{x}^\star\|_2+\frac{4\left(2+\sqrt{3/2}\right)}{\epsilon}\|\nabla f\left(\mathbf{x}^\star\right)|_{\mathcal{I}}\|_2,$$

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

Remark. Note that the the condition $\mu_{4s}\leq\sqrt{3/2}$ is imposed merely to have a contraction factor of $1/2$. In fact having $\mu_{4s}<\sqrt{2}$ would be sufficient to have a valid contraction factor. Furthermore, Theorem 1 indicates that $\nabla f\left(\mathbf{x}^\star\right)$ controls the accuracy of the output of GraSP. In particular, if the sparse minimum $\mathbf{x}^\star$ is also an unconstrained local minimum of f then $\nabla f\left(\mathbf{x}^\star\right)=0$ and the error floor vanishes. This result is similar to the s-term approximation guarantees in CS in case of nearly sparse signals or noisy measurements [1, 5].

IV. Discussion

In this paper we introduce the Gradient Support Pursuit (GraSP) algorithm to solve a wide range of sparsity-constrained optimization problems. We also propose the Stable Hessian Property (SHP), which allows us to provide theoretical guarantees on accuracy of the solution obtained by GraSP.

In contrast with ℓ_1-regularization techniques GraSP allows direct control of the sparsity of the solution which is critical in applications such as feature selection. The error bounds obtained in statistical estimation problems in general are not absolute constants and they generally depend on the true statistical optimum. Therefore, at large error bounds ℓ_1-regularization might not guarantee sufficiently sparse solutions. Furthermore, since GraSP operates on a small subset of coordinates in each iteration it provides more computational flexibility compared to standard ℓ_1-regularization techniques. Our results also show that if the SHP condition holds with proper constant, the algorithm shows a linear rate of convergence up to an approximation error.

Studying GraSP in statistical estimation framework, relaxing the SHP to an entirely local condition, and extending the results to nonsmooth cost functions, are interesting problems that we are investigating as parts of our future work.

Appendix

We first provide a few propositions and lemmas to analyze how the algorithm operates on its current estimate $\widehat{\mathbf{x}}$. These results lead to an iteration invariant property on the estimation error which is the basis for proving Theorem 1. Due to length limitations we omit the proofs of Propositions 1 and 2.

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

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

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

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

To simplify notation we use (4) and (5) and introduce functions

$$\alpha_k\left(\mathbf{p},\mathbf{q}\right)=\int_0^1A_k\left(t\mathbf{q}+\left(1-t\right)\mathbf{p}\right)dt$$

and

$$\beta_k\left(\mathbf{p},\mathbf{q}\right)=\int_0^1B_k\left(t\mathbf{q}+\left(1-t\right)\mathbf{p}\right)dt.$$

We also define $\gamma_k\left(\mathbf{p},\mathbf{q}\right):=\alpha_k\left(\mathbf{p},\mathbf{q}\right)-\beta_k\left(\mathbf{p},\mathbf{q}\right)$. Furthermore, we use the shorthand $\mathbf{z}^\star=\nabla f\left(\mathbf{x}^\star\right)$.

Lemma 1. *Let $\widehat{\boldsymbol{\Delta}}=\widehat{\mathbf{x}}-\mathbf{x}^\star$ and $\mathcal{R}=\mathrm{supp}\left(\widehat{\boldsymbol{\Delta}}\right)$. Then the current error vector $\widehat{\boldsymbol{\Delta}}$ obeys*

$$\begin{aligned}\|\widehat{\boldsymbol{\Delta}}|_{\Omega^c}\|_2\leq&\frac{\gamma_{4s}\left(\widehat{\mathbf{x}},\right)+\gamma_{2s}\left(\widehat{\mathbf{x}},\mathbf{x}^\star\right)}{2\beta_{2s}\left(\widehat{\mathbf{x}},\mathbf{x}^\star\right)}\|\widehat{\boldsymbol{\Delta}}\|_2\\&+\frac{\|\mathbf{z}^\star|_{\mathcal{R}\backslash\Omega}\|_2+\|\mathbf{z}^\star|_{\Omega\backslash\mathcal{R}}\|_2}{\beta_{2s}\left(\widehat{\mathbf{x}},\mathbf{x}^\star\right)}.\end{aligned}$$

