Analysis and Applications, Vol. 21, No. 4 (2023) 901–929

DOI: 10.1142/S0219530522500105

World Scientific
www.worldscientific.com

Sparse regularization with the ℓ_0 norm

Yuesheng Xu
Department of Mathematics and Statistics
Old Dominion University, Norfolk, Virginia 23529, USA
y1xu@odu.edu

Received 23 November 2021
Accepted 30 May 2022
Published 23 July 2022

We consider a minimization problem whose objective function is the sum of a fidelity term, not necessarily convex, and a regularization term defined by a positive regularization parameter λ multiple of the ℓ_0 norm composed with a linear transform. This problem has wide applications in compressed sensing, sparse machine learning and image reconstruction. The goal of this paper is to understand what choices of the regularization parameter can dictate the level of sparsity under the transform for a global minimizer of the resulting regularized objective function. This is a critical issue but it has been left unaddressed. We address it from a geometric viewpoint with which the sparsity partition of the image space of the transform is introduced. Choices of the regularization parameter are specified to ensure that a global minimizer of the corresponding regularized objective function achieves a prescribed level of sparsity under the transform. Results are obtained for the spacial sparsity case in which the transform is the identity map, a case that covers several applications of practical importance, including machine learning, image/signal processing and medical image reconstruction.

Keywords: Sparse regularization; sparse optimization; the ℓ_0 norm.

Mathematics Subject Classification 2020: 90C26, 90C30

1. Introduction

The aim of this work is to understand a global minimizer of regularization problems whose objective functions have the form of a fidelity term plus a regularization term involving the ℓ_0 norm. Regularization problems of this type appear frequently in recent studies of machine learning [16, 17, 21, 29], computer graphics [8, 27], signal processing [6, 15, 35], image processing [24, 25, 34], medical imaging [37] and statistics [10, 36]. Many published results have demonstrated that the use of the ℓ_0 norm in regularization models promotes sparsity for the regularized solutions or the transformed regularized solutions. Most of the existing works focus on developing numerical algorithms and considering convergence issues of the developed algorithms. It remains to be understood how choices of the regularization

parameter balance the sparsity of a global minimizer of the regularization problem and its approximation to a global minimizer of the fidelity function. It is the goal of this paper to provide mathematical understanding on how the use of the ℓ_0 norm as a regularization term promotes sparsity of the regularized solutions or the transformed regularized solutions.

We now describe precisely the problem to be considered in this paper. Let d be a fixed positive integer. For $x \in \mathbb{R}^d$, we use $\|x\|_0$ to denote the number of the nonzero components of x. Although $\|\cdot\|_0$ does not satisfy the axiom of vector norms, it is widely referred to as the ℓ_0 norm in the sparse optimization community. We follow the custom of the community to call it the ℓ_0 norm. Let m be another positive integer, which may be equal to d or may be different from d. Suppose that $g : \mathbb{R}^m \to \mathbb{R}$ is a given function and M is a real $d \times m$ matrix. For a parameter $\lambda > 0$, we define the function

$$f(x) := g(x) + \lambda \|Mx\|_0, \quad x \in \mathbb{R}^m \tag{1.1}$$

and consider the related regularization problem

$$\min\{f(x) : x \in \mathbb{R}^m\}. \tag{1.2}$$

Here, $\lambda > 0$ is a regularization parameter. Its choices may impose sparsity of a global minimizer of the corresponding function (1.1). Clearly, the function f defined by Eq. (1.1) depends on the parameter λ and the transform matrix M. Although for conciseness of notation, we do not label the dependence of f on λ or M in its notation, we always assume that f depends on these quantities.

In the context of regularization, the function g appearing in (1.1) is the data fidelity term derived from a linear [12] or nonlinear ill-posed problem [9], and as well as a fidelity term plus a usual ℓ_2 regularization [30, 31]. It may also describe a network [13] in machine learning. For more linear ill-posed problems, see [3–5]. The function g that appears in application is often convex, (for example, the least squares error). It can also be non-convex. For instance, fidelity terms for deep learning are non-convex [28]. It can be differentiable or non-differentiable. In this paper, in order to enlarge the applicability of the established theory, we consider a wide class of fidelity terms g, without imposing convexity or differentiability.

