Smoothing and Decomposition for Analysis Sparse Recovery

Zhao Tan, *Student Member, IEEE*, Yonina C. Eldar, *Fellow, IEEE*, Amir Beck, and Arye Nehorai, *Fellow, IEEE*

***Abstract*—We consider algorithms and recovery guarantees for the analysis sparse model in which the signal is sparse with respect to a highly coherent frame. We consider the use of a monotone version of the fast iterative shrinkage-thresholding algorithm (MFISTA) to solve the analysis sparse recovery problem. Since the proximal operator in MFISTA does not have a closed-form solution for the analysis model, it cannot be applied directly. Instead, we examine two alternatives based on smoothing and decomposition transformations that relax the original sparse recovery problem, and then implement MFISTA on the relaxed formulation. We refer to these two methods as smoothing-based and decomposition-based MFISTA. We analyze the convergence of both algorithms and establish that smoothing-based MFISTA converges more rapidly when applied to general nonsmooth optimization problems. We then derive a performance bound on the reconstruction error using these techniques. The bound proves that our methods can recover a signal sparse in a redundant tight frame when the measurement matrix satisfies a properly adapted restricted isometry property. Numerical examples demonstrate the performance of our methods and show that smoothing-based MFISTA converges faster than the decomposition-based alternative in real applications, such as MRI image reconstruction.**

***Index Terms*—Analysis model, convergence analysis, fast iterative shrinkage-thresholding algorithm, restricted isometry property, smoothing and decomposition, sparse recovery.**

I. Introduction

LOW-DIMENSIONAL signal recovery exploits the fact that many natural signals are inherently low dimensional, although they may have high ambient dimension. Prior information about the low-dimensional space can be exploited to aid in recovery of the signal of interest. Sparsity is one of the popular forms of prior information, and is the prior that underlies the growing field of compressive sensing [1]–[4]. Recovery of sparse inputs has found many applications in areas such as imaging, speech, radar signal processing, sub-Nyquist sampling and more. A typical sparse recovery problem has the following linear form:

$$\boldsymbol{b} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{w}, \tag{1}$$

in which $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ is a measurement matrix, $\boldsymbol{b} \in \mathbb{R}^{m}$ is the measurement vector, and $\boldsymbol{w} \in \mathbb{R}^{m}$ represents the noise term. Our goal is to recover the signal $\boldsymbol{x} \in \mathbb{R}^{n}$. Normally we have $m < n$, which indicates that the inverse problem is ill-posed and has infinitely many solutions. To find a unique solution, prior information on $\boldsymbol{x}$ must be incorporated.

In the synthesis approach to sparse recovery, it is assumed that $\boldsymbol{x}$ can be expressed as a sparse combination of known dictionary elements, represented as columns of a matrix $\boldsymbol{D} \in \mathbb{R}^{n\times p}$ with $p \geq n$. That is $\boldsymbol{x} = \boldsymbol{D}\boldsymbol{\alpha}$ with $\boldsymbol{\alpha}$ sparse, i.e., the number of non-zero elements in $\boldsymbol{\alpha}$ is far less than the length of $\boldsymbol{\alpha}$. The main methods for solving this problem can be classified into two categories. One includes greedy methods, such as iterative hard thresholding [5] and orthogonal matching pursuit [6]. The other is based on relaxation-type methods, such as basis pursuit [7] and LASSO [8]. These methods can stably recover a sparse signal $\boldsymbol{\alpha}$ when the matrix $\boldsymbol{A}\boldsymbol{D}$ satisfies the restricted isometry property (RIP) [9]–[11].

Recently, an alternative approach has became popular, which is known as the analysis method [12], [13]. In this framework, we are given an analysis dictionary $\boldsymbol{D}^*(\boldsymbol{D} \in \mathbb{R}^{n\times p})$ under which $\boldsymbol{D}^*\boldsymbol{x}$ is sparse. Assuming, for example, that the ℓ_2 norm of the noise $\boldsymbol{w}$ is bounded by ε, the recovery problem can be formulated as

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \|\boldsymbol{D}^*\boldsymbol{x}\|_0 \quad \text{subject to } \|\boldsymbol{b}-\boldsymbol{A}\boldsymbol{x}\|_2 \leq \varepsilon. \tag{2}$$

Since this problem is NP hard, several greedy algorithms have been proposed to approximate it, such as thresholding [14] and subspace pursuit [15].

Alternatively, the nonconvex ℓ_0 norm can be approximated by the convex ℓ_1 norm leading to the following relaxed problem, referred to as analysis basis pursuit (ABP):

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \|\boldsymbol{D}^*\boldsymbol{x}\|_1 \quad \text{subject to } \|\boldsymbol{b}-\boldsymbol{A}\boldsymbol{x}\|_2 \leq \varepsilon. \tag{3}$$

ABP is equivalent to the unconstrained optimization

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \frac{1}{2}\|\boldsymbol{b}-\boldsymbol{A}\boldsymbol{x}\|_2^2 + \lambda\|\boldsymbol{D}^*\boldsymbol{x}\|_1, \tag{4}$$

Manuscript received August 22, 2013; revised December 03, 2013; accepted January 24, 2014. Date of publication February 06, 2014; date of current version March 10, 2014. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Antonio De Maio. The work of Z. Tan and A. Nehorai was supported by the AFOSR Grant FA9550-11-1-0210, NSF Grant CCF-1014908, and ONR Grant N000141310050. The work of Y. C. Eldar was supported in part by the Israel Science Foundation under Grant no. 170/10, in part by the Ollendorf Foundation, in part by a Magnet grant Metro450 from the Israel Ministry of Industry and Trade, and in part by the Intel Collaborative Research Institute for Computational Intelligence (ICRI-CI). The work of A. Beck was partially supported by the Israel Science Foundation under grant ISF No.253/12.

Z. Tan and A. Nehorai are with the Preston M. Green Department of Electrical and Systems Engineering Department, Washington University in St. Louis, St. Louis, MO, 63130 USA (e-mail: tanz@ese.wustl.edu; nehorai@ese.wustl.edu).

Y. C. Eldar is with the Department of Electrical Engineering, Technion—Israel Institute of Technology, Haifa 32000, Israel (e-mail: yonina@ee.technion.ac.il).

A. Beck is with the Department of Industrial Engineering and Management, Technion—Israel Institute of Technology, Haifa 32000, Israel (e-mail: becka@ie.technion.ac.il).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2014.2304932

which we call analysis LASSO (ALASSO). The equivalence is in the sense that for any $\varepsilon > 0$ there exists a λ for which the optimal solutions of ABP and ALASSO are identical.

Both optimization problems ABP and ALASSO can be solved using interior point methods [16]. However, when the problem dimension grows, these techniques become very slow since they require solutions of linear systems. Another suggested approach is based on alternating direction method of multipliers (ADMM) [17], [18]. The efficiency of this method highly depends on nice structure of the matrices $\boldsymbol{A}$. Fast versions of first-order algorithms, such as the fast iterative shrinkage-thresholding algorithm (FISTA) [19], are more favorable in dealing with large dimensional data since they do not require $\boldsymbol{A}$ to have any structure. The difficulty in directly applying first-order techniques to ABP (3) and ALASSO (4) is the fact that the nonsmooth term $\|\boldsymbol{D}^*\boldsymbol{x}\|_1$ is inseparable. A generalized iterative soft-thresholding algorithm was proposed in [20] to tackle this difficulty. However, this approach converges relatively slow as we will show in one of our numerical examples. A common alternative is to transform the nondifferentiable problem into a smooth counterpart. In [21], the authors used Nesterov's smoothing-based method [22] in conjunction with continuation (NESTA) to solve ABP (3), under the assumption that the matrix $\boldsymbol{A}^*\boldsymbol{A}$ is an orthogonal projector. In [23], a smoothed version of ALASSO (4) is solved using a nonlinear conjugate gradient descent algorithm. To avoid imposing conditions on $\boldsymbol{A}$, we focus in this paper on the ALASSO formulation (4).

It was shown in [24] that one can apply any fast first-order method that achieves an ε-optimal solution within $O\left(\frac{1}{\sqrt{\varepsilon}}\right)$ iterations, to an ε smooth-approximation of the general nonsmooth problem and obtain an algorithm with $O\left(\frac{1}{\varepsilon}\right)$ iterations. In this paper, we choose a monotone version of FISTA (MFISTA) [25] as our fast first-order method, whose objective function values are guaranteed to be non-increasing. We apply the smoothing approach together with MFISTA leading to the smoothing-based MFISTA (SFISTA) algorithm. We also propose a decomposition-based MFISTA method (DFISTA) to solve the analysis sparse recovery problem. The decomposition idea is to introduce an auxiliary variable $\boldsymbol{z}$ in (4) so that MFISTA can be applied in a simple and explicit manner. This decomposition approach can be traced back to [26], and has been widely used for solving total variation problems in the context of image reconstruction [27].

Both smoothing and decomposition based algorithms for nonsmooth optimization problems are very popular in the literature. One of the main goals of this paper is to examine their respective performance. We show that SFISTA requires lower computational complexity to reach a predetermined accuracy. Our results can be applied to a general model, and are not restricted to the analysis sparse recovery problem.

In the context of analysis sparse recovery, we show in Section II-C that both smoothing and decomposition techniques solve the following optimization problem:

$$\min_{\boldsymbol{x}\in\mathbb{R}^n,\boldsymbol{z}\in\mathbb{R}^p} \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|_2^2+\lambda\|\boldsymbol{z}\|_1+\frac{1}{2}\rho\|\boldsymbol{z}-\boldsymbol{D}^*\boldsymbol{x}\|_2^2, \qquad (5)$$

which we refer to as relaxed ALASSO (RALASSO). Another contribution of this paper is in proving recovery guarantees for RALASSO (5). With the introduction of the restricted isometry property adapted to $\boldsymbol{D}$ (D-RIP) [12], previous work [12], [28] studied recovery guarantees based on ABP (3) and ALASSO (4). Here we combine the techniques in [9] and [28], and obtain a performance bound on RALASSO (5). We show that when $\sigma_{2s} < 0.1907$ and $\|\boldsymbol{D}^*\boldsymbol{A}^*\boldsymbol{w}\|_\infty \le \frac{\lambda}{2}$, the solution $\hat{\boldsymbol{x}}_\rho$ of RALASSO (5) satisfies

$$\|\hat{\boldsymbol{x}}_\rho-\boldsymbol{x}\|_2 \le C_0\sqrt{s}\lambda+C_1\frac{\|\boldsymbol{D}^*\boldsymbol{x}-(\boldsymbol{D}^*\boldsymbol{x})_s\|_1}{\sqrt{s}}+C_2\frac{\lambda p}{\sqrt{s}\rho}, \quad (6)$$

where p is the number of rows in $\boldsymbol{D}^*$, C_0, C_1, C_2 are constants, and we use $(\boldsymbol{x})_s$ to denote the vector consisting of the largest s entries of $|\boldsymbol{x}|$. As a special case, choosing $\rho \to \infty$ extends the bound in (6) and obtains the reconstruction bound for ALASSO (4) as long as $\sigma_{2s} < 0.1907$, which improves upon the results of [28].

The paper is organized as follows. In Section II, we introduce some mathematical preliminaries, and present SFISTA and DFISTA for solving RALASSO (5). We analyze the convergence behavior of these two algorithms in Section III, and show that SFISTA converges faster than DFISTA for a general model. Performance guarantees on RALASSO (5) are developed in Section IV. Finally, in Section V we test our techniques on numerical experiments to demonstrate the effectiveness of our algorithms in solving the analysis recovery problem. We show that SFISTA performs favorably in comparison with DFISTA. A continuation method is also introduced to further accelerate the convergence speed.

Throughout the paper, we use capital italic bold letters to represent matrices and lowercase italic bold letters to represent vectors. For a given matrix $\boldsymbol{D}$, $\boldsymbol{D}^*$ denotes the conjugate matrix. We denote by $\boldsymbol{D}^*_{\mathcal{T}}$ the matrix that maintains the rows in $\boldsymbol{D}^*$ with indices in set $\mathcal{T}$, while setting all other rows to zero. Given a vector $\boldsymbol{x}$, $\|\boldsymbol{x}\|_1, \|\boldsymbol{x}\|_2$ are the ℓ_1, ℓ_2 norms respectively, $\|\boldsymbol{x}\|_0$ counts the number of nonzero components which will be referred to as the ℓ_0 norm although it is not a norm, and $\|\boldsymbol{x}\|_\infty$ denotes the maximum absolute value of the elements in $\boldsymbol{x}$. We use $\boldsymbol{x}[i]$ to represent the ith element of $\boldsymbol{x}$. For a matrix $\boldsymbol{A}$, $\|\boldsymbol{A}\|_2$ is the induced spectral norm, and $\|\boldsymbol{A}\|_{p,q} = \max \frac{\|\boldsymbol{A}\boldsymbol{x}\|_p}{\|\boldsymbol{x}\|_q}$. Finally, $\mathrm{Re}\langle\boldsymbol{a},\boldsymbol{b}\rangle = \frac{\langle\boldsymbol{a},\boldsymbol{b}\rangle+\langle\boldsymbol{b},\boldsymbol{a}\rangle}{2}$. We use $\mathrm{argmin}\{f(\boldsymbol{x}) : \boldsymbol{x} = \boldsymbol{z}, \boldsymbol{y}\}$ to denote $\boldsymbol{z}$ or $\boldsymbol{y}$, whichever yields a smaller function value of $f(\boldsymbol{x})$.

II. Smoothing and Decomposition for Analysis Sparse Recovery

In this section we present the smoothing-based and decomposition-based methods for solving ALASSO (4). To do so, we first recall in Section II-A some results related to proximal gradient methods that will be essential to our presentation and analysis.

A. The Proximal Gradient Method

We begin this section with the definition of Moreau's proximal (or "prox") operator [29], which is the key step in defining the proximal gradient method.

Given a closed proper convex function $h : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$, the proximal operator of h is defined by

$$\operatorname{prox}_h(\boldsymbol{x}) = \operatorname*{arg\,min}_{\boldsymbol{u}\in\mathbb{R}^n} \left\{ h(\boldsymbol{u}) + \frac{1}{2}\|\boldsymbol{u} - \boldsymbol{x}\|_2^2 \right\}. \tag{7}$$

The proximal operator can be computed efficiently in many important instances. For example, it can be easily obtained when h is an l_p norm ($p \in [1, \infty)$), or an indicator of "simple" closed convex sets such as the box, unit-simplex and the ball. More examples of proximal operators as well as a wealth of properties can be found, for example, in [30], [31].