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

$$\|\mathbf{z}|_{\mathcal{R}\backslash\Omega}\|_2 \le \|\mathbf{z}|_{\Omega\backslash\mathcal{R}}\|_2. \tag{6}$$

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

$$\begin{aligned}\|\mathbf{z}|_{\mathcal{R}\backslash\Omega}\|_2 \ge& \|\nabla f(\widehat{\mathbf{x}})|_{\mathcal{R}\backslash\Omega} - \nabla f(\mathbf{x}^\star)|_{\mathcal{R}\backslash\Omega}\|_2 - \|\nabla f(\mathbf{x}^\star)|_{\mathcal{R}\backslash\Omega}\|_2\\ =& \left\|\int_0^1 \mathbf{P}_{\mathcal{R}\backslash\Omega}\mathbf{H}_f\left(\mathbf{x}^\star + t\widehat{\boldsymbol{\Delta}}\right)dt\widehat{\boldsymbol{\Delta}}\right\|_2 - \|\mathbf{z}^\star|_{\mathcal{R}\backslash\Omega}\|_2\\ \ge& \left\|\int_0^1 \mathbf{P}_{\mathcal{R}\backslash\Omega}\mathbf{H}_f\left(\mathbf{x}^\star + t\widehat{\boldsymbol{\Delta}}\right)\mathbf{P}^{\mathrm{T}}_{\mathcal{R}\backslash\Omega}dt\widehat{\boldsymbol{\Delta}}|_{\mathcal{R}\backslash\Omega}\right\|_2\\ &- \left\|\int_0^1 \mathbf{P}_{\mathcal{R}\backslash\Omega}\mathbf{H}_f\left(\mathbf{x}^\star + t\widehat{\boldsymbol{\Delta}}\right)\mathbf{P}^{\mathrm{T}}_{\Omega\cap\mathcal{R}}dt\widehat{\boldsymbol{\Delta}}|_{\Omega\cap\mathcal{R}}\right\|_2\\ &- \|\mathbf{z}^\star|_{\mathcal{R}\backslash\Omega}\|_2\end{aligned}$$

Since $\|\widehat{\boldsymbol{\Delta}}|_{\Omega\cap\mathcal{R}}\|_2 \le \|\widehat{\boldsymbol{\Delta}}|_2$ applying Propositions 1 and 2 yields

$$\begin{aligned}\|\mathbf{z}|_{\mathcal{R}\backslash\Omega}\|_2 \ge& \beta_{2s}(\widehat{\mathbf{x}}, \mathbf{x}^\star)\|\widehat{\boldsymbol{\Delta}}|_{\mathcal{R}\backslash\Omega}\|_2\\ &- \frac{\gamma_{2s}(\widehat{\mathbf{x}}, \mathbf{x}^\star)}{2}\|\widehat{\boldsymbol{\Delta}}\|_2 - \|\mathbf{z}^\star|_{\mathcal{R}\backslash\Omega}\|_2.\end{aligned} \tag{7}$$

Similarly, we have

$$\begin{aligned}\|\mathbf{z}|_{\Omega\backslash\mathcal{R}}\|_2 \le& \|\nabla f(\widehat{\mathbf{x}})|_{\Omega\backslash\mathcal{R}} - \nabla f(\mathbf{x}^\star)|_{\Omega\backslash\mathcal{R}}\|_2 + \|\mathbf{z}^\star|_{\Omega\backslash\mathcal{R}}\|_2\\ =& \left\|\int_0^1 \mathbf{P}_{\Omega\backslash\mathcal{R}}\mathbf{H}_f\left(\mathbf{x}^\star + t\widehat{\boldsymbol{\Delta}}\right)\mathbf{P}^{\mathrm{T}}_{\mathcal{R}}dt\widehat{\boldsymbol{\Delta}}|_{\mathcal{R}}\right\|_2 + \|\mathbf{z}^\star|_{\Omega\backslash\mathcal{R}}\|_2\\ \le& \frac{\gamma_{4s}(\widehat{\mathbf{x}}, \mathbf{x}^\star)}{2}\|\widehat{\boldsymbol{\Delta}}\|_2 + \|\mathbf{z}^\star|_{\Omega\backslash\mathcal{R}}\|_2.\end{aligned} \tag{8}$$