The matrix M that appears in the regularization term is often chosen as a mathematical transform such as a discrete cosine transform [26], a wavelet transform [7, 14, 18–20] or a framelet transform [2, 22], depending on specific applications. It can also be a difference matrix (for example, the ℓ_0-TV). For TV-regularization, the readers are referred to [23]. The matrix M does not have to be a square matrix. However, we confine ourselves to matrices of full rank, since most of mathematical transforms used frequently in applications have this property and the case with matrices of arbitrary rank may be treated by employing the singular value decomposition, on which we will comment at the end of the last section.

The regularization problem (1.2) often raises in the scenarios that the function g has a global minimizer which itself may not be sparse while a sparse minimizer

is desirable. Bringing forward such a model enables us to find a global minimizer of f having the desired sparsity under the transform while keeping it as close to the global minimizer of g as possible. A desirable solution of the problem (1.2) is the one achieving the desired sparsity and being close to the global minimizer of the fidelity term g. For this reason, we shall assume that the function g has a global minimizer in $\mathbb{R}^m$.

We are also interested in minimization problems of sparse regularization in the spacial domain, that is, the special case of (1.1) with $d = m$ and $M = I$, the identity matrix. In this case, the function f has the spacial form

$$f(x) := g(x) + \lambda \|x\|_0, \quad x \in \mathbb{R}^d. \tag{1.3}$$

Although the model (1.3) has its practical importance, we shall not present special results for this case since they can be obtained from general results by restricting $M = I$.

Motivated from *approximately sparse* regularization such as regularization with the envelope of the ℓ_0-norm and capped-ℓ_1, we introduce the function

$$f(x, y) := g(x, y) + \lambda \|x\|_0, \quad \text{for all } (x, y) \in \mathbb{R}^d \times \mathbb{R}^{d'}, \tag{1.4}$$

where $g : \mathbb{R}^d \times \mathbb{R}^{d'} \to \mathbb{R}$ is a continuous function, not necessarily convex. Typical examples of function f in the form (1.4) include the objective functions in wavelet inpainting with the ℓ_0 sparse regularization [25], inverting an incomplete Fourier transform [33] and medical image reconstruction [37]. In these applications, the function g is a sum of two or three convex functions which measure the data fidelity and define other convex constraints. We shall study what choices of the positive parameter λ will balance the sparsity of global minimizers of function f and its approximation to global minimizers of function g. We do not intend to provide practical methods for choices of the regularization parameter λ, and rather, we supply a mathematical understanding of the relation among choices of the regularization parameter, and global/local minimizers of the two functions f and g. We are also interested in understanding the relation between local minimizers of these two functions when the regularization parameter λ is fixed.

Our key approach is the understanding of the "surface" geometry of the function f defined by (1.1) or (1.3). When $\lambda = 0$, f clearly reduces to the function g. We regard the surface determined by the function g as the original "landscape" and imagine its animation controlled by the parameter $\lambda > 0$. At the moment when we start to increase the value of the parameter λ from 0 to a positive number, the original "landscape" begins to change like vertical fractures of the earth crust during an earthquake. The parts of the landscape corresponding to $Mx = 0$ will stay in their original positions and other parts will lift upward according to $\lambda \|Mx\|_0$. This geometry motivates us to partition the space $\mathbb{R}^d$ according to the values of the ℓ_0 norm of the vectors in the space, that is, the sparsity levels. This sparsity partition of the Euclidean space will enable us to understand how the value of the parameter

λ will determine the sparsity level of a global minimizer of f. We shall introduce the sparsity partition of the space $\mathbb{R}^d$, the image space of the transform M, and understand how this partition will result in a partition of the preimage space $\mathbb{R}^m$. Through these partitions we shall be able to visualize the animation as the value of the parameter λ increases. As a result, we can clearly determine how large the value of λ will be in order to achieve a desired level of sparsity for a global minimizer of f and at the same time to keep the minimizer as close to the global minimizer of the function g as possible.

We organize this paper in five sections. In Sec. 2, we introduce a partition of the image space of a transform M according to the levels of sparsity and consider its corresponding partition of the preimage space. We study both algebraic and topological properties of the sets in these partitions. We devote Sec. 3 to a study of choices of the parameter $\lambda > 0$ that ensure desired levels of sparsity under the transform of a global minimizer of function f having the form (1.1). Several necessary conditions of a global minimizer of f are presented. In Sec. 4, for functions f having the form (1.4), we investigate the same issues as those considered in Sec. 3 for function f in the form (1.1). We also present a relation between local minimizers of minimization problem (4.1) and its reduced minimization problem without the term involving the ℓ_0-norm. We briefly discuss in Sec. 5 extension of the results presented in Secs. 3 and 4 involving matrix M and potential practical uses of the main results of this paper.