The proximal operator can be used in order to compute smooth approximations of convex functions. Specifically, let h be a closed, proper, convex function, and let $\mu > 0$ be a given parameter. Define

$$h_\mu(\boldsymbol{x}) = \min_{\boldsymbol{u}\in\mathbb{R}^n} \left\{ h(\boldsymbol{u}) + \frac{1}{2\mu}\|\boldsymbol{u} - \boldsymbol{x}\|_2^2 \right\}. \tag{8}$$

It is easy to see that

$$h_\mu(\boldsymbol{x}) = h(\operatorname{prox}_{\mu h}(\boldsymbol{x})) + \frac{1}{2\mu}\|\boldsymbol{x} - \operatorname{prox}_{\mu h}(\boldsymbol{x})\|_2^2. \tag{9}$$

The function h_μ is called the *Moreau envelope of* h and has the following important properties (see [29] for further details):

- $h_\mu(\boldsymbol{x}) \le h(\boldsymbol{x})$.
- h_μ is continuously differentiable and its gradient is Lipschitz continuous with constant $1/\mu$.
- The gradient of h_μ is given by

$$\nabla h_\mu(\boldsymbol{x}) = \frac{1}{\mu}(\boldsymbol{x} - \operatorname{prox}_{\mu h}(\boldsymbol{x})). \tag{10}$$

One important usage of the proximal operator is in the proximal gradient method that is aimed at solving the following composite problem:

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \{F(\boldsymbol{x}) + G(\boldsymbol{x})\}. \tag{11}$$

Here $F : \mathbb{R}^n \to \mathbb{R}$ is a continuously differentiable convex function with a continuous gradient that has Lipschitz constant $L_{\nabla F}$:

$$\|\nabla F(\boldsymbol{x}) - \nabla F(\boldsymbol{y})\|_2 \le L_{\nabla F}\|\boldsymbol{x} - \boldsymbol{y}\|_2, \quad \text{for all } \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n,$$

and $G : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is an extended-valued, proper, closed and convex function. The *proximal gradient method* for solving (11) takes the following form (see [19], [32]):

Proximal Gradient Method For Solving (11)

```
Input: An upper bound $L \ge L_{\nabla F}$.
Step 0. Take $\boldsymbol{x}_0 \in \mathbb{R}^n$.
Step k. ($k \ge 1$)
    Compute $\boldsymbol{x}_k = \operatorname{prox}_{\frac{1}{L}G}\left(\boldsymbol{x}_{k-1} - \frac{1}{L}\nabla F(\boldsymbol{x}_{k-1})\right)$.
```

The main disadvantage of the proximal gradient method is that it suffers from a relatively slow $O(1/k)$ rate of convergence of the function values. An accelerated version is the *fast proximal gradient method*, also known in the literature as *fast iterative shrinkage thresholding algorithm* (FISTA) [19], [32]. When $G \equiv 0$, the problem is smooth, and FISTA coincides with Nesterov's optimal gradient method [33]. In this paper we implement a monotone version of FISTA (MFISTA) [25], which guarantees that the objective function value is non-increasing along the iterations.

Monotone FISTA Method (MFISTA) For Solving (11)

```
Input: An upper bound $L \ge L_{\nabla F}$.
Step 0. Take $\boldsymbol{y}_1 = \boldsymbol{x}_0, t_1 = 1$.
Step k. ($k \ge 1$) Compute
    $\boldsymbol{z}_k = \operatorname{prox}_{\frac{1}{L}G}\left(\boldsymbol{y}_k - \frac{1}{L}\nabla F(\boldsymbol{y}_k)\right)$.
    $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$.
    $\boldsymbol{x}_k = \operatorname{argmin}\{F(\boldsymbol{x}) + G(\boldsymbol{x}) : \boldsymbol{x} = \boldsymbol{z}_k, \boldsymbol{x}_{k-1}\}$.
    $\boldsymbol{y}_{k+1} = \boldsymbol{x}_k + \frac{t_k}{t_{k+1}}(\boldsymbol{z}_k - \boldsymbol{x}_k) + \frac{t_k - 1}{t_{k+1}}(\boldsymbol{x}_k - \boldsymbol{x}_{k-1})$.
```

The rate of convergence of the sequence generated by MFISTA is $O(1/k^2)$.

Theorem II.1: [25] Let $\{\boldsymbol{x}_k\}_{k\ge 0}$ *be the sequence generated by MFISTA, and let* $\hat{\boldsymbol{x}}$ *be an optimal solution of (11). Then*

$$F(\boldsymbol{x}_k) + G(\boldsymbol{x}_k) - F(\hat{\boldsymbol{x}}) - G(\hat{\boldsymbol{x}}) \le \frac{2L_{\nabla F}\|\boldsymbol{x}_0 - \hat{\boldsymbol{x}}\|_2^2}{(k+1)^2}. \tag{12}$$

B. The General Nonsmooth Model

The general optimization model we consider in this paper is

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \{H(\boldsymbol{x}) = f(\boldsymbol{x}) + g(\boldsymbol{D}^*\boldsymbol{x})\}, \tag{13}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a continuously differentiable convex function with a Lipschitz continuous gradient $L_{\nabla f}$. The function $g : \mathbb{R}^p \to \mathbb{R} \cup \{\infty\}$ is a closed, proper convex function which is not necessarily smooth, and $\boldsymbol{D}^* \in \mathbb{R}^{p\times n}$ is a given matrix. In addition, we assume that g is Lipschitz continuous with parameter L_g:

$$|g(\boldsymbol{z}) - g(\boldsymbol{v})| \le L_g\|\boldsymbol{z} - \boldsymbol{v}\|_2 \quad \text{for all } \boldsymbol{z}, \boldsymbol{v} \in \mathbb{R}^p.$$

This is equivalent to saying that the subgradients of g over $\mathbb{R}^p$ are bounded by L_g:

$$\|g'(\boldsymbol{z})\|_2 \le L_g \ \text{ for any } \boldsymbol{x} \in \mathbb{R}^n \ \text{ and } g'(\boldsymbol{z}) \in \partial g(\boldsymbol{z}).$$

An additional assumption we make throughout is that the proximal operator of $\alpha g(\boldsymbol{z})$ for any $\alpha > 0$ can be easily computed.

Directly applying MFISTA to (13) requires computing the proximal operator of $g(\boldsymbol{D}^*\boldsymbol{x})$. Despite the fact that we assume that it is easy to compute the proximal operator of $g(\boldsymbol{z})$, it is in general difficult to compute that of $\alpha g(\boldsymbol{D}^*\boldsymbol{x})$. Therefore we need to transform the problem before utilizing MFISTA, in order to avoid this computation.

When considering ALASSO, $f(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2$ and $g(\boldsymbol{D}^*\boldsymbol{x}) = \lambda\|\boldsymbol{D}^*\boldsymbol{x}\|_1$. The Lipschitz constants are given by $L_{\nabla f} = \|\boldsymbol{A}\|_2^2$ and $L_g = \lambda\sqrt{p}$. The proximal operator of $\alpha g(\boldsymbol{z}) = \alpha\lambda\|\boldsymbol{z}\|_1$ can be computed as

$$\operatorname{prox}_{\alpha g}(\boldsymbol{z}) = \Gamma_{\lambda\alpha}(\boldsymbol{z}) = [|\boldsymbol{z}| - \lambda\alpha]_+ \operatorname{sgn}(\boldsymbol{z}), \tag{14}$$

where for brevity, we denote the soft shrinkage operator by $\Gamma_{\lambda\alpha}(\boldsymbol{z})$. Here $[\boldsymbol{z}]_+$ denotes the vector whose components are given by the maximum between z_i and 0. Note, however, that there is no explicit expression for the proximal operator of $g(\boldsymbol{D}^*\boldsymbol{x}) = \lambda\|\boldsymbol{D}^*\boldsymbol{x}\|_1$, i.e., there is no closed form solution to

$$\operatorname*{arg\,min}_{\boldsymbol{u}\in\mathbb{R}^n} \left\{ \alpha\lambda\|\boldsymbol{D}^*\boldsymbol{u}\|_1 + \frac{1}{2}\|\boldsymbol{u}-\boldsymbol{x}\|_2^2 \right\}. \tag{15}$$

In the next subsection, we introduce two popular approaches for transforming the problem (13): smoothing and decomposition. We will show in Sections II-D and II-E that both transformations lead to algorithms which only require computation of the proximal operator of $g(\boldsymbol{z})$, and not that of $g(\boldsymbol{D}^*\boldsymbol{x})$.

C. The Smoothing and Decomposition Transformations

The first approach to transform (13) is the smoothing method in which the nonsmooth function $g(\boldsymbol{z})$ is replaced by its Moreau envelope $g_\mu(\boldsymbol{z})$, which can be seen as a smooth approximation. By letting $\boldsymbol{z} = \boldsymbol{D}^*\boldsymbol{x}$, the smoothed problem becomes

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \{H_\mu(\boldsymbol{x}) = f(\boldsymbol{x}) + g_\mu(\boldsymbol{D}^*\boldsymbol{x})\}, \tag{16}$$

to which MFISTA can be applied since it only requires evaluating the proximal operator of $g(\boldsymbol{z})$. From the general properties of the Moreau envelope, and from the fact that the norms of the subgradients of g are bounded above by L_g, we can deduce that there exists some β_1, $\beta_2 > 0$ such that $\beta_1 + \beta_2 = L_g$ and $g(\boldsymbol{z}) - \beta_1\mu \le g_\mu(\boldsymbol{z}) \le g(\boldsymbol{z}) + \beta_2\mu$ for all $\boldsymbol{z} \in \mathbb{R}^p$ (see [22], [24]). This shows that a smaller μ leads to a finer approximation.

The second approach for transforming the problem is the decomposition method in which we consider:

$$\min_{\boldsymbol{x}\in\mathbb{R}^n,\boldsymbol{z}\in\mathbb{R}^p} \left\{ G_\rho(\boldsymbol{x},\boldsymbol{z}) = f(\boldsymbol{x}) + g(\boldsymbol{z}) + \frac{\rho}{2}\|\boldsymbol{z}-\boldsymbol{D}^*\boldsymbol{x}\|_2^2 \right\}. \tag{17}$$

With $\rho \to \infty$, this problem is equivalent to the following constrained formulation of the original problem (13):

$$\begin{aligned} \min\quad & \{f(\boldsymbol{x}) + g(\boldsymbol{z})\} \\ \text{s.t.}\quad & \boldsymbol{z} = \boldsymbol{D}^*\boldsymbol{x}, \quad \boldsymbol{x}\in\mathbb{R}^n, \boldsymbol{z}\in\mathbb{R}^p. \end{aligned} \tag{18}$$

Evidently, there is a close relationship between the approximate models (16) and (17). Indeed, fixing $\boldsymbol{x}$ and minimizing the objective function of (17) with respect to $\boldsymbol{z}$ we obtain

$$\begin{aligned} &\min_{\boldsymbol{x}\in\mathbb{R}^n,\boldsymbol{z}\in\mathbb{R}^p} \left\{ f(\boldsymbol{x}) + g(\boldsymbol{z}) + \frac{\rho}{2}\|\boldsymbol{z}-\boldsymbol{D}^*\boldsymbol{x}\|_2^2 \right\} \\ &\qquad = \min_{\boldsymbol{x}\in\mathbb{R}^n} \left\{ f(\boldsymbol{x}) + g_{\frac{1}{\rho}}(\boldsymbol{D}^*\boldsymbol{x}) \right\}. \end{aligned} \tag{19}$$

Therefore, the two models are equivalent in the sense that their optimal solution set (limited to $\boldsymbol{x}$) is the same when $\mu = \frac{1}{\rho}$. For analysis sparse recovery, both transformations lead to RALASSO (5). However, as we shall see, the resulting smoothing-based and decomposition-based algorithms and their analysis are very different.

D. The Smoothing-Based Method

Since (16) is a smooth problem we can apply an optimal first-order method such as MFISTA with $F = H_\mu = f(\boldsymbol{x}) + g_\mu(\boldsymbol{D}^*\boldsymbol{x})$ and $G \equiv 0$ in (11). The Lipschitz constant of H_μ is given by $L_{\nabla f} + \frac{\|\boldsymbol{D}\|_2^2}{\mu}$, and according to (10) the gradient of $\nabla g_\mu(\boldsymbol{D}^*\boldsymbol{x})$ is equal to $\frac{1}{\mu}\boldsymbol{D}(\boldsymbol{D}^*\boldsymbol{x} - \operatorname{prox}_{\mu g}(\boldsymbol{D}^*\boldsymbol{x}))$. The expression $\operatorname{prox}_{\mu g}(\boldsymbol{D}^*\boldsymbol{x})$ is calculated by first computing $\operatorname{prox}_{\mu g}(\boldsymbol{z})$, and then letting $\boldsymbol{z} = \boldsymbol{D}^*\boldsymbol{x}$.

Returning to the analysis sparse recovery problem, after smoothing we obtain

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \left\{ H_\mu(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|_2^2 + g_\mu(\boldsymbol{D}^*\boldsymbol{x}) \right\}, \tag{20}$$

where

$$\begin{aligned} g_\mu(\boldsymbol{D}^*\boldsymbol{x}) &= \min_{\boldsymbol{u}} \left\{ \lambda\|\boldsymbol{u}\|_1 + \frac{1}{2\mu}\|\boldsymbol{u}-\boldsymbol{D}^*\boldsymbol{x}\|_2^2 \right\} \\ &= \sum_{i=1}^{p} \lambda\mathcal{H}_{\lambda\mu}((\boldsymbol{D}^*\boldsymbol{x})[i]). \end{aligned}$$

The function $\mathcal{H}_\alpha(x)$ with parameter $\alpha > 0$ is the so-called Huber function [34], and is given by

$$\mathcal{H}_\alpha(x) = \begin{cases} \frac{1}{2\alpha}x^2 & \text{if } |x| < \alpha \\ |x| - \frac{\alpha}{2} & \text{otherwise.} \end{cases} \tag{21}$$

From (14), the gradient of $g_\mu(\boldsymbol{D}^*\boldsymbol{x})$ is equal to

$$\nabla g_\mu(\boldsymbol{D}^*\boldsymbol{x}) = \frac{1}{\mu}\boldsymbol{D}(\boldsymbol{D}^*\boldsymbol{x} - \Gamma_{\lambda\mu}(\boldsymbol{D}^*\boldsymbol{x})). \tag{22}$$

Applying MFISTA to (20), results in the SFISTA algorithm, summarized in Algorithm 1.