Combining (6), (7), and (8) we obtain

$$\begin{aligned}\beta_{2s}(\widehat{\mathbf{x}}, \mathbf{x}^\star)\|\widehat{\boldsymbol{\Delta}}|_{\mathcal{R}\backslash\Omega}\|_2 - \frac{\gamma_{2s}(\widehat{\mathbf{x}}, \mathbf{x}^\star)}{2}\|\widehat{\boldsymbol{\Delta}}\|_2 - \|\mathbf{z}^\star|_{\mathcal{R}\backslash\Omega}\|_2\\ \le \frac{\gamma_{4s}(\widehat{\mathbf{x}}, \mathbf{x}^\star)}{2}\|\widehat{\boldsymbol{\Delta}}\|_2 + \|\mathbf{z}^\star|_{\Omega\backslash\mathcal{R}}\|_2.\end{aligned}$$

Since $\mathcal{R} = \operatorname{supp}\left(\widehat{\boldsymbol{\Delta}}\right)$, we have $\|\widehat{\boldsymbol{\Delta}}|_{\mathcal{R}\backslash\Omega}\|_2 = \|\widehat{\boldsymbol{\Delta}}|_{\Omega^c}\|_2$. Hence,

$$\begin{aligned}\|\widehat{\boldsymbol{\Delta}}|_{\Omega^c}\|_2 \le& \frac{\gamma_{4s}(\widehat{\mathbf{x}}, \mathbf{x}^\star) + \gamma_{2s}(\widehat{\mathbf{x}}, \mathbf{x}^\star)}{2\beta_{2s}(\widehat{\mathbf{x}}, \mathbf{x}^\star)}\|\widehat{\boldsymbol{\Delta}}\|_2\\ &+ \frac{\|\mathbf{z}^\star|_{\mathcal{R}\backslash\Omega}\|_2 + \|\mathbf{z}^\star|_{\Omega\backslash\mathcal{R}}\|_2}{\beta_{2s}(\widehat{\mathbf{x}}, \mathbf{x}^\star)}.\end{aligned}$$

■

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

$$\mathbf{b} = \arg\min\ f(\mathbf{x})\ \text{ s.t. } \mathbf{x}|_{\mathcal{T}^c} = \mathbf{0} \tag{9}$$

let $\widetilde{\boldsymbol{\Delta}} = \mathbf{b} - \mathbf{x}^\star$. *Then we have*

$$\|\widetilde{\boldsymbol{\Delta}}\|_2 \le \frac{\|\mathbf{z}^\star|_{\mathcal{T}}\|_2}{\beta_{3s}(\mathbf{b}, \mathbf{x}^\star)} + \left(1 + \frac{\gamma_{4s}(\mathbf{b}, \mathbf{x}^\star)}{2\beta_{3s}(\mathbf{b}, \mathbf{x}^\star)}\right)\|\mathbf{x}^\star|_{\mathcal{T}^c}\|_2.$$

Proof: By definition $\mathbf{z}^\star = \nabla f(\mathbf{x}^\star)$ thus we have

$$\mathbf{z}^\star - \nabla f(\mathbf{b}) = -\int_0^1 \mathbf{H}_f\left(\mathbf{x}^\star + t\widetilde{\boldsymbol{\Delta}}\right)dt\,\widetilde{\boldsymbol{\Delta}}.$$