2. Sparsity Partition of the Euclidean Space

We introduce in this section a partition of the space $\mathbb{R}^d$, the image space of the linear transform M, according to levels of sparsity, and study its corresponding partition of the preimage space $\mathbb{R}^m$. For the purpose of understanding the sparsity of a global minimizer of the function f defined by (1.1), we present algebraic and topological properties of the sets in the partitions.

It is convenient to introduce the level of sparsity for a vector in $\mathbb{R}^d$. To this end, for a positive integer d, we define two index sets $\mathbb{N}_d := \{1, 2, \ldots, d\}$ and $\mathbb{Z}_d := \{0, 1, \ldots, d-1\}$. Precisely, a vector $x \in \mathbb{R}^d$ is said to have sparsity of level $\ell \in \mathbb{Z}_{d+1}$ if x has exactly ℓ number of nonzero components. Clearly, the zero vector has sparsity of level 0 and a vector whose components are all nonzero have sparsity of level d. Vectors having sparsity of level d are not sparse. Sparse vectors are those located on the coordinate axes or coordinate planes of space $\mathbb{R}^d$. For example, in $\mathbb{R}^3$, vectors on the three coordinate axes but not at the origin have sparsity of level 1, vectors on the three coordinate planes but not on the three coordinate axes have sparsity of level 2 and vectors not on the three coordinate planes have sparsity of level 3. Most vectors in the space $\mathbb{R}^d$ are not sparse. In fact, the set of the sparse vectors in $\mathbb{R}^d$ has zero measure.

We now define the sparsity partition of $\mathbb{R}^d$. We need the canonical basis for the space $\mathbb{R}^d$. For each $j \in \mathbb{N}_d$, by $e_j \in \mathbb{R}^d$, we denote the unit vector with 1 for the

jth component and 0 otherwise. The vectors e_j, $j \in \mathbb{N}_d$, form the canonical basis for $\mathbb{R}^d$. Let

$$A_0 := \{0 \in \mathbb{R}^d\},$$

$$A_\ell := \left\{ \sum_{j \in \mathbb{N}_\ell} x_{k_j} e_{k_j} : x_{k_j} \in \mathbb{R}\backslash\{0\}, \text{ for } 1 \le k_1 < k_2 < \cdots < k_\ell \le d \right\}, \quad \text{for } \ell \in \mathbb{N}_d. \tag{2.1}$$

In the next proposition, we show that the sets A_ℓ, $\ell \in \mathbb{Z}_{d+1}$, defined by (2.1) indeed form a partition for the space $\mathbb{R}^d$.

Proposition 2.1. *If the sets A_ℓ, $\ell \in \mathbb{Z}_{d+1}$, are defined by* (2.1), *then*

(i) *they are mutually disjoint,*
(ii) *they form a partition for the space $\mathbb{R}^d$, that is,*

$$\mathbb{R}^d = \bigcup_{\ell \in \mathbb{Z}_{d+1}} A_\ell. \tag{2.2}$$

Proof. (i) It suffices to show that

$$A_j \cap A_{j'} = \emptyset, \quad \text{for all } j, j' \in \mathbb{Z}_{d+1} \quad \text{with } j \ne j'. \tag{2.3}$$

Without loss of generality, we assume that $j < j'$. Suppose that $x \in A_j \cap A_{j'}$. By the definition of A_j, there exist $1 \le k_1 < k_2 < \cdots < k_j \le d$ and $x_{k_i} \in \mathbb{R}\backslash\{0\}$ such that

$$x = \sum_{i \in \mathbb{N}_j} x_{k_i} e_{k_i}, \tag{2.4}$$

and by the definition of $A_{j'}$, there exist $1 \le k'_1 < k'_2 < \cdots < k'_{j'} \le d$ and $x'_{k'_i} \in \mathbb{R}\backslash\{0\}$ such that

$$x = \sum_{i \in \mathbb{N}_{j'}} x'_{k'_i} e_{k'_i}. \tag{2.5}$$

Subtracting Eq. (2.4) from (2.5) yields

$$\sum_{i \in \mathbb{N}_{j'}} x'_{k'_i} e_{k'_i} - \sum_{i \in \mathbb{N}_j} x_{k_i} e_{k_i} = 0. \tag{2.6}$$