Algorithm 1: Smoothing-based MFISTA (SFISTA)

```
Input: An upper bound $L \ge \|\boldsymbol{A}\|_2^2 + \frac{\|\boldsymbol{D}\|_2^2}{\mu}$.
Step 0. Take $\boldsymbol{y}_1 = \boldsymbol{x}_0, t_1 = 1$.
Step k. ($k \ge 1$) Compute
    $\nabla f(\boldsymbol{y}_k) = \boldsymbol{A}^*(\boldsymbol{A}\boldsymbol{y}_k - \boldsymbol{b})$.
    $\nabla g_\mu(\boldsymbol{D}^*\boldsymbol{x}_{k-1}) = \frac{1}{\mu}\boldsymbol{D}(\boldsymbol{D}^*\boldsymbol{x}_{k-1} - \Gamma_{\lambda\mu}(\boldsymbol{D}^*\boldsymbol{x}_{k-1}))$.
    $\boldsymbol{z}_k = \boldsymbol{y}_k - \frac{1}{L}(\nabla f(\boldsymbol{y}_k) + \nabla g_\mu(\boldsymbol{D}^*\boldsymbol{x}_{k-1}))$.
    $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$.
    $\boldsymbol{x}_k = \operatorname{argmin}\{H_\mu(\boldsymbol{x}) : \boldsymbol{x} = \boldsymbol{z}_k, \boldsymbol{x}_{k-1}\}$.
    $\boldsymbol{y}_{k+1} = \boldsymbol{x}_k + \frac{t_k}{t_{k+1}}(\boldsymbol{z}_k - \boldsymbol{x}_k) + \frac{t_k - 1}{t_{k+1}}(\boldsymbol{x}_k - \boldsymbol{x}_{k-1})$.
```

E. The Decomposition-Based Method

We can also employ MFISTA on the decomposition model

$$\min_{\boldsymbol{x}\in\mathbb{R}^n,\boldsymbol{z}\in\mathbb{R}^p} \{G_\rho(\boldsymbol{x},\boldsymbol{z}) = F_\rho(\boldsymbol{x},\boldsymbol{z}) + G(\boldsymbol{x},\boldsymbol{z})\}, \tag{23}$$

where we take the smooth part as $F_\rho(\boldsymbol{x},\boldsymbol{z}) = f(\boldsymbol{x}) + \frac{\rho}{2}\|\boldsymbol{z} - \boldsymbol{D}^*\boldsymbol{x}\|_2^2$ and the nonsmooth part as $G(\boldsymbol{x},\boldsymbol{z}) = g(\boldsymbol{z})$. In order to apply MFISTA to (17), we need to compute the proximal operator of αG for a given constant $\alpha > 0$, which is given by

$$\operatorname{prox}_{\alpha G}(\boldsymbol{x},\boldsymbol{z}) = \begin{pmatrix} \boldsymbol{x} \\ \operatorname{prox}_{\alpha g}(\boldsymbol{z}) \end{pmatrix}. \tag{24}$$

In RALASSO (5), $G(\boldsymbol{x},\boldsymbol{z}) = \lambda\|\boldsymbol{z}\|_1$ and $F_\rho(\boldsymbol{x},\boldsymbol{z}) = \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|_2^2 + \frac{1}{2}\rho\|\boldsymbol{z} - \boldsymbol{D}^*\boldsymbol{x}\|_2^2$. Therefore,

$$\mathrm{prox}_{\alpha G}(\boldsymbol{x},\boldsymbol{z}) = \begin{pmatrix} \boldsymbol{x} \\ \Gamma_{\lambda\alpha}(\boldsymbol{z}) \end{pmatrix}. \tag{25}$$

The Lipschitz constant of ∇F is equal to $(\|\boldsymbol{A}\|_2^2+\rho(1+\|\boldsymbol{D}\|_2^2))$. By applying MFISTA directly, we have the DFISTA algorithm, stated in Algorithm 2.

Algorithm 2: Decomposition-based MFISTA (DFISTA)

```
Input: An upper bound $L \geq (\|\boldsymbol{A}\|_2^2 + \rho(1 + \|\boldsymbol{D}\|_2^2))$.
Step 0. Take $\boldsymbol{u}_1 = \boldsymbol{x}_0, \boldsymbol{v}_1 = \boldsymbol{z}_0, t_1 = 1$.
Step k. ($k \geq 1$) Compute
$\nabla_{\boldsymbol{x}} F_\rho(\boldsymbol{u}_k, \boldsymbol{v}_k) = \boldsymbol{A}^*(\boldsymbol{A}\boldsymbol{u}_k - \boldsymbol{b}) + \rho\boldsymbol{D}(\boldsymbol{D}^*\boldsymbol{u}_k - \boldsymbol{v}_k)$.
$\nabla_{\boldsymbol{z}} F_\rho(\boldsymbol{u}_k, \boldsymbol{v}_k)) = \rho(\boldsymbol{v}_k - \boldsymbol{D}^*\boldsymbol{u}_k)$.
$\boldsymbol{p}_k = \boldsymbol{u}_k - \frac{1}{L}\nabla_{\boldsymbol{x}} F_\rho(\boldsymbol{u}_k, \boldsymbol{v}_k)$.
$\boldsymbol{q}_k = \Gamma_{\frac{\lambda}{L}}(\boldsymbol{v}_k - \frac{1}{L}\nabla_{\boldsymbol{z}} F_\rho(\boldsymbol{u}_k, \boldsymbol{v}_k))$.
$t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$.
$(\boldsymbol{x}_k, \boldsymbol{z}_k)$
$= \mathrm{argmin}\{G_\rho(\boldsymbol{x},\boldsymbol{z}) : (\boldsymbol{x},\boldsymbol{z}) = (\boldsymbol{p}_k, \boldsymbol{q}_k), (\boldsymbol{x}_{k-1}, \boldsymbol{z}_{k-1})\}$.
$\boldsymbol{u}_{k+1} = \boldsymbol{x}_k + \frac{t_k}{t_{k+1}}(\boldsymbol{p}_k - \boldsymbol{x}_k) + \frac{t_k - 1}{t_{k+1}}(\boldsymbol{x}_k - \boldsymbol{x}_{k-1})$.
$\boldsymbol{v}_{k+1} = \boldsymbol{z}_k + \frac{t_k}{t_{k+1}}(\boldsymbol{q}_k - \boldsymbol{z}_k) + \frac{t_k - 1}{t_{k+1}}(\boldsymbol{z}_k - \boldsymbol{z}_{k-1})$.
```

III. Convergence Analysis

In this section we analyze the convergence behavior of both the smoothing-based and decomposition-based methods. Convergence of smoothing algorithms has been treated in [22], [24]. In order to make the paper self contained, we quote the main results here. We then analyze the convergence of the decomposition approach. Both methods require the same type of operations at each iteration: the computation of the gradient of the smooth function f, and of the proximal operator corresponding to αg, which means that they have the same computational cost per iteration. However, we show that smoothing converges faster than decomposition based methods. Specifically, the smoothing-based algorithm is guaranteed to generate an ε-optimal solution within $O(1/\varepsilon)$ iterations, whereas the decomposition-based approach requires $O(1/\varepsilon^{1.5})$ iterations. We prove the results by analyzing SFISTA and DFISTA for the general problem (13), however, the same analysis can be easily extended to other optimal first-order methods, such as the one described in [22].

A. Convergence of the Smoothing-Based Method

For SFISTA the sequence $\{\boldsymbol{x}_k\}$ satisfies the following relationship [25]:

$$H_\mu(\boldsymbol{x}_k) - H_\mu(\hat{\boldsymbol{x}}_\mu) \leq \frac{2\left(L_{\nabla f} + \frac{\|\boldsymbol{D}\|_2^2}{\mu}\right)\Lambda_1}{(k+1)^2}, \tag{26}$$

where Λ_1 is an upper bound on the expression $\|\hat{\boldsymbol{x}}_\mu - \boldsymbol{x}_0\|_2$ with $\hat{\boldsymbol{x}}_\mu$ being an arbitrary optimal solution of the smoothed problem (16), and $\boldsymbol{x}_0$ is the initial point of the algorithm. Of course, this rate of convergence is problematic since we are more interested in bounding the expression $H(\boldsymbol{x}_k) - \hat{H}$ rather than the expression $H_\mu(\boldsymbol{x}_k) - H_\mu(\hat{\boldsymbol{x}}_\mu)$, which is in terms of the smoothed problem. Here, $\hat{H}$ stands for the optimal value for original nonsmooth problem (13). For that, we can use the following result from [24].

Theorem III.1: [24] Let $\{\boldsymbol{x}_k\}$ be the sequence generated by applying MFISTA to the problem (16). *Let $\boldsymbol{x}_0$ be the initial point and let $\hat{\boldsymbol{x}}$ denote the optimal solution of* (13). *An ε-optimal solution of* (13), *i.e., $|H(\boldsymbol{x}_k) - H(\hat{\boldsymbol{x}})| \leq \varepsilon$, is obtained in the smoothing-based method using MFISTA after at most*

$$K = 2\|\boldsymbol{D}\|_2\sqrt{L_g\Lambda_1}\frac{1}{\varepsilon} + \sqrt{L_{\nabla f}\Lambda_1}\frac{1}{\sqrt{\varepsilon}} \tag{27}$$

iterations with μ chosen as

$$\mu = \sqrt{\frac{\|\boldsymbol{D}\|_2^2}{L_g}}\frac{\varepsilon}{\sqrt{\|\boldsymbol{D}\|_2^2 L_g} + \sqrt{\|\boldsymbol{D}\|_2^2 L_g + L_{\nabla f}\varepsilon}}, \tag{28}$$

in which L_g and $L_{\nabla f}$ are the Lipschitz constants of g and the gradient function of f in (13), *and $\Lambda_1 = \|\boldsymbol{x}_0 - \hat{\boldsymbol{x}}_\mu\|_2$. We use $\hat{\boldsymbol{x}}_\mu$ to denote the optimal solution of problem* (16).

Remarks: For analysis sparse recovery using SFISTA, $L_g = \lambda p^{\frac{1}{2}}$ and $L_{\nabla f} = \|\boldsymbol{A}\|_2^2$, which can be plugged into the expressions in the theorem.

B. Convergence of the Decomposition-Based Method

A key property of the decomposition model (17) is that its minimal value is bounded above by the optimal value $\hat{H}$ in the original problem (13).

Lemma III.1: Let $\hat{G}_\rho$ be the optimal value of problem (17) *and $\hat{H}$ be the optimal value of problem* (13). *Then $\hat{G}_\rho \leq \hat{H}$.*

Proof: The proof follows from adding the constraint $\boldsymbol{z} = \boldsymbol{D}^*\boldsymbol{x}$ to the optimization:

$$\begin{aligned} \hat{G}_\rho &= \min_{\boldsymbol{x}\in\mathbb{R}^n, \boldsymbol{z}\in\mathbb{R}^p} \left\{ f(\boldsymbol{x}) + g(\boldsymbol{z}) + \frac{\rho}{2}\|\boldsymbol{z} - \boldsymbol{D}^*\boldsymbol{x}\|_2^2 \right\} \\ &\leq \min_{\boldsymbol{x}\in\mathbb{R}^n, \boldsymbol{z}\in\mathbb{R}^p, \boldsymbol{z}=\boldsymbol{D}^*\boldsymbol{x}} \left\{ f(\boldsymbol{x}) + g(\boldsymbol{z}) + \frac{\rho}{2}\|\boldsymbol{z} - \boldsymbol{D}^*\boldsymbol{x}\|_2^2 \right\} \\ &= \min_{\boldsymbol{x}\in\mathbb{R}^n} \left\{ f(\boldsymbol{x}) + g(\boldsymbol{D}^*\boldsymbol{x}) \right\}, \end{aligned} \tag{29}$$

which is equal to $\hat{H}$.

The next theorem is our main convergence result establishing that an ε-optimal solution can be reached after $O(1/\varepsilon^{1.5})$ iterations. By assuming that the functions f and g are nonnegative, which is not an unusual assumption, we have the following theorem.

Theorem III.2: Let $\{\boldsymbol{x}_k, \boldsymbol{z}_k\}$ be the sequences generated by applying MFISTA to (17) *with both f and g both being nonnegative functions. The initial point is taken as $(\boldsymbol{x}_0, \boldsymbol{z}_0)$ with $\boldsymbol{z}_0 = \boldsymbol{D}^*\boldsymbol{x}_0$. Let $\hat{\boldsymbol{x}}$ denote the optimal solution of the original problem* (13). *An ε-optimal solution of problem* (13), *i.e., $|H(\boldsymbol{x}_k) - H(\hat{\boldsymbol{x}})| \leq \varepsilon$, is obtained using the decomposition-based method after at most*

$$K = \max\left\{ \frac{16\sqrt{(1 + \|\boldsymbol{D}\|^2\Lambda_2 H(\boldsymbol{x}_0))}L_g}{\varepsilon^{1.5}}, \frac{2\sqrt{L_{\nabla f}\Lambda_2}}{\sqrt{\varepsilon}} \right\} \tag{30}$$

iterations of MFISTA with ρ chosen as

$$\rho = \left(\frac{L_g \sqrt{2H(\boldsymbol{x}_0)} K^2}{2(1 + \|\boldsymbol{D}\|^2)\Lambda_2} \right)^{2/3}. \tag{31}$$

Here L_g and $L_{\nabla f}$ are the Lipschitz constants for g and the gradient function of f in (13)*, and $\Lambda_2 = \|\boldsymbol{x}_0 - \hat{\boldsymbol{x}}_\rho\|_2^2 + \|\boldsymbol{z}_0 - \hat{\boldsymbol{z}}_\rho\|_2^2$. We use $\hat{\boldsymbol{x}}_\rho, \hat{\boldsymbol{z}}_\rho$ to denote the optimal solutions to* (17)*.*

Proof: Since the monotone version of FISTA is applied we have

$$\begin{aligned} &f(\boldsymbol{x}_k) + g(\boldsymbol{z}_k) + \frac{\rho}{2}\|\boldsymbol{z}_k - \boldsymbol{D}^*\boldsymbol{x}_k\|_2^2 \\ &= G_\rho(\boldsymbol{x}_k, \boldsymbol{z}_k) \le G_\rho(\boldsymbol{x}_0, \boldsymbol{z}_0) = f(\boldsymbol{x}_0) + g(\boldsymbol{D}^*\boldsymbol{x}_0) = H(\boldsymbol{x}_0). \end{aligned} \tag{32}$$

With the assumption that f and g are nonnegative, it follows that

$$\frac{\rho}{2}\|\boldsymbol{z}_k - \boldsymbol{D}^*\boldsymbol{x}_k\|_2^2 \le H(\boldsymbol{x}_0),$$

and therefore

$$\|\boldsymbol{z}_k - \boldsymbol{D}^*\boldsymbol{x}_k\|_2 \le \sqrt{\frac{2H(\boldsymbol{x}_0)}{\rho}}. \tag{33}$$

The gradient of $f(\boldsymbol{x}) + \frac{\rho}{2}\|\boldsymbol{z} - \boldsymbol{D}^*\boldsymbol{x}\|_2^2$, is Lipschitz continuous with parameter $(L_{\nabla f} + \rho(1 + \|\boldsymbol{D}\|_2^2))$. According to [25], by applying MFISTA, we obtain a sequence $\{(\boldsymbol{x}_k, \boldsymbol{z}_k)\}$ satisfying

$$G_\rho(\boldsymbol{x}_k, \boldsymbol{z}_k) - \hat{G}_\rho \le \frac{2(L_{\nabla f} + \rho(1 + \|\boldsymbol{D}\|_2^2))\Lambda_2}{k^2}.$$

Using lemma III.1 and the notation

$$A = 2L_{\nabla f}\Lambda_2, B = 2(1 + \|\boldsymbol{D}\|_2^2)\Lambda_2,$$

we have

$$G_\rho(\boldsymbol{x}_k, \boldsymbol{z}_k) - \hat{H} \le \frac{A + \rho B}{k^2}. \tag{34}$$

We therefore conclude that

$$\begin{aligned} H(\boldsymbol{x}_k) &= f(\boldsymbol{x}_k) + g(\boldsymbol{D}^*\boldsymbol{x}_k) \\ &= f(\boldsymbol{x}_k) + g(\boldsymbol{z}_k) + g(\boldsymbol{D}^*\boldsymbol{x}_k) - g(\boldsymbol{z}_k) \\ &\le G_\rho(\boldsymbol{x}_k, \boldsymbol{z}_k) + L_g\|\boldsymbol{z}_k - \boldsymbol{D}^*\boldsymbol{x}_k\|_2 \\ &\le \hat{H} + \frac{A + \rho B}{k^2} + L_g\|\boldsymbol{z}_k - \boldsymbol{D}^*\boldsymbol{x}_k\|_2 \\ &\le \hat{H} + \frac{A + \rho B}{k^2} + L_g\sqrt{\frac{2H(\boldsymbol{x}_0)}{\rho}}. \end{aligned}$$

The first inequality follows from the Lipschitz condition for the function g, the second inequality is obtained from (34), and the last inequality is a result of (33).