Furthermore, since $\mathbf{b}$ is the solution to (9) we must have $\nabla f(\mathbf{b})|_{\mathcal{T}} = 0$. Therefore,

$$\begin{aligned}\mathbf{z}^\star|_{\mathcal{T}} =& -\int_0^1 \mathbf{P}_{\mathcal{T}}\mathbf{H}_f\left(\mathbf{x}^\star + t\widetilde{\boldsymbol{\Delta}}\right)dt\,\widetilde{\boldsymbol{\Delta}}\\ =& -\int_0^1 \mathbf{P}_{\mathcal{T}}\mathbf{H}_f\left(\mathbf{x}^\star + t\widetilde{\boldsymbol{\Delta}}\right)\mathbf{P}^{\mathrm{T}}_{\mathcal{T}}dt\,\widetilde{\boldsymbol{\Delta}}|_{\mathcal{T}}\\ &- \int_0^1 \mathbf{P}_{\mathcal{T}}\mathbf{H}_f\left(\mathbf{x}^\star + t\widetilde{\boldsymbol{\Delta}}\right)\mathbf{P}^{\mathrm{T}}_{\mathcal{T}^c}dt\,\widetilde{\boldsymbol{\Delta}}|_{\mathcal{T}^c}.\end{aligned} \tag{10}$$

Since f has μ_{4s}-SHP and $|\mathcal{T}| \le 3s$, functions $A_{3s}(\cdot)$ and $B_{3s}(\cdot)$, defined using (4) and (5), exist such that for all $t \in [0, 1]$ we have

$$\begin{aligned}B_{3s}\left(\mathbf{x}^\star + t\widetilde{\boldsymbol{\Delta}}\right) \le& \lambda_{\min}\left(\mathbf{P}_{\mathcal{T}}\mathbf{H}_f\left(\mathbf{x}^\star + t\widetilde{\boldsymbol{\Delta}}\right)\mathbf{P}^{\mathrm{T}}_{\mathcal{T}}\right)\\ \le& \lambda_{\max}\left(\mathbf{P}_{\mathcal{T}}\mathbf{H}_f\left(\mathbf{x}^\star + t\widetilde{\boldsymbol{\Delta}}\right)\mathbf{P}^{\mathrm{T}}_{\mathcal{T}}\right) \le A_{3s}\left(\mathbf{x}^\star + t\widetilde{\boldsymbol{\Delta}}\right).\end{aligned}$$

If $\mathbf{W}$ denotes the matrix $\int_0^1 \mathbf{P}_{\mathcal{T}}\mathbf{H}_f\left(\mathbf{x}^\star + t\widetilde{\boldsymbol{\Delta}}\right)\mathbf{P}^{\mathrm{T}}_{\mathcal{T}}dt$ it follows from Proposition 1 that

$$\beta_{3s}(\mathbf{b}, \mathbf{x}^\star) \le \lambda_{\min}(\mathbf{W}) \le \lambda_{\max}(\mathbf{W}) \le \alpha_{3s}(\mathbf{b}, \mathbf{x}^\star).$$

Consequently $\mathbf{W}$ is invertible and

$$\frac{1}{\alpha_{3s}(\mathbf{b}, \mathbf{x}^\star)} \le \lambda_{\min}\left(\mathbf{W}^{-1}\right) \le \lambda_{\max}\left(\mathbf{W}^{-1}\right) \le \frac{1}{\beta_{3s}(\mathbf{b}, \mathbf{x}^\star)}. \tag{11}$$

Therefore, by multiplying (10) by $\mathbf{W}^{-1}$ and using the fact that $\widetilde{\boldsymbol{\Delta}}|_{\mathcal{T}^c} = -\mathbf{x}^\star|_{\mathcal{T}^c}$ we obtain

$$\mathbf{W}^{-1}\mathbf{z}^\star|_{\mathcal{T}} = -\widetilde{\boldsymbol{\Delta}}|_{\mathcal{T}} + \mathbf{W}^{-1}\int_0^1 \mathbf{P}_{\mathcal{T}}\mathbf{H}_f\left(\mathbf{x}^\star + t\widetilde{\boldsymbol{\Delta}}\right)\mathbf{P}^{\mathrm{T}}_{\mathcal{T}^c}dt\,\mathbf{x}^\star|_{\mathcal{T}^c}.$$