We introduce two index sets $\mathbb{I} := \{k_1, k_2, \ldots, k_j\}$ and $\mathbb{I}' := \{k'_1, k'_2, \ldots, k'_{j'}\}$. Since $j < j'$, we observe that $\mathbb{I} \ne \mathbb{I}'$. It follows that there exists an index $k'_t \in \mathbb{I}'$ but $k'_t \notin \mathbb{I}$. Since e_j, $j \in \mathbb{N}_d$, are linearly independent, according to (2.6), we conclude that $x'_{k'_t} = 0$. This contradicts the hypothesis that $x'_{k'_t} \ne 0$ and confirms (2.3).

(ii) Assume that $x \in \mathbb{R}^d$. Let $\ell := \|x\|_0$. Then, $\ell \in \mathbb{Z}_{d+1}$. Thus, we have that $x \in A_\ell$. This ensures that

$$\mathbb{R}^d \subseteq \bigcup_{\ell \in \mathbb{Z}_{d+1}} A_\ell.$$

Clearly, we have that

$$\bigcup_{\ell\in\mathbb{Z}_{d+1}} A_\ell \subseteq \mathbb{R}^d.$$

These two inclusions imply the validity of Eq. (2.2), which together with part (i) of this proposition confirms that the sets A_ℓ, $\ell \in \mathbb{Z}_{d+1}$, form a partition for $\mathbb{R}^d$. □

We illustrate Proposition 2.1 by $\mathbb{R}^2$. Clearly, for $d = 2$, $\mathbb{R}^2 = A_0 \cup A_1 \cup A_2$, where

$$A_0 := \{(0,0)\}, \quad A_1 := \{(x,0) : x \in \mathbb{R}\backslash\{0\}\} \cup \{(0,y) : y \in \mathbb{R}\backslash\{0\}\},$$

and

$$A_2 := \{(x,y) : (x,y) \in \mathbb{R}\times\mathbb{R}, x \neq 0 \text{ and } y \neq 0\}.$$

That is, A_1 contains points on the two axes except the origin and A_2 contains the four quadrants of the two-dimensional plane.

We remark that according to (2.1), for each $\ell \in \mathbb{Z}_{d+1}$, A_ℓ is the set of all vectors in $\mathbb{R}^d$ having sparsity of level ℓ. According to Proposition 2.1, the space $\mathbb{R}^d$ has the sparsity partition A_j, $j \in \mathbb{Z}_{d+1}$, which groups the vectors in $\mathbb{R}^d$ according to their sparsity levels. We further observe that the sets A_ℓ, $\ell \in \mathbb{N}_d$, are closed under the operation of nonzero scalar multiplication, but not closed under the operation of addition. For example, $e_1, e_2 \in A_1$ but $e_1 + e_2 \in A_2$.

It is also convenient to define the set of vectors in $\mathbb{R}^d$ whose sparsity levels do not exceed ℓ. For $\ell \in \mathbb{Z}_{d+1}$, we let

$$\Omega_\ell := \bigcup_{j\in\mathbb{Z}_{\ell+1}} A_j.$$

Clearly, Ω_ℓ is the set of vectors in $\mathbb{R}^d$ whose sparsity levels do not exceed ℓ. Moreover, we have that

$$\Omega_0 = A_0, \quad \Omega_{j+1} = \Omega_j \cup A_{j+1}, \quad j \in \mathbb{Z}_{d+1} \quad \text{and} \quad \Omega_d = \mathbb{R}^d. \tag{2.7}$$

These equations yield that

$$A_d = \mathbb{R}^d\backslash\Omega_{d-1}.$$

The set A_d consists of the vectors in $\mathbb{R}^d$ whose components are all nonzero. By the definition of the sets Ω_j and properties of A_j, we see that Ω_j for $j \in \mathbb{N}_{d-1}$ are closed under the operation of nonzero scalar multiplication, but not closed under the operation of addition.