We now seek the "best" ρ that minimizes the upper bound, or equivalently, minimizes the term

$$\frac{A + \rho B}{k^2} + L_g\sqrt{\frac{2H(\boldsymbol{x}_0)}{\rho}} = \frac{A}{k^2} + C\rho + \frac{D}{\sqrt{\rho}}, \tag{35}$$

where $C = \frac{B}{k^2}$ and $D = L_g\sqrt{2H(\boldsymbol{x}_0)}$. Setting the derivative to zero, the optimal value of ρ is $\rho = \left(\frac{D}{2C}\right)^{2/3}$, and

$$H(\boldsymbol{x}_k) \le \hat{H} + \frac{A}{k^2} + 2C^{1/3}D^{2/3}. \tag{36}$$

Therefore, to obtain an ε-optimal solution, it is enough that

$$\frac{A}{k^2} \le \frac{\varepsilon}{2}, \quad \frac{2B^{1/3}D^{2/3}}{k^{2/3}} \le \frac{\varepsilon}{2}, \tag{37}$$

or

$$\begin{aligned} k &\ge \max\left\{ \frac{4^{3/2}B^{1/2}D}{\varepsilon^{1.5}}, \frac{\sqrt{2A}}{\sqrt{\varepsilon}} \right\} \\ &= \max\left\{ \frac{16\sqrt{(1 + \|\boldsymbol{D}\|^2\Lambda_2 H(\boldsymbol{x}_0))}L_g}{\varepsilon^{1.5}}, \frac{2\sqrt{L_{\nabla f}\Lambda_2}}{\sqrt{\varepsilon}} \right\}, \end{aligned} \tag{38}$$

completing the proof.

Remarks:

1. As in SFISTA, when treating the analysis sparse recovery problem, $L_g = \lambda p^{\frac{1}{2}}$ and $L_{\nabla f} = \|\boldsymbol{A}\|_2^2$, which again can be plugged into the expressions in the theorem.
2. MFISTA is applied in SFISTA and DFISTA to guarantee a mathematical rigorous proof, i.e., the existence of (32). In real application, FISTA without monotone operations can also be applied to yield corresponding smoothing and decomposition based algorithms.

Comparing the results of smoothing-based and decomposition-based methods, we immediately conclude that the smoothing-based method is preferable. First, it requires only $O(1/\varepsilon)$ iterations to obtain an ε-optimal solution whereas the decomposition approach necessitates $O(1/\varepsilon^{3/2})$ iterations. Note that both bounds are better than the bound $O(1/\varepsilon^2)$ corresponding to general sub-gradient schemes for nonsmooth optimization. Second, the bound in the smoothing approach depends on $\sqrt{L_g}$, and not on L_g, as when using decomposition methods. This is important since, for example, when $g(\boldsymbol{z}) = \|\boldsymbol{z}\|_1$, we have $L_g = p^{\frac{1}{2}}$. In the smoothing approach the dependency on p is of the form $p^{\frac{1}{4}}$ and not $p^{\frac{1}{2}}$, as when using the decomposition algorithm.

IV. Performance Bounds

We now turn to analyze the recovery performance of analysis LASSO when smoothing and decomposition are applied. As we have seen, both transformations lead to the same RALASSO problem in (5). Our main result in this section shows that the reconstruction obtained by solving RALASSO is stable when $\boldsymbol{D}^*\boldsymbol{x}$ has rapidly decreasing coefficients and the noise in the model (1) is small enough. Our performance bound also depends on the choice of parameter ρ in the objective function. Before stating the main theorems, we first introduce a definition and some useful lemmas, whose proofs are detailed in the Appendix.

To ensure stable recovery, we require that the matrix $\boldsymbol{A}$ satisfies the D-RIP:

Definition IV.1: (D-RIP) [12]. The measurement matrix $\boldsymbol{A}$ obeys the restricted isometry property adapted to $\boldsymbol{D}$ with constant σ_s if

$$(1 - \sigma_s)\|\boldsymbol{v}\|_2^2 \le \|\boldsymbol{A}\boldsymbol{v}\|_2^2 \le (1 + \sigma_s)\|\boldsymbol{v}\|_2^2 \tag{39}$$

holds for all $\boldsymbol{v} \in \Sigma_s = \{\boldsymbol{y} : \boldsymbol{y} = \boldsymbol{D}\boldsymbol{x} \text{ and } \|\boldsymbol{x}\|_0 \leq s\}$. *In other words,* Σ_s *is the union of subspaces spanned by all subsets of* s *columns of* $\boldsymbol{D}$.

The following lemma provides a useful inequality for matrices satisfying D-RIP.

Lemma IV.1: Let $\boldsymbol{A}$ *satisfy the D-RIP with parameter* σ_{2s}, *and assume that* $\boldsymbol{u}, \boldsymbol{v} \in \Sigma_s$. *Then,*

$$\mathrm{Re}\langle \boldsymbol{A}\boldsymbol{u}, \boldsymbol{A}\boldsymbol{v}\rangle \geq -\sigma_{2s}\|\boldsymbol{u}\|_2\|\boldsymbol{v}\|_2 + \mathrm{Re}\langle \boldsymbol{u}, \boldsymbol{v}\rangle. \tag{40}$$

In the following, $\hat{\boldsymbol{x}}_\rho$ denotes the optimal solution of RALASSO (5) and $\boldsymbol{x}$ is the original signal in the linear model (1); we also use $\boldsymbol{h}$ to represent the reconstruction error $\boldsymbol{h} = \hat{\boldsymbol{x}}_\rho - \boldsymbol{x}$. Let $\mathcal{T}$ be the indices of coefficients with s largest magnitudes in the vector $\boldsymbol{D}^*\boldsymbol{x}$, and denote the complement of $\mathcal{T}$ by $\mathcal{T}^c$. Setting $\mathcal{T}_0 = \mathcal{T}$, we decompose $\mathcal{T}_0^c$ into sets of size s where $\mathcal{T}_1$ denotes the locations of the s largest coefficients in $\boldsymbol{D}^*_{\mathcal{T}^c}\boldsymbol{x}$, $\mathcal{T}_2$ denote the next s largest coefficients and so on. Finally, we let $\mathcal{T}_{01} = \mathcal{T}_0 \cup \mathcal{T}_1$.

Using the result of Lemma IV.1 and the inequality $\|\boldsymbol{D}^*_{\mathcal{T}_0}\boldsymbol{h}\|_2 + \|\boldsymbol{D}^*_{\mathcal{T}_1}\boldsymbol{h}\|_2 \leq \sqrt{2}\|\boldsymbol{D}^*_{\mathcal{T}_{01}}\boldsymbol{h}\|_2$ since $\mathcal{T}_0$ and $\mathcal{T}_1$ are disjoint, we have the following lemma.

Lemma IV.2: ***(D-RIP property)*** *Let* $\boldsymbol{h} = \hat{\boldsymbol{x}}_\rho - \boldsymbol{x}$ *be the reconstruction error in RALASSO* (5). *We assume that* $\boldsymbol{A}$ *satisfies the D-RIP with parameter* σ_{2s} *and* $\boldsymbol{D}$ *is a tight frame. Then,*

$$\begin{aligned}&\mathrm{Re}\langle \boldsymbol{A}\boldsymbol{h}, \boldsymbol{A}\boldsymbol{D}\boldsymbol{D}^*_{\mathcal{T}_{01}}\boldsymbol{h}\rangle \\ &\geq (1-\sigma_{2s})\|\boldsymbol{D}^*_{\mathcal{T}_{01}}\boldsymbol{h}\|_2^2 - \sqrt{2}s^{-\frac{1}{2}}\sigma_{2s}\|\boldsymbol{D}^*_{\mathcal{T}_{01}}\boldsymbol{h}\|_2\|\boldsymbol{D}^*_{\mathcal{T}^c}\boldsymbol{h}\|_1.\end{aligned} \tag{41}$$

Finally, the lemmas below show that the reconstruction error $\boldsymbol{h}$ and $\|\boldsymbol{D}^*_{\mathcal{T}^c}\boldsymbol{h}\|_1$ can not be very large.

Lemma IV.3: ***(Optimality condition)*** *The optimal solution* $\hat{\boldsymbol{x}}_\rho$ *for RALASSO* (5) *satisfies*

$$\|\boldsymbol{D}^*\boldsymbol{A}^*\boldsymbol{A}\boldsymbol{h}\|_\infty \leq \left(\frac{1}{2} + \|\boldsymbol{D}^*\boldsymbol{D}\|_{1,1}\right)\lambda. \tag{42}$$

Lemma IV.4: ***(Cone constraint)*** *The optimal solution* $\hat{\boldsymbol{x}}_\rho$ *for RALASSO* (5) *satisfies the following cone constraint,*

$$\|\boldsymbol{D}^*_{\mathcal{T}^c}\boldsymbol{h}\|_1 \leq \frac{\lambda}{\rho}p + 3\|\boldsymbol{D}^*_{\mathcal{T}}\boldsymbol{h}\|_1 + 4\|\boldsymbol{D}^*_{\mathcal{T}^c}\boldsymbol{x}\|_1. \tag{43}$$

We are now ready to state our main result.

Theorem IV.1: Let $\boldsymbol{A}$ *be an* $m \times n$ *measurement matrix,* $\boldsymbol{D}$ *an arbitrary* $n \times p$ *tight frame, and let* $\boldsymbol{A}$ *satisfy the D-RIP with* $\sigma_{2s} < 0.1907$. *Consider the measurement* $\boldsymbol{b} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{w}$, *where* $\boldsymbol{w}$ *is noise that satisfies* $\|\boldsymbol{D}^*\boldsymbol{A}^*\boldsymbol{w}\|_\infty \leq \frac{\lambda}{2}$. *Then the solution* $\hat{\boldsymbol{x}}_\rho$ *to RALASSO* (5) *satisfies*

$$\|\hat{\boldsymbol{x}}_\rho - \boldsymbol{x}\|_2 \leq C_0\sqrt{s}\lambda + C_1\frac{\|\boldsymbol{D}^*\boldsymbol{x} - (\boldsymbol{D}^*\boldsymbol{x})_s\|_1}{\sqrt{s}} + C_2\frac{\lambda p}{\sqrt{s}\rho}, \tag{44}$$

for the decomposition transformation and

$$\|\hat{\boldsymbol{x}}_\rho - \boldsymbol{x}\|_2 \leq C_0\sqrt{s}\lambda + C_1\frac{\|\boldsymbol{D}^*\boldsymbol{x} - (\boldsymbol{D}^*\boldsymbol{x})_s\|_1}{\sqrt{s}} + C_2\frac{\lambda\mu p}{\sqrt{s}}, \tag{45}$$

for the smoothing transformation. Here $(\boldsymbol{D}^*\boldsymbol{x})_s$ *is the vector consisting of the largest* s *entries of* $\boldsymbol{D}^*\boldsymbol{x}$ *in magnitude,* C_1 *and* C_2 *are constants depending on* σ_{2s}, *and* C_0 *depends on* σ_{2s} *and* $\|\boldsymbol{D}^*\boldsymbol{D}\|_{1,1}$.

Proof: The proof follows mainly from the ideas in [9], [28], and proceeds in two steps. First, we try to show that $\boldsymbol{D}^*\boldsymbol{h}$ inside $\mathcal{T}_{01}$ is bounded by the terms of $\boldsymbol{D}^*\boldsymbol{h}$ outside the set $\mathcal{T}$. Then we show that $\boldsymbol{D}^*_{\mathcal{T}^c}\boldsymbol{h}$ is essentially small.

From Lemma IV.2,

$$\begin{aligned}&\mathrm{Re}\langle \boldsymbol{A}\boldsymbol{h}, \boldsymbol{A}\boldsymbol{D}\boldsymbol{D}^*_{\mathcal{T}_{01}}\boldsymbol{h}\rangle \\ &\geq (1-\sigma_{2s})\|\boldsymbol{D}^*_{\mathcal{T}_{01}}\boldsymbol{h}\|_2^2 - \sqrt{2}s^{-\frac{1}{2}}\sigma_{2s}\|\boldsymbol{D}^*_{\mathcal{T}_{01}}\boldsymbol{h}\|_2\|\boldsymbol{D}^*_{\mathcal{T}^c}\boldsymbol{h}\|_1.\end{aligned} \tag{46}$$

Using the fact that $\mathrm{Re}\langle \boldsymbol{x}, \boldsymbol{y}\rangle \leq |\langle \boldsymbol{x}, \boldsymbol{y}\rangle| \leq \|\boldsymbol{x}\|_1\|\boldsymbol{y}\|_\infty$, we obtain that

$$\begin{aligned}\mathrm{Re}\langle \boldsymbol{A}\boldsymbol{h}, \boldsymbol{A}\boldsymbol{D}\boldsymbol{D}^*_{\mathcal{T}_{01}}\boldsymbol{h}\rangle &= \mathrm{Re}\langle \boldsymbol{D}^*\boldsymbol{A}^*\boldsymbol{A}\boldsymbol{h}, \boldsymbol{D}^*_{\mathcal{T}_{01}}\boldsymbol{h}\rangle \\ &\leq \|\boldsymbol{D}^*\boldsymbol{A}^*\boldsymbol{A}\boldsymbol{h}\|_\infty\|\boldsymbol{D}^*_{\mathcal{T}_{01}}\boldsymbol{h}\|_1 \\ &\leq \sqrt{2s}c_0\lambda\|\boldsymbol{D}^*_{\mathcal{T}_{01}}\boldsymbol{h}\|_2,\end{aligned} \tag{47}$$

with $c_0 = \frac{1}{2} + \|\boldsymbol{D}^*\boldsymbol{D}\|_{1,1}$. The second inequality is a result of Lemma IV.3 and the fact that $\|\boldsymbol{D}^*_{\mathcal{T}_{01}}\boldsymbol{h}\|_1 \leq \sqrt{2s}\|\boldsymbol{D}^*_{\mathcal{T}_{01}}\boldsymbol{h}\|_2$, in which $2s$ is the number of nonzero terms in $\boldsymbol{D}^*_{\mathcal{T}_{01}}\boldsymbol{h}$. Combining (46) and (47), we get

$$\|\boldsymbol{D}^*_{\mathcal{T}_{01}}\boldsymbol{h}\|_2 \leq \frac{\sqrt{2s}\lambda c_0 + \sqrt{2}s^{-\frac{1}{2}}\sigma_{2s}\|\boldsymbol{D}^*_{\mathcal{T}^c}\boldsymbol{h}\|_1}{1 - \sigma_{2s}}. \tag{48}$$

Then the second step bounds $\|\boldsymbol{D}^*_{\mathcal{T}^c}\boldsymbol{h}\|_1$. From (48),

$$\begin{aligned}\|\boldsymbol{D}^*_{\mathcal{T}}\boldsymbol{h}\|_1 &\leq \sqrt{s}\|\boldsymbol{D}^*_{\mathcal{T}}\boldsymbol{h}\|_2 \leq \sqrt{s}\|\boldsymbol{D}^*_{\mathcal{T}_{01}}\boldsymbol{h}\|_2 \\ &\leq \frac{\sqrt{2}\lambda s c_0 + \sqrt{2}\sigma_s\|\boldsymbol{D}^*_{\mathcal{T}^c}\boldsymbol{h}\|_1}{1 - \sigma_{2s}}.\end{aligned} \tag{49}$$

Finally, using Lemma IV.4 and (49),

$$\|\boldsymbol{D}^*_{\mathcal{T}^c}\boldsymbol{h}\|_1 \leq \frac{\lambda}{\rho}p + \frac{3\sqrt{2}\lambda s c_0 + 3\sqrt{2}\sigma_{2s}\|\boldsymbol{D}^*_{\mathcal{T}^c}\boldsymbol{h}\|_1}{1 - \sigma_{2s}} + 4\|\boldsymbol{D}^*_{\mathcal{T}^c}\boldsymbol{x}\|_1. \tag{50}$$