Triangle inequality, (11), and Proposition 2 then yield

$$\begin{aligned}\|\widetilde{\boldsymbol{\Delta}}|_{\mathcal{T}}\|_2 \le& \|\mathbf{W}^{-1}\mathbf{z}^\star|_{\mathcal{T}}\|_2\\ &+ \left\|\mathbf{W}^{-1}\int_0^1 \mathbf{P}_{\mathcal{T}}\mathbf{H}_f\left(\mathbf{x}^\star + t\widetilde{\boldsymbol{\Delta}}\right)\mathbf{P}^{\mathrm{T}}_{\mathcal{T}^c\cap\mathcal{S}^\star}dt\mathbf{x}^\star|_{\mathcal{T}^c\cap\mathcal{S}^\star}\right\|_2\\ \le& \frac{\|\mathbf{z}^\star|_{\mathcal{T}}\|_2}{\beta_{3s}(\mathbf{b}, \mathbf{x}^\star)} + \frac{\gamma_{4s}(\mathbf{b}, \mathbf{x}^\star)}{2\beta_{3s}(\mathbf{b}, \mathbf{x}^\star)}\|\mathbf{x}^\star|_{\mathcal{T}^c}\|_2,\end{aligned}$$

where $\mathcal{S}^\star = \operatorname{supp}(\mathbf{x}^\star)$. Finally, we obtain

$$\begin{aligned}\|\widetilde{\boldsymbol{\Delta}}\|_2 \le& \|\mathbf{x}^\star|_{\mathcal{T}^c}\|_2 + \|\widetilde{\boldsymbol{\Delta}}|_{\mathcal{T}}\|_2\\ \le& \frac{\|\mathbf{z}^\star|_{\mathcal{T}}\|_2}{\beta_{3s}(\mathbf{b}, \mathbf{x}^\star)} + \left(1 + \frac{\gamma_{4s}(\mathbf{b}, \mathbf{x}^\star)}{2\beta_{3s}(\mathbf{b}, \mathbf{x}^\star)}\right)\|\mathbf{x}^\star|_{\mathcal{T}^c}\|_2.\end{aligned}$$

■

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

$$\begin{aligned}\|\mathbf{b}_s-\mathbf{x}^\star\|_2 \leq &\frac{\gamma_{4s}(\widehat{\mathbf{x}},\mathbf{x}^\star)+\gamma_{2s}(\widehat{\mathbf{x}},\mathbf{x}^\star)}{\beta_{2s}(\widehat{\mathbf{x}},\mathbf{x}^\star)}\left[1+\frac{\gamma_{4s}\left(\mathbf{b},\mathbf{x}^\star\right)}{2\beta_{3s}\left(\mathbf{b},\mathbf{x}^\star\right)}\right]\|\widehat{\boldsymbol{\Delta}}\|_2\\ &+\left[2+\frac{\gamma_{4s}(\mathbf{b},\mathbf{x}^\star)}{\beta_{3s}(\mathbf{b},\mathbf{x}^\star)}\right]\frac{\|\mathbf{z}^\star|_{\mathcal{R}\backslash\Omega}\|_2+\|\mathbf{z}^\star|_{\Omega\backslash\mathcal{R}}\|_2}{\beta_{2s}(\widehat{\mathbf{x}},\mathbf{x}^\star)}+\frac{2\|\mathbf{z}^\star|_{\mathcal{T}}\|_2}{\beta_{3s}(\mathbf{b},\mathbf{x}^\star)}.\end{aligned}$$

Proof: Since $\Omega \subseteq \mathcal{T}$ we have $\mathcal{T}^c \subseteq \Omega^c$. Thus, $\|\mathbf{x}^\star|_{\mathcal{T}^c}\|_2 = \|\widehat{\boldsymbol{\Delta}}|_{\mathcal{T}^c}\|_2 \leq \|\widehat{\boldsymbol{\Delta}}|_{\Omega^c}\|_2$. Then using Lemma 1 we obtain

$$\begin{aligned}\|\mathbf{x}^\star|_{\mathcal{T}^c}\|_2 \leq &\frac{\gamma_{4s}\left(\widehat{\mathbf{x}},\mathbf{x}^\star\right)+\gamma_{2s}\left(\widehat{\mathbf{x}},\mathbf{x}^\star\right)}{2\beta_{2s}\left(\widehat{\mathbf{x}},\mathbf{x}^\star\right)}\|\widehat{\boldsymbol{\Delta}}\|_2\\ &+\frac{\|\mathbf{z}^\star|_{\mathcal{R}\backslash\Omega}\|_2+\|\mathbf{z}^\star|_{\Omega\backslash\mathcal{R}}\|_2}{\beta_{2s}\left(\widehat{\mathbf{x}},\mathbf{x}^\star\right)}.\end{aligned}\tag{12}$$