We now consider a partition of the space $\mathbb{R}^m$, the preimage space of the transform M, induced by the sparsity partition of $\mathbb{R}^d$. Suppose that

$$M\mathbb{R}^m = \mathbb{R}^d. \tag{2.8}$$

When condition (2.8) is satisfied, we say that M is of full rank. We introduce $d+1$ subsets B_j, $j \in \mathbb{Z}_{d+1}$, of the preimage space $\mathbb{R}^m$ according to the sparsity partition

A_j, $j \in \mathbb{Z}_{d+1}$ by

$$B_j := \{x \in \mathbb{R}^m : Mx \in A_j\}.$$

Because $A_0 = \{0\}$, the set B_0 is the null space of matrix M. Moreover, we have the following simple fact.

Proposition 2.2. *If M is a $d \times m$ full rank matrix, then*

$$MB_j = A_j, \quad \text{for all } j \in \mathbb{Z}_{d+1}.$$

Proof. Let $j \in \mathbb{Z}_{d+1}$ be fixed. We assume that $y \in MB_j$. Thus, there exists $x \in B_j$ such that $y = Mx$. By the definition of B_j, we have that $Mx \in A_j$. Hence, $y \in A_j$. This implies the inclusion $MB_j \subseteq A_j$.

Conversely, we let $y \in A_j$. By Proposition 2.1, the sets A_j, $j \in \mathbb{R}^d$, form a partition for the space $\mathbb{R}^d$ and thus, $y \in \mathbb{R}^d$. Since M is of full rank, according to Eq. (2.8), there exists $x \in \mathbb{R}^m$ such that $y = Mx$. Since $Mx \in A_j$, we find that $x \in B_j$. Thus, we have that $y \in MB_j$. This yields the inclusion $A_j \subseteq MB_j$. We therefore establish the desired equation of this proposition. □

Proposition 2.2 clearly reveals that for each $j \in \mathbb{Z}_{d+1}$, the set B_j is the preimage set of A_j, the set of the vectors in $\mathbb{R}^d$ having sparsity of level j, under the transform M. However, vectors in B_j do not necessarily have sparsity of level j. In the next proposition, we show that the sets B_j, $j \in \mathbb{Z}_{d+1}$, form a partition for the preimage space $\mathbb{R}^m$ of the transform M.

Proposition 2.3. *If M is a $d \times m$ full rank matrix, then the sets B_j, $j \in \mathbb{Z}_{d+1}$, form a partition for the space $\mathbb{R}^m$.*

Proof. It suffices to establish that

$$\mathbb{R}^m = \bigcup_{j \in \mathbb{Z}_{d+1}} B_j \tag{2.9}$$

and

$$B_j \cap B_{j'} = \emptyset \quad \text{for all } j, j' \in \mathbb{Z}_{d+1} \text{ with } j \neq j'. \tag{2.10}$$

To show (2.9), we let $x \in \mathbb{R}^m$. By the hypothesis on matrix M, we see that Eq. (2.8) holds and thus, $Mx \in \mathbb{R}^d$. Employing the sparsity partition A_j, $j \in \mathbb{Z}_{d+1}$, for the space $\mathbb{R}^d$, we see that there exists $j \in \mathbb{Z}_{d+1}$ such that $Mx \in A_j$. By the definition of the set B_j, we conclude that $x \in B_j$. Hence, we have that

$$\mathbb{R}^m \subseteq \bigcup_{j \in \mathbb{Z}_{d+1}} B_j.$$

By the definition of the sets B_j, each of these sets is contained in $\mathbb{R}^m$. Thus,

$$\bigcup_{j \in \mathbb{Z}_{d+1}} B_j \subseteq \mathbb{R}^m.$$

Consequently, Eq. (2.9) holds true.

It remains to prove Eq. (2.10). Suppose that $x \in B_j \cap B_{j'}$ for a fixed pair of indices $j, j' \in \mathbb{Z}_{d+1}$, with $j \neq j'$. By the definition of the set B_j, we have that $Mx \in A_j$ and by the definition of the set $B_{j'}$, we have that $Mx \in A_{j'}$. According to Proposition 2.1, the two sets A_j and $A_{j'}$ are disjoint. This clearly implies that $Mx \in \emptyset$, a contradiction. This yields Eq. (2.10). □

In the remaining part of this section, we study useful topological properties of the sets that we introduced earlier in this section.