Since $\sigma_{2s} < 0.1907$, we have $1 - (1 + 3\sqrt{2})\sigma_{2s} > 0$. Rearranging terms, the above inequality becomes

$$\begin{aligned}&\|\boldsymbol{D}^*_{\mathcal{T}^c}\boldsymbol{h}\|_1 \\ &\leq \frac{1 - \sigma_{2s}}{1-(1+3\sqrt{2})\sigma_{2s}}\frac{\lambda}{\rho}p + \frac{3\sqrt{2}\lambda s c_0 + 4(1-\sigma_{2s})\|\boldsymbol{D}^*_{\mathcal{T}^c}\boldsymbol{x}\|_1}{1 - (1 + 3\sqrt{2})\sigma_{2s}}.\end{aligned} \tag{51}$$

We now derive the bound on the reconstruction error. Using the results of (48) and (51), we get

$$\begin{aligned}\|\boldsymbol{h}\|_2 &= \|\boldsymbol{D}^*\boldsymbol{h}\|_2 \leq \|\boldsymbol{D}^*_{\mathcal{T}_{01}}\boldsymbol{h}\|_2 + \sum_{j\geq 2}\|\boldsymbol{D}^*_{\mathcal{T}_j}\boldsymbol{h}\|_2 \\ &\leq \frac{\sqrt{2s}\lambda c_0 + \sqrt{2}s^{-\frac{1}{2}}\sigma_{2s}\|\boldsymbol{D}^*_{\mathcal{T}^c}\boldsymbol{h}\|_1}{1 - \sigma_{2s}} + s^{-\frac{1}{2}}\|\boldsymbol{D}^*_{\mathcal{T}^c}\boldsymbol{h}\|_1 \\ &= \frac{c_0\lambda\sqrt{2s}}{1 - \sigma_{2s}} + \frac{((\sqrt{2}-1)\sigma_{2s} + 1)s^{-\frac{1}{2}}\|\boldsymbol{D}^*_{\mathcal{T}^c}\boldsymbol{h}\|_1}{1 - \sigma_{2s}} \\ &\leq C_0\sqrt{s}\lambda + C_1\frac{\|\boldsymbol{D}^*\boldsymbol{x} - (\boldsymbol{D}^*\boldsymbol{x})_s\|_1}{\sqrt{s}} + C_2\frac{\lambda p}{\sqrt{s}\rho}.\end{aligned} \tag{52}$$

The first equality follows from the assumption that $\boldsymbol{D}$ is a tight frame so that $\boldsymbol{D}\boldsymbol{D}^* = \boldsymbol{I}$. The first inequality is the result of the triangle inequality. The second inequality follows from (48) and the fact that $\sum_{j\geq 2}\|\boldsymbol{D}^*_{\mathcal{T}_j}\boldsymbol{h}\|_2 \leq s^{-\frac{1}{2}}\|\boldsymbol{D}^*_{\mathcal{T}^c}\boldsymbol{h}\|_1$, which is proved in (58) in the Appendix. The constants in the final result are given by

$$C_0 = \frac{4\sqrt{2}c_0}{1-(1+3\sqrt{2})\sigma_{2s}},$$
$$C_1 = \frac{4((\sqrt{2}-1)\sigma_{2s}+1)}{1-(1+3\sqrt{2})\sigma_{2s}},$$
$$C_2 = \frac{(\sqrt{2}-1)\sigma_{2s}+1}{1-(1+3\sqrt{2})\sigma_{2s}}.$$

To obtain the error bound for the smoothing transformation we replace ρ with $1/\mu$ in the result. □

Choosing $\rho \to \infty$ in RALASSO (5) leads to the ALASSO problem for which $\boldsymbol{z} = \boldsymbol{D}^*\boldsymbol{x}$. We then have the following result.

Theorem IV.2: Let $\boldsymbol{A}$ *be an* $m \times n$ *measurement matrix,* $\boldsymbol{D}$ *an arbitrary* $n \times p$ *tight frame, and let* $\boldsymbol{A}$ *satisfy the D-RIP with* $\sigma_{2s} < 0.1907$. *Consider the measurement* $\boldsymbol{b} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{w}$, *where* $\boldsymbol{w}$ *is noise that satisfies* $\|\boldsymbol{D}^*\boldsymbol{A}^*\boldsymbol{w}\|_\infty \leq \frac{\lambda}{2}$. *Then the solution* $\hat{\boldsymbol{x}}$ *to ALASSO* (4) *satisfies*

$$\|\hat{\boldsymbol{x}} - \boldsymbol{x}\|_2 \leq C_0\sqrt{s}\lambda + C_1\frac{\|\boldsymbol{D}^*\boldsymbol{x} - (\boldsymbol{D}^*\boldsymbol{x})_s\|_1}{\sqrt{s}}, \tag{53}$$

where $(\boldsymbol{D}^*\boldsymbol{x})_s$ *is the vector consisting of the largest* s *entries of* $\boldsymbol{D}^*\boldsymbol{x}$ *in magnitude,* C_1 *is a constant depending on* σ_{2s}, *and* C_0 *depends on* σ_{2s} *and* $\|\boldsymbol{D}^*\boldsymbol{D}\|_{1,1}$.

Remarks:

1. When the noise in the system is zero, we can set λ as a positive value which is arbitrarily close to zero. The solution $\hat{\boldsymbol{x}}$ then satisfies $\|\hat{\boldsymbol{x}} - \boldsymbol{x}\| \leq C_1\frac{\|\boldsymbol{D}^*\boldsymbol{x} - (\boldsymbol{D}^*\boldsymbol{x})_s\|_1}{\sqrt{s}}$, which parallels the result for the noiseless synthesis model in [9].
2. When $\boldsymbol{D}^*$ is a tight frame, we have $\boldsymbol{D}\boldsymbol{D}^* = \boldsymbol{I}$. Therefore by letting $\boldsymbol{v} = \boldsymbol{D}^*\boldsymbol{x}$, we can reformulate the original analysis model as
$$\min_{\boldsymbol{v}} \frac{1}{2}\|\boldsymbol{A}\boldsymbol{D}\boldsymbol{v} - \boldsymbol{b}\|_2^2 + \lambda\|\boldsymbol{v}\|_1. \tag{54}$$
Assuming that the noise term satisfies the l_2 norm constraint $\|\boldsymbol{w}\|_2 \leq \varepsilon$, we have
$$\|\boldsymbol{D}^*\boldsymbol{A}^*\boldsymbol{w}\|_\infty \leq \|\boldsymbol{D}^*\boldsymbol{A}^*\boldsymbol{w}\|_2 \leq \|\boldsymbol{D}^*\boldsymbol{A}^*\|_2\|\boldsymbol{w}\|_2 \leq \varepsilon\|\boldsymbol{D}^*\boldsymbol{A}^*\|_2. \tag{55}$$
When $\boldsymbol{A}$ satisfies D-RIP with $\sigma_{2s} < 0.1907$, by letting $\lambda = 2\varepsilon\|\boldsymbol{D}^*\boldsymbol{A}^*\|_2$ we have
$$\|\hat{\boldsymbol{v}} - \boldsymbol{v}\|_2 \leq \|\boldsymbol{D}^*\|_2\|\hat{\boldsymbol{x}} - \boldsymbol{x}\|_2 \leq \tilde{C}_0\varepsilon + \tilde{C}_1\frac{\|\boldsymbol{v} - (\boldsymbol{v})_s\|_1}{\sqrt{s}}. \tag{56}$$
This result has a form similar to the reconstruction error bound shown in [9]. However, the specific constants are different since in [9] the matrix $\boldsymbol{A}\boldsymbol{D}$ is required to satisfy the RIP, whereas in our paper we require only that the D-RIP is satisfied.
3. A similar performance bound is introduced in [28] and shown to be valid when $\sigma_{3s} < 0.25$. Using Corollary 3.4 in [35], this is equivalent to $\sigma_{2s} < 0.0833$. Thus the results in Theorem IV.2 allow for a looser constraint on ALASSO recovery.
4. The performance bound of Theorem IV.1 implies that a larger choice of ρ, or a smaller parameter μ, leads to a smaller reconstruction error bound. This trend is intuitive since large ρ or small μ results in smaller model inaccuracy. However, a larger ρ or a smaller μ leads to a larger Lipschitz constant and thus results in slower convergence according to Theorem II.1. The idea of parameter continuation [36] can be introduced to both ρ and μ to accelerate the convergence while obtaining a desired reconstruction accuracy. More details will be given in the next section.

V. Numerical Results

In the numerical examples, we use both randomly generated data and MRI image reconstruction to demonstrate that SFISTA performs better than DFISTA. In the last example we also introduce a continuation technique to further speed up convergence of the smoothing-based method. We further compare SFISTA with the existing methods in [18], [20], [23] using MRI image reconstruction, and show its advantages.

A. Randomly Generated Data in a Noiseless Case

In this simulation, the entries in the $m \times n$ measurement matrix $\boldsymbol{A}$ were randomly generated according to a normal distribution. The $n \times p$ matrix $\boldsymbol{D}$ is a random tight frame. First we generated a $p \times n$ matrix whose elements follow an i.i.d. Gaussian distribution. Then QR factorization was performed on this random matrix to yield the tight frame $\boldsymbol{D}$ with $\boldsymbol{D}\boldsymbol{D}^* = \boldsymbol{I}$ ($\boldsymbol{D}^*$ comprises the first n columns from $\boldsymbol{Q}$, which was generated from the QR factorization).

In the simulation we let $n = 120$ and $p = 144$, and we also set the values of m and the number of zero terms named l in $\boldsymbol{D}^*\boldsymbol{x}$ according to the following formula:

$$m = \alpha n, \quad l = n - \beta m. \tag{57}$$

We varied α and β from 0.1 to 1, with a step size 0.05. We set $\lambda = 0.004$, $\mu = 10^{-3}\lambda^{-1}$ for the smoothing-based method, and $\rho = 10^3\lambda$ for the decomposition-based method. L is set to be $\|\boldsymbol{A}\|_2^2 + \frac{\|\boldsymbol{D}\|_2^2}{\mu}$ for smoothing and $\|\boldsymbol{A}\|_2^2 + \rho(1 + \|\boldsymbol{D}\|_2^2)$ for decomposition. For every combination of α and β, we ran a Monte Carlo simulation 50 times. Each algorithm ran for 3000 iterations, and we computed the average reconstruction error. The reconstruction error is defined by $\frac{\|\hat{\boldsymbol{x}} - \boldsymbol{x}\|}{\|\boldsymbol{x}\|}$, in which $\hat{\boldsymbol{x}}$ is the reconstructed signal using smoothing or decomposition and $\boldsymbol{x}$ is the original signal in (1).

The average reconstruction error for smoothing and decomposition are plotted in Figs. 1 and 2, respectively. White pixels present low reconstruction error whereas black pixels mean high error. Evidently, see that with same number of iterations, SFISTA results in a better reconstruction than DFISTA.

B. MRI Image Reconstruction in a Noisy Case

The next numerical experiment was performed on a noisy 256×256 Shepp Logan phantom. The image scale was normalized to $[0, 1]$. The additive noise followed a zero-mean Gaussian distribution with standard deviation $\sigma = 0.001$. Due to the high cost of sampling in MRI, we only observed a limited number of radial lines of the phantom's 2D discrete Fourier transform.

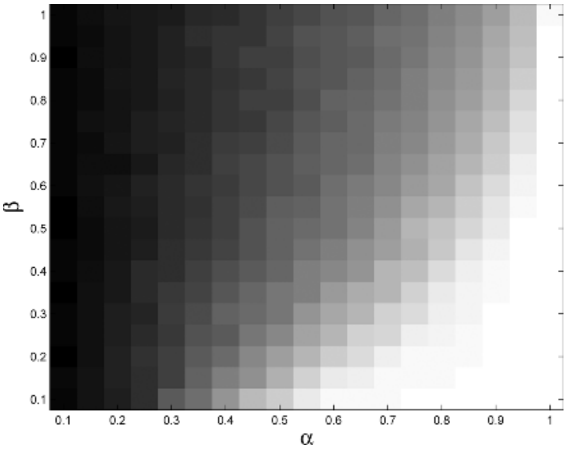

Fig. 1. Reconstruction error of SFISTA.

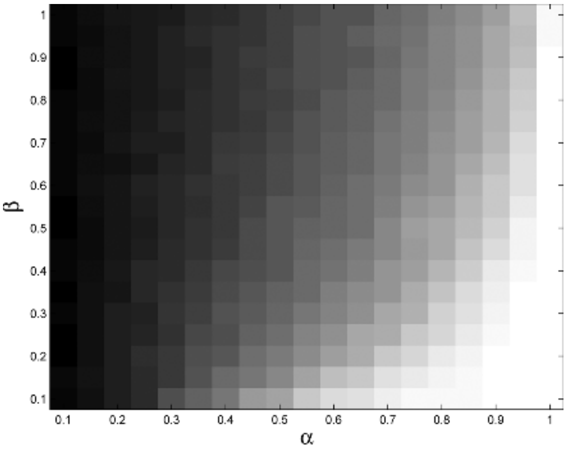

Fig. 2. Reconstruction error of DFISTA.

The matrix $\boldsymbol{D}^*$ consists of all vertical and horizontal gradients, which leads to a sparse $\boldsymbol{D}^*\boldsymbol{x}$. We let $\lambda = 0.001$ in the optimization. We tested this MRI scenario with μ values of $10^{-2}\lambda^{-1}, 10^{-3}\lambda^{-1}, 10^{-4}\lambda^{-1}$ for SFISTA and $\rho = 10^2\lambda, \rho = 10^3\lambda, 10^4\lambda$ for DFISTA. L is set to be $\|\boldsymbol{A}\|_2^2 + \frac{\|\boldsymbol{D}\|_2^2}{\mu}$ for SFISTA and $\|\boldsymbol{A}\|_2^2 + \rho(1 + \|\boldsymbol{D}\|_2^2)$ for DFISTA. We took the samples along 15 radial lines to test these two methods.

In Fig. 3 we plot the objective $\frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2 + \lambda\|\boldsymbol{D}^*\boldsymbol{x}\|_1$ as a function of the iteration number. It can be seen that the objective function of SFISTA decreases more rapidly than DFISTA. Furthermore, with smaller ρ and larger μ, DFISTA and SFISTA converge faster. Then we computed the reconstruction error. Here we see that smaller μ and larger ρ lead to a more accurate reconstruction. We can see that SFISTA converges faster than DFISTA, which follows the convergence results in Section III.

Next, we compared SFISTA with the nonlinear conjugate gradient descend (CGD) algorithm proposed in [23]. The CGD also needs to introduce a smoothing transformation to approximate the term $\|\boldsymbol{D}^*\boldsymbol{x}\|_1$, and in this simulation the Moreau envelop with $\mu = 10^{-4}\lambda^{-1}$ was used to smooth this term. We can see from Fig. 5 that SFISTA converges faster than the CGD in terms of CPU time. CGD is slower because in each iteration, a backtracking line-search is required, which reduces the algorithm efficiency.