Furthermore,

$$\|\mathbf{b}_s-\mathbf{x}^\star\|_2 \leq \|\mathbf{x}^\star-\mathbf{b}\|_2+\|\mathbf{b}_s-\mathbf{b}\|_2 \leq 2\|\mathbf{x}^\star-\mathbf{b}\|_2 = 2\|\widetilde{\boldsymbol{\Delta}}\|_2$$

because $\mathbf{x}^\star$ is s-sparse and $\mathbf{b}_s$ is the best s-term approximation of $\mathbf{b}$. Therefore, using Lemma 2,

$$\|\mathbf{b}_s-\mathbf{x}^\star\|_2 \leq \frac{2\|\mathbf{z}^\star|_{\mathcal{T}}\|_2}{\beta_{3s}\left(\mathbf{b},\mathbf{x}^\star\right)}+\left(2+\frac{\gamma_{4s}\left(\mathbf{b},\mathbf{x}^\star\right)}{\beta_{3s}\left(\mathbf{b},\mathbf{x}^\star\right)}\right)\|\mathbf{x}^\star|_{\mathcal{T}^c}\|_2.\tag{13}$$

Combining (12) and (13) we obtain

$$\begin{aligned}\|\mathbf{b}_s-\mathbf{x}^\star\|_2 \leq &\frac{\gamma_{4s}\left(\widehat{\mathbf{x}},\mathbf{x}^\star\right)+\gamma_{2s}\left(\widehat{\mathbf{x}},\mathbf{x}^\star\right)}{\beta_{2s}\left(\widehat{\mathbf{x}},\mathbf{x}^\star\right)}\left[1+\frac{\gamma_{4s}\left(\mathbf{b},\mathbf{x}^\star\right)}{2\beta_{3s}\left(\mathbf{b},\mathbf{x}^\star\right)}\right]\|\widehat{\boldsymbol{\Delta}}\|_2\\ &+2\left(1+\frac{\gamma_{4s}\left(\mathbf{b},\mathbf{x}^\star\right)}{2\beta_{3s}\left(\mathbf{b},\mathbf{x}^\star\right)}\right)\frac{\|\mathbf{z}^\star|_{\mathcal{R}\backslash\Omega}\|_2+\|\mathbf{z}^\star|_{\Omega\backslash\mathcal{R}}\|_2}{\beta_{2s}\left(\widehat{\mathbf{x}},\mathbf{x}^\star\right)}\\ &+2\frac{\|\mathbf{z}^\star|_{\mathcal{T}}\|_2}{\beta_{3s}\left(\mathbf{b},\mathbf{x}^\star\right)}.\end{aligned}$$

■

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

$$\begin{aligned}\frac{\alpha_k\left(\mathbf{p},\mathbf{q}\right)}{\beta_k\left(\mathbf{p},\mathbf{q}\right)} &= \frac{\int_0^1 A_k\left(t\mathbf{q}+\left(1-t\right)\mathbf{p}\right)dt}{\int_0^1 B_k\left(t\mathbf{q}+\left(1-t\right)\mathbf{p}\right)dt}\\ &\leq \frac{\int_0^1 \mu_k B_k\left(t\mathbf{q}+\left(1-t\right)\mathbf{p}\right)dt}{\int_0^1 B_k\left(t\mathbf{q}+\left(1-t\right)\mathbf{p}\right)dt} = \mu_k,\end{aligned}$$