Proposition 2.4. *The following statements holds true*:

(i) *The set A_0 is closed, the sets A_ℓ, $\ell \in \mathbb{N}_{d-1}$, are neither closed nor open, and A_d is open.*
(ii) *For $j \in \mathbb{Z}_d$, Ω_j are closed sets.*

Proof. (i) Since A_0 contains only one point 0, it is closed. It is straightforward to see that the sets A_j, $j \in \mathbb{N}_{d-1}$, are not open since in every neighborhood of a vector in A_j contains vectors that are not in A_j. We now show that the sets A_j, $j \in \mathbb{N}_{d-1}$, are not closed either. To this end, we consider a sequence of vectors x_n, $n \in \mathbb{Z}$, in $\mathbb{R}^d$ whose first $j-1$ components are all equal to 1, last $d-j$ components are all equal to zero and jth component is $1/n$, that is,

$$x_n := (1, \ldots, 1, 1/n, 0, \ldots, 0), \quad \text{for all } n \in \mathbb{Z}.$$

Clearly, $x_n \in A_j$, for all $n \in \mathbb{Z}$ and $\|x_n - \hat{x}\|_2 = 1/n \to 0$ as $n \to \infty$, where $\hat{x} \in \mathbb{R}^d$ whose first $j-1$ components are all equal to 1 and last $d-j+1$ components are all equal to zero, that is, $\hat{x} := (1, \ldots, 1, 0, 0, \ldots, 0)$. In other words, x_n converges to a vector in A_{j-1} not in A_j. Therefore, A_j, $j \in \mathbb{N}_{d-1}$, are not closed.

It remains to show that A_d is open. Suppose that $\hat{x} \in A_d$. Then, we have that $\hat{x} = (\hat{t}_1, \hat{t}_2, \ldots, \hat{t}_d)$ with $\hat{t}_j \neq 0$, for all $j \in \mathbb{N}_d$. Hence, for all $j \in \mathbb{N}_d$ there exists $\epsilon > 0$ such that for all $t'_j \in (\hat{t}_j - \epsilon/d^{1/2}, \hat{t}_j + \epsilon/d^{1/2})$, we have that $t'_j \neq 0$. Let $x' := (t'_1, t'_2, \ldots, t'_d)$. We observe that $x' \in A_d$. That is, the open ball

$$\mathcal{B}_o(\hat{x}, \epsilon) := \{x \in \mathbb{R}^d : \|x - \hat{x}\|_2 < \epsilon\}$$

is contained in A_d. Thus, A_d is an open set.

(ii) For a fixed $j \in \mathbb{Z}_d$, we assume that a sequence x_n, $n \in \mathbb{Z}$, in Ω_j converges to a point $x \in \mathbb{R}^d$ as $n \to \infty$, and we show that $x \in \Omega_j$ by contradiction. Assume, to the contrary, that $x \notin \Omega_j$. By the second equation of (2.7), we have that $\Omega_j \subset \Omega_{j+1}$. Without loss of generality, we assume that $x \in \Omega_{j+1}$. Hence, $x \in A_{j+1}$. That is, x has exactly $j+1$ nonzero components. Therefore, for sufficiently large n, x_n has at least $j+1$ nonzero components. This contradicts the assumption that $x_n \in \Omega_j$, which implies that x has at most j nonzero components. This contradiction proves that Ω_j is closed. □

The next result translates the openness of A_d to its preimage set B_d.

Proposition 2.5. *If M is a $d \times m$ full rank matrix, then the set B_d is an open set in $\mathbb{R}^m$.*

Proof. By Proposition 2.2, we have that $MB_d = A_d$. According to Proposition 2.4, A_d is open. Note that M can be viewed as a continuous mapping from $\mathbb{R}^m$ to $\mathbb{R}^d$. Hence, we see that B_d is open since the preimage of an open set under a continuous mapping is open. □

It is clear that the ℓ_0 norm is not a continuous function in the sense that the condition $\|x_n - \hat{x}\|_2 \to 0$ does not guarantee that $\|x_n\|_0 \to \|\hat{x}\|_0$. This is seen from the example $x_n := (1/n, 1/n, \ldots, 1/n)$ and $\hat{x} = 0$.