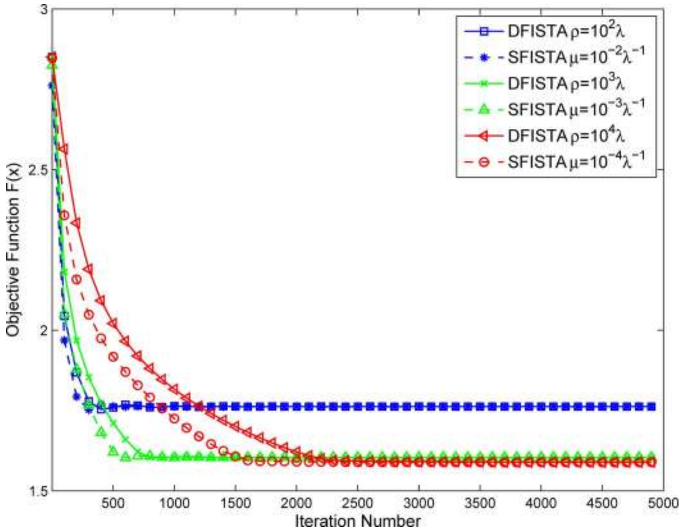

Fig. 3. The objective function for MRI reconstruction on Shepp Logan.

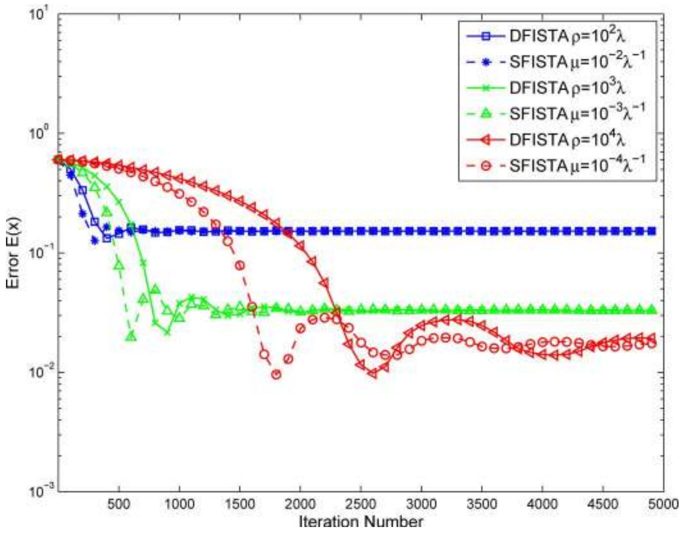

Fig. 4. Reconstruction error for SFISTA and DFISTA with different parameters.

C. Acceleration by Continuation

Algorithm 3: Continuation with SFISTA

Input: $\boldsymbol{x}$, the starting parameter $\mu = \mu_0$, the ending parameter μ_f and $\gamma > 1$.

Step 1. run SFISTA with μ and initial point $\boldsymbol{x}$.

Step 2. Get the solution $\boldsymbol{x}^*$ and let $\boldsymbol{x} = \boldsymbol{x}^*, \mu = \mu/\gamma$.

Until. $\mu \le \mu_f$.

To accelerate convergence and increase the accuracy of reconstruction, we consider continuation on the parameter μ for SFISTA, or on ρ for DFISTA. From Theorem IV.1, we see that smaller μ results in a smaller reconstruction error. At the same time, smaller μ leads to a larger Lipschitz constant $L_{\nabla F}$ in Theorem II.1, and thus results in slower convergence. The idea of continuation is to solve a sequence of similar problems while using the previous solution as a warm start. Taking the smoothing-based method as an example, we can run SFISTA

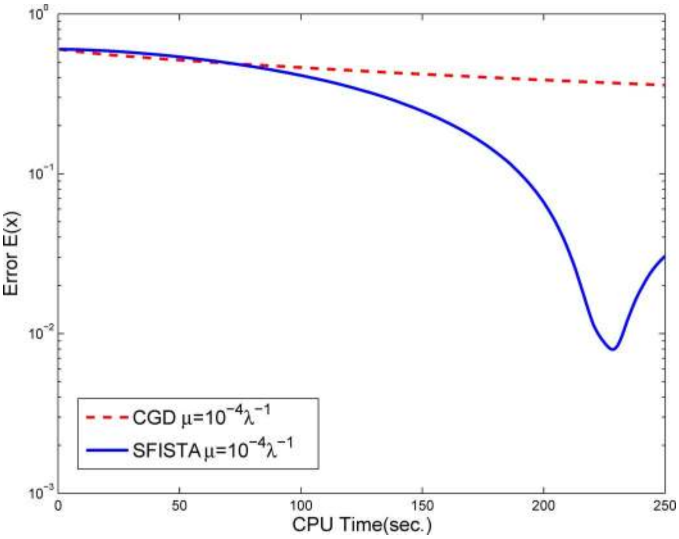

Fig. 5. Reconstruction error for SFISTA and CGD with respect to CPU time.

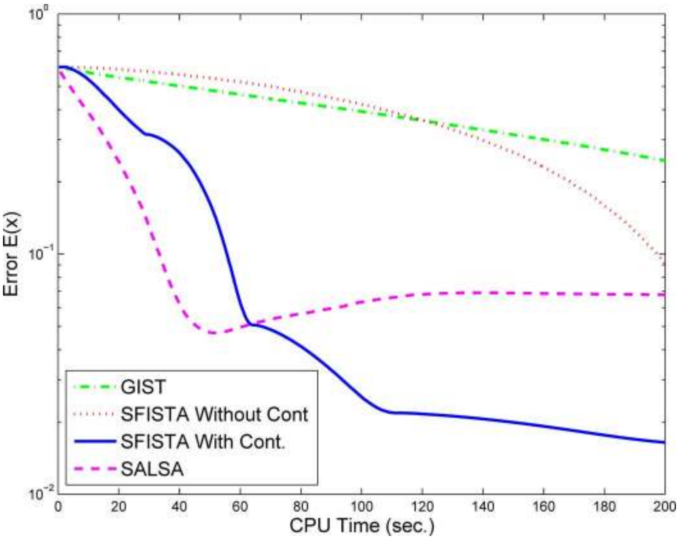

Fig. 6. Convergence comparison among SFISTA with and without continuation, GIST and SALSA.

with $\mu_1 \geq \mu_2 \geq \mu_3, \ldots \geq \mu_f$. The continuation method is given in Algorithm 3. The algorithm for applying continuation on DFISTA is the same.

We tested the algorithm on the Shepp Logan image from the previous subsection with the same setting, using SFISTA with $\mu_f = 10^{-4}\lambda^{-1}$ and standard SFISTA with $\mu = 10^{-4}\lambda^{-1}$. We implemented the generalized iterative soft-thresholding algorithm (GIST) from [20]. We also included an ADMM-based method, i.e., the split augmented Lagrangian shrinkage algorithm (SALSA) [18]. SALSA requires solving the proximal operator of $\|\boldsymbol{D}^*\boldsymbol{x}\|_1$, which is nontrivial. In this simulation, we implemented 40 iterations of the Fast GP algorithm [25] to approximate this proximal operator. Without solving the proximal operator exactly, the ADMM-based method can converge very fast while the accuracy of reconstruction is compromised as we show in Fig. 6. In this figure we plot the reconstruction error for these four algorithms. It also shows that continuation helps speed up the convergence and exhibits better performance then GIST. The reconstructed Shepp Logan phantom using continuation is presented in Fig. 7, with reconstruction error 3.17%.

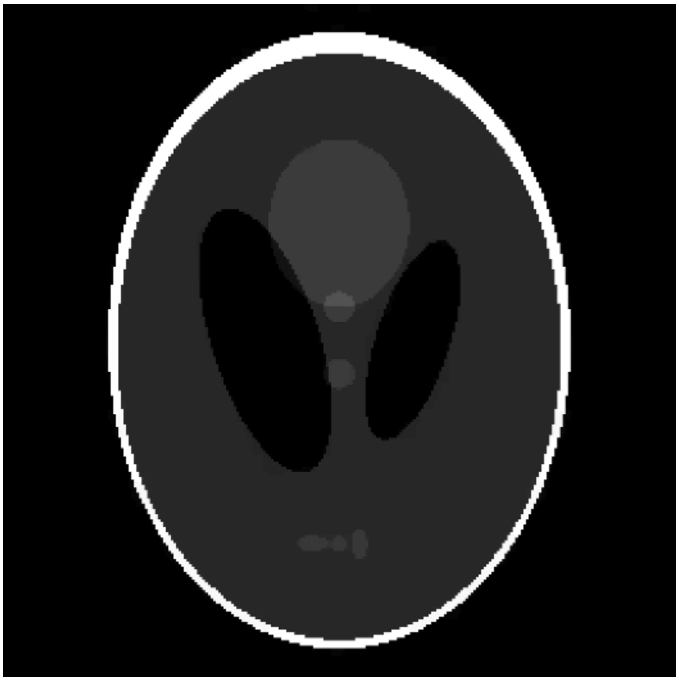

Fig. 7. Reconstructed Shepp Logan with SFISTA using continuation.

VI. Conclusion

In this paper, we proposed methods based on MFISTA to solve the analysis LASSO optimization problem. Since the proximal operator in MFISTA for $\|\boldsymbol{D}^*\boldsymbol{x}\|_1$ does not have a closed-form solution, we presented two methods, SFISTA and DFISTA, using smoothing and decomposition respectively, to transform the original sparse recovery problem into a smooth counterpart. We analyzed the convergence of SFISTA and DFISTA and showed that SFISTA converges faster in general nonsmooth optimization problems. We also derived a bound on the performance for both approaches assuming a tight frame and D-RIP. Our methods were demonstrated via several simulations. With the application of parameter continuation, these two algorithms are suitable to solve large scale problems.

Appendix

Proof of Lemma IV.1: Without loss of generality we assume that $\|\boldsymbol{u}\|_2 = 1$ and $\|\boldsymbol{v}\|_2 = 1$. By the definition of D-RIP, we have

$$\begin{aligned}\mathrm{Re}\langle \boldsymbol{Au}, \boldsymbol{Av}\rangle &= \frac{1}{4}\{\|\boldsymbol{Au}+\boldsymbol{Av}\|_2^2 - \|\boldsymbol{Au}-\boldsymbol{Av}\|_2^2\}\\ &\geq \frac{1}{4}\{(1-\sigma_{2s})\|\boldsymbol{u}+\boldsymbol{v}\|_2^2 - (1+\sigma_{2s})\|\boldsymbol{u}-\boldsymbol{v}\|_2^2\}\\ &= -\sigma_{2s} + \mathrm{Re}\langle \boldsymbol{u}, \boldsymbol{v}\rangle.\end{aligned}$$

Now it is easy to extend this equation to get the desired result.

Proof of Lemma IV.2: From the definition of $\mathcal{T}_j$ we have

$$\|\boldsymbol{D}^*_{\mathcal{T}_j}\boldsymbol{h}\|_2 \leq s^{-\frac{1}{2}}\|\boldsymbol{D}^*_{\mathcal{T}_{j-1}}\boldsymbol{h}\|_1$$

for all $j \ge 2$. Summing $j = 2, 3, \ldots$ leads to

$$\sum_{j\ge 2}\|\boldsymbol{D}^*_{\mathcal{T}_j}\boldsymbol{h}\|_2 \le s^{-\frac{1}{2}}\sum_{j\ge 1}\|\boldsymbol{D}^*_{\mathcal{T}_j}\boldsymbol{h}\|_1 = s^{-\frac{1}{2}}\|\boldsymbol{D}^*_{\mathcal{T}^c}\boldsymbol{h}\|_1. \tag{58}$$

Now, considering the fact that $\boldsymbol{D}$ is a tight frame, i.e., $\boldsymbol{D}\boldsymbol{D}^* = \boldsymbol{I}$, and that the D-RIP holds,

$$\begin{aligned}
&\mathrm{Re}\langle \boldsymbol{A}\boldsymbol{h}, \boldsymbol{A}\boldsymbol{D}\boldsymbol{D}^*_{\mathcal{T}_{01}}\boldsymbol{h}\rangle\\
&= \mathrm{Re}\langle \boldsymbol{A}\boldsymbol{D}\boldsymbol{D}^*_{\mathcal{T}_{01}}\boldsymbol{h}, \boldsymbol{A}\boldsymbol{D}\boldsymbol{D}^*_{\mathcal{T}_{01}}\boldsymbol{h}\rangle + \sum_{j\ge 2}\mathrm{Re}\langle \boldsymbol{A}\boldsymbol{D}\boldsymbol{D}^*_{\mathcal{T}_j}\boldsymbol{h}, \boldsymbol{A}\boldsymbol{D}\boldsymbol{D}^*_{\mathcal{T}_{01}}\boldsymbol{h}\rangle\\
&\ge (1-\sigma_{2s})\|\boldsymbol{D}\boldsymbol{D}^*_{\mathcal{T}_{01}}\boldsymbol{h}\|_2^2 + \sum_{j\ge 2}\mathrm{Re}\langle \boldsymbol{A}\boldsymbol{D}\boldsymbol{D}^*_{\mathcal{T}_j}\boldsymbol{h}, \boldsymbol{A}\boldsymbol{D}\boldsymbol{D}^*_{\mathcal{T}_0}\boldsymbol{h}\rangle\\
&\quad + \sum_{j\ge 2}\mathrm{Re}\langle \boldsymbol{A}\boldsymbol{D}\boldsymbol{D}^*_{\mathcal{T}_j}\boldsymbol{h}, \boldsymbol{A}\boldsymbol{D}\boldsymbol{D}^*_{\mathcal{T}_1}\boldsymbol{h}\rangle
\end{aligned}$$

Using the result from Lemma IV.1, we can bound the last two terms in the above inequality; hence, we derive

$$\begin{aligned}
&\mathrm{Re}\langle \boldsymbol{A}\boldsymbol{h}, \boldsymbol{A}\boldsymbol{D}\boldsymbol{D}^*_{\mathcal{T}_{01}}\boldsymbol{h}\rangle\\
&\ge (1-\sigma_{2s})\|\boldsymbol{D}\boldsymbol{D}^*_{\mathcal{T}_{01}}\boldsymbol{h}\|_2^2 + \sum_{j\ge 2}\mathrm{Re}\langle \boldsymbol{D}\boldsymbol{D}^*_{\mathcal{T}_j}\boldsymbol{h}, \boldsymbol{D}\boldsymbol{D}^*_{\mathcal{T}_0}\boldsymbol{h}\rangle\\
&\quad + \sum_{j\ge 2}\mathrm{Re}\langle \boldsymbol{D}\boldsymbol{D}^*_{\mathcal{T}_j}\boldsymbol{h}, \boldsymbol{D}\boldsymbol{D}^*_{\mathcal{T}_1}\boldsymbol{h}\rangle\\
&\quad - \sigma_{2s}\|\boldsymbol{D}\boldsymbol{D}^*_{\mathcal{T}_0}\boldsymbol{h}\|_2\sum_{j\ge 2}\|\boldsymbol{D}\boldsymbol{D}^*_{\mathcal{T}_j}\boldsymbol{h}\|_2\\
&\quad - \sigma_{2s}\|\boldsymbol{D}\boldsymbol{D}^*_{\mathcal{T}_1}\boldsymbol{h}\|_2\sum_{j\ge 2}\|\boldsymbol{D}\boldsymbol{D}^*_{\mathcal{T}_j}\boldsymbol{h}\|_2\\
&= (1-\sigma_{2s})\|\boldsymbol{D}\boldsymbol{D}^*_{\mathcal{T}_{01}}\boldsymbol{h}\|_2^2 + \mathrm{Re}\left\langle \sum_{j\ge 2}\boldsymbol{D}\boldsymbol{D}^*_{\mathcal{T}_j}\boldsymbol{h}, \boldsymbol{D}\boldsymbol{D}^*_{\mathcal{T}_{01}}\boldsymbol{h}\right\rangle\\
&\quad - \sigma_{2s}(\|\boldsymbol{D}\boldsymbol{D}^*_{\mathcal{T}_0}\boldsymbol{h}\|_2 + \|\boldsymbol{D}\boldsymbol{D}^*_{\mathcal{T}_1}\boldsymbol{h}\|_2)\sum_{j\ge 2}\|\boldsymbol{D}\boldsymbol{D}^*_{\mathcal{T}_j}\boldsymbol{h}\|_2
\end{aligned} \tag{59}$$