and thereby $\frac{\gamma_k(\mathbf{p},\mathbf{q})}{\beta_k(\mathbf{p},\mathbf{q})} \leq \mu_k - 1$. If $\widehat{\boldsymbol{\Delta}}^{(i)} = \widehat{\mathbf{x}}^{(i)} - \mathbf{x}^\star$ denotes the error vector in the i-th iteration of the algorithm then it follows from Lemma 3 that the estimation error obeys

$$\begin{aligned}\|\widehat{\boldsymbol{\Delta}}^{(i)}\|_2 \leq &2\left(\mu_{4s}-1\right)\left(1+\frac{\mu_{4s}-1}{2}\right)\|\widehat{\boldsymbol{\Delta}}^{(i-1)}\|_2\\ &+2\left(1+\frac{\mu_{4s}-1}{2}\right)\frac{\|\mathbf{z}^\star|_{\mathcal{R}\backslash\Omega}\|_2+\|\mathbf{z}^\star|_{\Omega\backslash\mathcal{R}}\|_2}{\beta_{2s}\left(\widehat{\mathbf{x}}^{(i-1)},\mathbf{x}^\star\right)}\\ &+2\frac{\|\mathbf{z}^\star|_{\mathcal{T}}\|_2}{\beta_{3s}\left(\mathbf{b}^{(i-1)},\mathbf{x}^\star\right)}\\ &\leq \left(\mu_{4s}^2-1\right)\|\widehat{\boldsymbol{\Delta}}^{(i-1)}\|_2+\frac{2\left(\mu_{4s}+2\right)}{\epsilon}\|\mathbf{z}^\star|_{\mathcal{I}}\|_2\end{aligned}$$

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

$$\|\widehat{\boldsymbol{\Delta}}^{(i)}\|_2 \leq \frac{\|\widehat{\boldsymbol{\Delta}}^{(i-1)}\|_2}{2}+\frac{2\left(2+\sqrt{3/2}\right)}{\epsilon}\|\mathbf{z}^\star|_{\mathcal{I}}\|_2.$$

The theorem follows by applying this inequality recursively and using the fact that $\|\widehat{\boldsymbol{\Delta}}^{(0)}\|_2 = \|\mathbf{x}^\star\|_2$.

References

[1] E. J. Candès, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on Pure and Applied Mathematics*, vol. 59, no. 8, pp. 1207–1223, 2006.

[2] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 33–61, 1998.

[3] E. J. Candès, "The restricted isometry property and its implications for compressed sensing," *Comptes Rendus Mathematique*, vol. 346, no. 9-10, pp. 589 – 592, 2008.

[4] T. Blumensath and M. E. Davies, "Iterative hard thresholding for compressed sensing," *Applied and Computational Harmonic Analysis*, vol. 27, pp. 265–274, Nov. 2009.

[5] D. Needell and J. A. Tropp, "CoSaMP: iterative signal recovery from incomplete and inaccurate samples," *Applied and Computational Harmonic Analysis*, vol. 26, no. 3, pp. 301–321, 2009.

[6] W. Dai and O. Milenkovic, "Subspace pursuit for compressive sensing signal reconstruction," *IEEE Transactions on Information Theory*, vol. 55, no. 5, pp. 2230–2249, 2009.

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

[8] S. Negahban, P. Ravikumar, M. Wainwright, and B. Yu, "A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers," in *Advances in Neural Information Processing Systems 22* (Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, eds.), pp. 1348–1356, 2009.

[9] S. Van De Geer, "High-dimensional generalized linear models and the lasso," *The Annals of Statistics*, vol. 36, no. 2, pp. 614–645, 2008.

[10] 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*, vol. 9 of *JMLR Workshop and Conference Proceedings*, pp. 381–388, 2010.

[11] S. Shalev-Shwartz, N. Srebro, and T. Zhang, "Trading accuracy for sparsity in optimization problems with sparsity constraints," *SIAM Journal on Optimization*, vol. 20, no. 6, pp. 2807–2832, 2010.

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

[13] A. Maleki and D. Donoho, "Optimally tuned iterative reconstruction algorithms for compressed sensing," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 4, pp. 330 –341, april 2010.