By definition of $\mathcal{T}_j$, we have

$$\begin{aligned}
\mathrm{Re}\left\langle \sum_{j\ge 2}\boldsymbol{D}\boldsymbol{D}^*_{\mathcal{T}_j}\boldsymbol{h}, \boldsymbol{D}\boldsymbol{D}^*_{\mathcal{T}_{01}}\boldsymbol{h}\right\rangle &= \mathrm{Re}\langle \boldsymbol{h} - \boldsymbol{D}\boldsymbol{D}^*_{\mathcal{T}_{01}}\boldsymbol{h}, \boldsymbol{D}\boldsymbol{D}^*_{\mathcal{T}_{01}}\boldsymbol{h}\rangle\\
&= \|\boldsymbol{D}^*_{\mathcal{T}_{01}}\boldsymbol{h}\|_2^2 - \|\boldsymbol{D}\boldsymbol{D}^*_{\mathcal{T}_{01}}\boldsymbol{h}\|_2^2.
\end{aligned}$$

Combining this equation with (59) results in

$$\begin{aligned}
&\mathrm{Re}\langle \boldsymbol{A}\boldsymbol{h}, \boldsymbol{A}\boldsymbol{D}\boldsymbol{D}^*_{\mathcal{T}_{01}}\boldsymbol{h}\rangle\\
&\ge \|\boldsymbol{D}\boldsymbol{D}^*_{\mathcal{T}_{01}}\boldsymbol{h}\|_2^2 - \sigma_{2s}\|\boldsymbol{D}\boldsymbol{D}^*_{\mathcal{T}_{01}}\boldsymbol{h}\|_2^2 + \|\boldsymbol{D}^*_{\mathcal{T}_{01}}\boldsymbol{h}\|_2^2 - \|\boldsymbol{D}\boldsymbol{D}^*_{\mathcal{T}_{01}}\boldsymbol{h}\|_2^2\\
&\quad - \sigma_{2s}(\|\boldsymbol{D}\boldsymbol{D}^*_{\mathcal{T}_0}\boldsymbol{h}\|_2 + \|\boldsymbol{D}\boldsymbol{D}^*_{\mathcal{T}_1}\boldsymbol{h}\|_2)\sum_{j\ge 2}\|\boldsymbol{D}\boldsymbol{D}^*_{\mathcal{T}_j}\boldsymbol{h}\|_2.
\end{aligned}$$

Using the fact that when $\boldsymbol{D}$ is a tight frame, $\|\boldsymbol{D}\boldsymbol{D}^*_{\mathcal{T}_{01}}\boldsymbol{h}\|_2 \le \|\boldsymbol{D}^*_{\mathcal{T}_{01}}\boldsymbol{h}\|_2$, we have

$$\begin{aligned}
&\mathrm{Re}\langle \boldsymbol{A}\boldsymbol{h}, \boldsymbol{A}\boldsymbol{D}\boldsymbol{D}^*_{\mathcal{T}_{01}}\boldsymbol{h}\rangle\\
&\ge (1-\sigma_{2s})\|\boldsymbol{D}^*_{\mathcal{T}_{01}}\boldsymbol{h}\|_2^2 - \sigma_{2s}(\|\boldsymbol{D}^*_{\mathcal{T}_0}\boldsymbol{h}\|_2 + \|\boldsymbol{D}^*_{\mathcal{T}_1}\boldsymbol{h}\|_2)\sum_{j\ge 2}\|\boldsymbol{D}^*_{\mathcal{T}_j}\boldsymbol{h}\|_2.
\end{aligned}$$

Since $\|\boldsymbol{D}^*_{\mathcal{T}_0}\boldsymbol{h}\|_2 + \|\boldsymbol{D}^*_{\mathcal{T}_1}\boldsymbol{h}\|_2 \le \sqrt{2}\|\boldsymbol{D}^*_{\mathcal{T}_{01}}\boldsymbol{h}\|_2$ (because $\mathcal{T}_0$ and $\mathcal{T}_1$ are disjoint), we conclude that

$$\begin{aligned}
&\mathrm{Re}\langle \boldsymbol{A}\boldsymbol{h}, \boldsymbol{A}\boldsymbol{D}\boldsymbol{D}^*_{\mathcal{T}_{01}}\boldsymbol{h}\rangle\\
&\quad \ge (1-\sigma_{2s})\|\boldsymbol{D}^*_{\mathcal{T}_{01}}\boldsymbol{h}\|_2^2 - \sqrt{2}\sigma_{2s}\|\boldsymbol{D}^*_{\mathcal{T}_{01}}\boldsymbol{h}\|_2\sum_{j\ge 2}\|\boldsymbol{D}^*_{\mathcal{T}_j}\boldsymbol{h}\|_2,
\end{aligned}$$

which along with inequality (58) yields the desired result given by

$$\begin{aligned}
&\mathrm{Re}\langle \boldsymbol{A}\boldsymbol{h}, \boldsymbol{A}\boldsymbol{D}\boldsymbol{D}^*_{\mathcal{T}_{01}}\boldsymbol{h}\rangle\\
&\quad \ge (1-\sigma_{2s})\|\boldsymbol{D}^*_{\mathcal{T}_{01}}\boldsymbol{h}\|_2^2 - \sqrt{2}s^{-\frac{1}{2}}\sigma_{2s}\|\boldsymbol{D}^*_{\mathcal{T}_{01}}\boldsymbol{h}\|_2\|\boldsymbol{D}^*_{\mathcal{T}^c}\boldsymbol{h}\|_1.
\end{aligned}$$

Proof of Lemma IV.3: The subgradient optimality condition for RALASSO (5) can be stated as

$$\boldsymbol{A}^*(\boldsymbol{A}\hat{\boldsymbol{x}}_\rho - \boldsymbol{b}) + \rho\boldsymbol{D}(\boldsymbol{D}^*\hat{\boldsymbol{x}}_\rho - \hat{\boldsymbol{z}}_\rho) = 0, \tag{60}$$

$$\lambda\boldsymbol{v} + \rho(\hat{\boldsymbol{z}}_\rho - \boldsymbol{D}^*\hat{\boldsymbol{x}}_\rho) = 0, \tag{61}$$

where $\boldsymbol{v}$ is a subgradient of the function $\|\boldsymbol{z}\|_1$ and consequently $\|\boldsymbol{v}\|_\infty \le 1$. Combining (60) and (61), we have

$$\boldsymbol{A}^*(\boldsymbol{A}\hat{\boldsymbol{x}}_\rho - \boldsymbol{b}) = \lambda\boldsymbol{D}\boldsymbol{v}.$$

Multiplying both sides by $\boldsymbol{D}^*$, we get

$$\begin{aligned}
&\|\boldsymbol{D}^*\boldsymbol{A}^*(\boldsymbol{A}\hat{\boldsymbol{x}}_\rho - \boldsymbol{b})\|_\infty\\
&\quad = \lambda\|\boldsymbol{D}^*\boldsymbol{D}\boldsymbol{v}\|_\infty \le \lambda\|\boldsymbol{D}^*\boldsymbol{D}\|_{\infty,\infty} = \lambda\|\boldsymbol{D}^*\boldsymbol{D}\|_{1,1}.
\end{aligned} \tag{62}$$

The first inequality follows from the fact that $\|\boldsymbol{v}\|_\infty \le 1$. With the assumption that $\|\boldsymbol{D}^*\boldsymbol{A}^*\boldsymbol{w}\|_\infty \le \frac{\lambda}{2}$, and the triangle inequality, we have

$$\begin{aligned}
&\|\boldsymbol{D}^*\boldsymbol{A}^*\boldsymbol{A}\boldsymbol{h}\|_\infty\\
&\quad \le \|\boldsymbol{D}^*\boldsymbol{A}^*(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})\|_\infty + \|\boldsymbol{D}^*\boldsymbol{A}^*(\boldsymbol{A}\hat{\boldsymbol{x}}_\rho - \boldsymbol{b})\|_\infty\\
&\quad \le \left(\frac{1}{2} + \|\boldsymbol{D}^*\boldsymbol{D}\|_{1,1}\right)\lambda.
\end{aligned} \tag{63}$$

Proof of Lemma IV.4: Since $\hat{\boldsymbol{x}}_\rho$ and $\hat{\boldsymbol{z}}_\rho$ solve the optimization problem RALASSO (5), we have,

$$\begin{aligned}
&\frac{1}{2}\|\boldsymbol{A}\hat{\boldsymbol{x}}_\rho - \boldsymbol{b}\|_2^2 + \lambda\|\hat{\boldsymbol{z}}_\rho\|_1 + \frac{1}{2}\rho\|\boldsymbol{D}^*\hat{\boldsymbol{x}}_\rho - \hat{\boldsymbol{z}}_\rho\|_2^2\\
&\qquad \le \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2 + \lambda\|\boldsymbol{D}^*\boldsymbol{x}\|_1.
\end{aligned}$$

Since $\boldsymbol{b} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{w}$ and $\boldsymbol{h} = \hat{\boldsymbol{x}}_\rho - \boldsymbol{x}$, it follows that

$$\begin{aligned}
&\frac{1}{2}\|\boldsymbol{A}\boldsymbol{h} - \boldsymbol{w}\|_2^2 + \lambda\|\hat{\boldsymbol{z}}_\rho\|_1 + \frac{1}{2}\rho\|\boldsymbol{D}^*\hat{\boldsymbol{x}}_\rho - \hat{\boldsymbol{z}}_\rho\|_2^2\\
&\qquad \le \frac{1}{2}\|\boldsymbol{w}\|_2^2 + \lambda\|\boldsymbol{D}^*\boldsymbol{x}\|_1.
\end{aligned}$$

Expanding and rearranging the terms in the above equation, we get

$$\begin{aligned}
&\frac{1}{2}\|\boldsymbol{A}\boldsymbol{h}\|_2^2 + \lambda\|\hat{\boldsymbol{z}}_\rho\|_1 + \frac{1}{2}\rho\|\boldsymbol{D}^*\hat{\boldsymbol{x}}_\rho - \hat{\boldsymbol{z}}_\rho\|_2^2\\
&\qquad \le \mathrm{Re}\langle \boldsymbol{A}\boldsymbol{h}, \boldsymbol{w}\rangle + \lambda\|\boldsymbol{D}^*\boldsymbol{x}\|_1,
\end{aligned}$$

Using (61) to replace the terms with $\hat{\boldsymbol{z}}_\rho$, we have

$$\frac{1}{2}\|\boldsymbol{A}\boldsymbol{h}\|_2^2 + \lambda \left\| \boldsymbol{D}^*\hat{\boldsymbol{x}}_\rho - \frac{\lambda}{\rho}\boldsymbol{v} \right\|_1 + \frac{1}{2}\rho \left\| \frac{\lambda}{\rho}\boldsymbol{v} \right\|_2^2 \le \mathrm{Re}\langle \boldsymbol{A}\boldsymbol{h}, \boldsymbol{w}\rangle + \lambda\|\boldsymbol{D}^*\boldsymbol{x}\|_1.$$

Since $\|\boldsymbol{D}^*\hat{\boldsymbol{x}}_\rho - \frac{\lambda}{\rho}\boldsymbol{v}\|_1 \ge \|\boldsymbol{D}^*\hat{\boldsymbol{x}}_\rho\|_1 - \frac{\lambda}{\rho}\|\boldsymbol{v}\|_1$, we have

$$\begin{aligned} &\frac{1}{2}\|\boldsymbol{A}\boldsymbol{h}\|_2^2 + \lambda\|\boldsymbol{D}^*\hat{\boldsymbol{x}}_\rho\|_1 \\ &\le \frac{\lambda^2}{\rho}\|\boldsymbol{v}\|_1 - \frac{\lambda^2}{2\rho}\|\boldsymbol{v}\|_2^2 + \mathrm{Re}\langle \boldsymbol{A}\boldsymbol{h}, \boldsymbol{w}\rangle + \lambda\|\boldsymbol{D}^*\boldsymbol{x}\|_1 \\ &\le \frac{\lambda^2 p}{2\rho} + \mathrm{Re}\langle \boldsymbol{A}\boldsymbol{h}, \boldsymbol{w}\rangle + \lambda\|\boldsymbol{D}^*\boldsymbol{x}\|_1. \end{aligned} \tag{64}$$

The second inequality follows from the fact that $\frac{\lambda^2}{\rho}\|\boldsymbol{v}\|_1 - \frac{\lambda^2}{2\rho}\|\boldsymbol{v}\|_2^2$ is maximized when every element of $\boldsymbol{v} \in \mathbb{R}^p$ is 1. Now, with the assumption that $\boldsymbol{D}$ is a tight frame, we have the following relation:

$$\begin{aligned} \mathrm{Re}\langle \boldsymbol{A}\boldsymbol{h}, \boldsymbol{w}\rangle + \lambda\|\boldsymbol{D}^*\boldsymbol{x}\|_1 &= \mathrm{Re}\langle \boldsymbol{D}^*\boldsymbol{h}, \boldsymbol{D}^*\boldsymbol{A}^*\boldsymbol{w}\rangle + \lambda\|\boldsymbol{D}^*\boldsymbol{x}\|_1 \\ &\le \|\boldsymbol{D}^*\boldsymbol{h}\|_1\|\boldsymbol{D}^*\boldsymbol{A}^*\boldsymbol{w}\|_\infty + \lambda\|\boldsymbol{D}^*\boldsymbol{x}\|_1. \end{aligned}$$

This inequality follows from the fact that $\mathrm{Re}\langle \boldsymbol{x}, \boldsymbol{y}\rangle \le |\langle \boldsymbol{x}, \boldsymbol{y}\rangle| \le \|\boldsymbol{x}\|_1\|\boldsymbol{y}\|_\infty$. Using the assumption that $\|\boldsymbol{D}^*\boldsymbol{A}^*\boldsymbol{w}\|_\infty \le \frac{\lambda}{2}$, we get

$$\mathrm{Re}\langle \boldsymbol{A}\boldsymbol{h}, \boldsymbol{w}\rangle + \lambda\|\boldsymbol{D}^*\boldsymbol{x}\|_1 \le \frac{\lambda}{2}\|\boldsymbol{D}^*\boldsymbol{h}\|_1 + \lambda\|\boldsymbol{D}^*\boldsymbol{x}\|_1. \tag{65}$$

Applying inequalities (64) and (65), we have

$$\begin{aligned} \lambda\|\boldsymbol{D}^*\hat{\boldsymbol{x}}_\rho\|_1 &\le \frac{1}{2}\|\boldsymbol{A}\boldsymbol{h}\|_2^2 + \lambda\|\boldsymbol{D}^*\hat{\boldsymbol{x}}_\rho\|_1 \\ &\le \frac{\lambda^2}{2\rho}p + \mathrm{Re}\langle \boldsymbol{A}\boldsymbol{h}, \boldsymbol{w}\rangle + \lambda\|\boldsymbol{D}^*\boldsymbol{x}\|_1 \\ &\le \frac{\lambda^2}{2\rho}p + \frac{\lambda}{2}\|\boldsymbol{D}^*\boldsymbol{h}\|_1 + \lambda\|\boldsymbol{D}^*\boldsymbol{x}\|_1, \end{aligned}$$

which is the same as,

$$\|\boldsymbol{D}^*\hat{\boldsymbol{x}}_\rho\|_1 \le \frac{\lambda}{2\rho}p + \frac{1}{2}\|\boldsymbol{D}^*\boldsymbol{h}\|_1 + \|\boldsymbol{D}^*\boldsymbol{x}\|_1.$$

Since we have $\boldsymbol{h} = \hat{\boldsymbol{x}}_\rho - \boldsymbol{x}$, it follows that

$$\|\boldsymbol{D}^*\boldsymbol{h} + \boldsymbol{D}^*\boldsymbol{x}\|_1 \le \frac{\lambda}{2\rho}p + \frac{1}{2}\|\boldsymbol{D}^*\boldsymbol{h}\|_1 + \|\boldsymbol{D}^*\boldsymbol{x}\|_1,$$

and hence

$$\begin{aligned} &\|\boldsymbol{D}_\mathcal{T}^*\boldsymbol{h} + \boldsymbol{D}_\mathcal{T}^*\boldsymbol{x}\|_1 + \|\boldsymbol{D}_{\mathcal{T}^c}^*\boldsymbol{h} + \boldsymbol{D}_{\mathcal{T}^c}^*\boldsymbol{x}\|_1 \\ &\le \frac{\lambda}{2\rho}p + \frac{1}{2}\|\boldsymbol{D}_\mathcal{T}^*\boldsymbol{h}\|_1 + \frac{1}{2}\|\boldsymbol{D}_{\mathcal{T}^c}^*\boldsymbol{h}\|_1 + \|\boldsymbol{D}_\mathcal{T}^*\boldsymbol{x}\|_1 + \|\boldsymbol{D}_{\mathcal{T}^c}^*\boldsymbol{x}\|_1. \end{aligned}$$

Applying the triangle inequality to the left handside of above inequality, we results in

$$\begin{aligned} &-\|\boldsymbol{D}_\mathcal{T}^*\boldsymbol{h}\|_1 + \|\boldsymbol{D}_\mathcal{T}^*\boldsymbol{x}\|_1 + \|\boldsymbol{D}_{\mathcal{T}^c}^*\boldsymbol{h}\|_1 - \|\boldsymbol{D}_{\mathcal{T}^c}^*\boldsymbol{x}\|_1 \\ &\le \frac{\lambda}{2\rho}p + \frac{1}{2}\|\boldsymbol{D}_\mathcal{T}^*\boldsymbol{h}\|_1 + \frac{1}{2}\|\boldsymbol{D}_{\mathcal{T}^c}^*\boldsymbol{h}\|_1 + \|\boldsymbol{D}_\mathcal{T}^*\boldsymbol{x}\|_1 + \|\boldsymbol{D}_{\mathcal{T}^c}^*\boldsymbol{x}\|_1. \end{aligned}$$

After rearranging the terms, we have the following cone constraint,

$$\|\boldsymbol{D}_{\mathcal{T}^c}^*\boldsymbol{h}\|_1 \le \frac{\lambda}{\rho}p + 3\|\boldsymbol{D}_\mathcal{T}^*\boldsymbol{h}\|_1 + 4\|\boldsymbol{D}_{\mathcal{T}^c}^*\boldsymbol{x}\|_1. \tag{66}$$

References

[1] E. J. Candès and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 21–30, Mar. 2008.

[2] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[3] E. J. Candès and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?," *IEEE Trans. Inf. Theory*, vol. 52, no. 12, pp. 5406–5425, Dec. 2006.

[4] Y. C. Eldar and G. Kutyniok, *Compressed Sensing: Theory and Applications*. Cambridge, U.K.: Cambridge Univ. Press, 2012.

[5] T. Blumensath and M. E. Davies, "Iterative hard thresholding for compressed sensing," *Appl. Comput. Harmon. Anal.*, vol. 27, no. 3, pp. 265–274, 2009.

[6] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 4655–4666, Dec. 2007.

[7] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Rev.*, vol. 43, no. 3, pp. 129–159, Mar. 2001.

[8] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer, 2009.

[9] E. J. Candès, "The restricted isometry property and its implications for compressed sensing," *C. R. Acad. Sci. Paris, Ser. I*, vol. 346, pp. 589–592, 2008.

[10] P. Bickel, Y. Ritov, and A. Tsybakov, "Simultaneous analysis of LASSO and Dantzig selector," *Ann. Statist.*, vol. 37, pp. 1705–1732, 2009.

[11] T. Blumensath and M. E. Davies, "Normalized iterative hard thresholding: Guaranteed stability and performance," *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 2, pp. 298–309, 2010.

[12] E. J. Candès, Y. C. Eldar, D. Needell, and P. Randall, "Compressed sensing with coherent and redundant dictionaries," *Appl. Comput. Harmon. Anal.*, vol. 31, pp. 59–73, Jul. 2011.

[13] S. Nam, M. E. Davies, M. Elad, and R. Gribonval, "The cosparse analysis model and algorithms," *Appl. Comput. Harmon. Anal.*, vol. 34, no. 1, pp. 30–56, 2013.

[14] T. Peleg and M. Elad, "Performance guarantees of the thresholding algorithm for the cosparse analysis model," *IEEE Trans. Ind. Informat.*, vol. 59, no. 3, pp. 1832–1845, 2013.

[15] R. Giryes, S. Nam, M. Elad, R. Gribonval, and M. E. Davies, "Greedy-like algorithms for the cosparse analysis model," *Linear Algebra Appl.*, vol. 441, pp. 22–60, Jan. 2014.

[16] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[17] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via alternating direction method of multipliers," in *Found. Trends Mach. Learn.*, 2010, vol. 3, pp. 1–122.

[18] M. V. Afonso, J. M. Bioucas-Dias, and M. A. T. Figueiredo, "Fast image recovery using variable splitting and constrained optimization," *IEEE Trans. Image Process.*, vol. 19, no. 9, pp. 2345–2356, 2010.

[19] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Imag. Sci.*, vol. 2, no. 1, pp. 183–202, Jan. 2009.

[20] I. Loris and C. Verhoeven, "On a generalization of the iterative soft-thresholding algorithm for the case of non-separable penalty," *Inverse Problems*, vol. 27, no. 12, 2011.

[21] S. Becker, J. Bobin, and E. J. Candès, "NESTA: A fast and accurate first-order method for sparse recovery," *SIAM J. Imag. Sci.*, vol. 4, no. 1, pp. 1–39.

[22] Y. E. Nesterov, "Smooth minimization of non-smooth functions," *Math. Program.*, vol. 103, no. 1, pp. 127–152, 2005.

[23] M. Lustig, D. Donoho, and J. M. Pauly, "Sparse MRI: The application of compressed sensing for rapid MR imaging," *Magn. Reson. Med.*, vol. 58, no. 6, pp. 1182–1195, 2007.

[24] A. Beck and M. Teboulle, "Smoothing and first order methods: A unified framework," *SIAM J. Optim.*, vol. 22, no. 2, pp. 557–580, 2012.

[25] A. Beck and M. Teboulle, "Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems," *IEEE Trans. Image Process.*, vol. 18, no. 11, pp. 2419–2434, Nov. .

[26] R. Courant, "Variational methods for the solution of problems with equilibrium and vibration," *Bull. Amer. Math. Soc.*, vol. 49, pp. 1–23, 1943.

[27] Y. Wang, J. Yang, W. Yin, and Y. Zhang, "A new alternating minimization algorithm for total variation image reconstruction," *SIAM J. Imag. Sci.*, vol. 1, no. 3, pp. 248–272, 2008.

[28] J. Lin and S. Li, "Sparse recovery with coherent tight frame via analysis Dantzig selector and analysis lasso," *Appl. Comput. Harmon. Anal.*, Oct. 2013 [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1063520313000948

[29] J. J. Moreau, "Proximitéet dualité dans un espace Hilbertien," (in French) *Bull. Soc. Math. France*, vol. 93, pp. 273–299, 1965.

[30] J. J. Moreau, "Fonctions convexes duales et points proximaux dans un espace Hilbertien," (in French) *Comptes Rendus de l'Académie des Sci. (Paris), Série A*, vol. 255, pp. 2897–2899, 1962.

[31] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Space*. New York, NY, USA: Springer, 2011.

[32] A. Beck and M. Teboulle, "Gradient-based algorithms with applications to signal recovery problems," in *Convex Optimization in Signal Processing and Communications*, D. Palomar and Y. Eldar, Eds. Cambridge, U.K.: Cambridge Univ. Press, 2009, pp. 139–162.

[33] Y. E. Nesterov, "A method for solving the convex programming problem with convergence rate $O(1/k^2)$," *Dokl. Akad. Nauk SSSR*, vol. 269, no. 3, pp. 543–547, 1983.

[34] P. J. Huber, "Robust estimation of a location parameter," *Ann. Math. Statist.*, vol. 35, pp. 73–101, 1964.

[35] D. Needell and J. A. Tropp, "Cosamp: Iterative signal recovery from noisy samples," *Appl. Comput. Harmon. Anal.*, vol. 26, no. 3, pp. 301–321, 2008.

[36] M. A. T. Figueiredo, R. D. Nowak, and S. J. Wright, "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems," *IEEE J. Sel. Topics Signal Process.*, vol. 1, no. 4, pp. 586–597, 2007.

Zhao Tan (S'12) received the B.Sc. degree in electronic information science and technology from Fudan University of China in 2010. He received M.SC. degree in electrical engineering from Washington University in 2013. He is currently working towards a Ph.D. degree with the Preston M. Green Department of Electrical and Systems Engineering at Washington University in St. Louis, under the guidance of Dr. Arye Nehorai. His research interests are mainly in the areas of optimization algorithms, compressed sensing, sensor arrays, radar signal processing, and state estimation and scheduling in smart grid.

Yonina C. Eldar (S'98–M'02–SM'07–F'12) received the B.Sc. degree in physics and the B.Sc. degree in electrical engineering both from Tel-Aviv University (TAU), Tel-Aviv, Israel, in 1995 and 1996, respectively, and the Ph.D. degree in electrical engineering and computer science from the Massachusetts Institute of Technology (MIT), Cambridge, in 2002. From January 2002 to July 2002, she was a Postdoctoral Fellow at the Digital Signal Processing Group at MIT. She is currently a Professor in the Department of Electrical Engineering at the Technion–Israel Institute of Technology, Haifa and holds the Edwards Chair in Engineering. She is also a Research Affiliate with the Research Laboratory of Electronics at MIT and a Visiting Professor at Stanford University, Stanford, CA. Her research interests are in the broad areas of statistical signal processing, sampling theory and compressed sensing, optimization methods, and their applications to biology and optics.

Dr. Eldar was in the program for outstanding students at TAU from 1992 to 1996. In 1998, she held the Rosenblith Fellowship for study in electrical engineering at MIT, and in 2000, she held an IBM Research Fellowship. From 2002 to 2005, she was a Horev Fellow of the Leaders in Science and Technology program at the Technion and an Alon Fellow. In 2004, she was awarded the Wolf Foundation Krill Prize for Excellence in Scientific Research, in 2005 the Andre and Bella Meyer Lectureship, in 2007 the Henry Taub Prize for Excellence in Research, in 2008 the Hershel Rich Innovation Award, the Award for Women with Distinguished Contributions, the Muriel & David Jacknow Award for Excellence in Teaching, and the Technion Outstanding Lecture Award, in 2009 the Technion's Award for Excellence in Teaching, in 2010 the Michael Bruno Memorial Award from the Rothschild Foundation, and in 2011 the Weizmann Prize for Exact Sciences. In 2012 she was elected to the Young Israel Academy of Science and to the Israel Committee for Higher Education, and elected an IEEE Fellow. In 2013 she received the Technion's Award for Excellence in Teaching, the Hershel Rich Innovation Award, and the IEEE Signal Processing Technical Achievement Award. She received several best paper awards together with her research students and colleagues. She is a Signal Processing Society Distinguished Lecturer, and Editor in Chief of Foundations and Trends in Signal Processing. In the past, she was a member of the IEEE Signal Processing Theory and Methods and Bio Imaging Signal Processing technical committees, and served as an associate editor for the IEEE Transactions on Signal Processing, the EURASIP *Journal of Signal Processing*, the SIAM *Journal on Matrix Analysis and Applications*, and the SIAM *Journal on Imaging Sciences*.

Amir Beck received the B.Sc. degree in pure mathematics *(cum laude)* in 1991, the M.Sc. degree in operations research *(summa cum laude)*, and the Ph.D. degree in operations research all from Tel Aviv University (TAU).

From 2003 to 2005, he was a Postdoctoral Fellow at the Minerva Optimization Center, Technion, Haifa, Israel. He is currently an Associate Professor in the Department of Industrial Engineering at The Technion Israel Institute of Technology. His research interests are in continuous optimization, including theory, algorithmic analysis, and its applications. He has published numerous papers and has given invited lectures at international conferences. He was awarded the Salomon Simon Mani Award for Excellence in Teaching and the Henry Taub Research Prize award. He is in the editorial board of *Mathematics of Operations Research, Operations Research* and *Journal of Optimization Theory and Applications.* His research has been supported by various funding agencies, including the Israel Science Foundation, the German-Israeli Foundation, the Binational US-Israel foundation, the Israeli Science and Energy Ministries and the European community FP6 program.

Arye Nehorai (S'80–M'83–SM'90–F'94) is the Eugene and Martha Lohman Professor and Chair of the Preston M. Green Department of Electrical and Systems Engineering (ESE), Professor in the Department of Biomedical Engineering (by courtesy) and in the Division of Biology and Biomedical Studies (DBBS) at Washington University in St. Louis (WUSTL). He serves as Director of the Center for Sensor Signal and Information Processing at WUSTL. Under his leadership as department chair, the undergraduate enrollment has more than tripled in the last four years. Earlier, he was a faculty member at Yale University and the University of Illinois at Chicago. He received the B.Sc. and M.Sc. degrees from the Technion, Israel and the Ph.D. from Stanford University, California.

Dr. Nehorai served as Editor-in-Chief of the IEEE Transactions on Signal Processing from 2000 to 2002. From 2003 to 2005 he was the Vice President (Publications) of the IEEE Signal Processing Society (SPS), the Chair of the Publications Board, and a member of the Executive Committee of this Society. He was the founding editor of the special columns on Leadership Reflections in *IEEE Signal Processing Magazine* from 2003 to 2006.

Dr. Nehorai received the 2006 IEEE SPS Technical Achievement Award and the 2010 IEEE SPS Meritorious Service Award. He was elected Distinguished Lecturer of the IEEE SPS for a term lasting from 2004 to 2005. He received several best paper awards in IEEE journals and conferences. In 2001 he was named University Scholar of the University of Illinois. Dr. Nehorai was the Principal Investigator of the Multidisciplinary University Research Initiative (MURI) project titled Adaptive Waveform Diversity for Full Spectral Dominance from 2005 to 2010. He is a Fellow of the IEEE since 1994, Fellow of the Royal Statistical Society since 1996, and Fellow of AAAS since 2012.