It is important to understand how the sparsity of a vector in $\mathbb{R}^d$ influences the sparsity of vectors in its neighborhood. To this end, for a given index set $\mathcal{I} \subseteq \mathbb{N}_d$ we define a subspace of $\mathbb{R}^d$ by letting

$$\mathcal{C}_{\mathcal{I}} := \{x \in \mathbb{R}^d : S(x) \subseteq \mathcal{I}\}, \tag{2.11}$$

where $S(x)$ denotes the support of $x \in \mathbb{R}^d$, that is, $S(x) := \{i \in \mathbb{N}_d : x_i \neq 0\}$, for $x \in \mathbb{R}^d$. Clearly, $\mathcal{C}_{\mathcal{I}}$ is convex. It is convenient to define the set

$$\partial\mathcal{C}_{\mathcal{I}} := \{x \in \mathbb{R}^d : S(x) = \mathcal{I}\}. \tag{2.12}$$

We first establish a technical lemma.

Lemma 2.6. *If for some $\ell \in \mathbb{Z}_{d+1}$, $\hat{x} \in A_\ell$, then*

$$\operatorname{dist}(\hat{x}, A_\ell \backslash \partial\mathcal{C}_{S(\hat{x})}) > 0, \tag{2.13}$$

where

$$\operatorname{dist}(x, A) := \min\{\|x - z\|_2 : z \in A\}.$$

Proof. Since $\hat{x} \in A_\ell$, we may assume that $S(\hat{x}) = \{k_1, k_2, \ldots, k_\ell\}$, where $1 \leq k_1 < k_2 < \cdots < k_\ell \leq d$. It follows that $\hat{x}$ may be represented as

$$\hat{x} = \sum_{j \in \mathbb{N}_\ell} \hat{x}_{k_j} e_{k_j}, \quad \text{for some } \hat{x}_{k_j} \in \mathbb{R}\backslash\{0\}, \quad j \in \mathbb{N}_\ell.$$

For all $z \in A_\ell \backslash \partial\mathcal{C}_{S(\hat{x})}$, we have that $S(\hat{x}) \neq S(z)$, and there exist integers k'_j, $j \in \mathbb{N}_\ell$, with $1 \leq k'_1 < k'_2 < \cdots < k'_\ell \leq d$ such that

$$z = \sum_{j \in \mathbb{N}_\ell} z_{k'_j} e_{k'_j}, \quad \text{for some } z_{k'_j} \in \mathbb{R}\backslash\{0\}, \quad j \in \mathbb{N}_\ell.$$

Hence, there exists some $k_j \in S(\hat{x})$ but $k_j \notin S(z)$. This fact together with the above representations of $\hat{x}$ and z implies that for all $z \in A_\ell \backslash \partial\mathcal{C}_{S(\hat{x})}$, $\|x - z\|_2 \geq |\hat{x}_{k_j}| > 0$.

From this we conclude that

$$\operatorname{dist}(\hat{x}, A_\ell \backslash \partial \mathcal{C}_{S(\hat{x})}) \geq |\hat{x}_{k_j}| > 0,$$

which completes the proof of this lemma. □

With the help of Lemma 2.6, we prove the following proposition. We define the closed ball with the center $\hat{x} \in \mathbb{R}^d$ and radius $\epsilon > 0$ by

$$\mathcal{B}(\hat{x}, \epsilon) := \{x \in \mathbb{R}^d : \|x - \hat{x}\|_2 \leq \epsilon\}.$$

Proposition 2.7. *The following statements hold true*:

(i) *If for some* $\ell \in \mathbb{Z}_{d+1}$, $\hat{x} \in A_\ell$, *then there exists* $\delta > 0$ *such that for all* $x \in \mathcal{B}(\hat{x}, \delta)$, *there holds* $x \in \cup_{j=\ell}^{d} A_j$, *that is*, $\|x\|_0 \geq \|\hat{x}\|_0$.

(ii) *If for some* $\ell \in \mathbb{Z}_{d+1}$, $\hat{x} \in A_\ell$, *then there exists* $\delta > 0$ *such that for all* $x \in \mathcal{B}(\hat{x}, \delta) \backslash \mathcal{C}_{\mathcal{I}}$ *with* $\mathcal{I} := S(\hat{x})$, *there holds* $x \in A_j$, *for some* $j \geq \ell + 1$, *that is*, $\|x\|_0 \geq \|\hat{x}\|_0 + 1$.

Proof. (i) We prove this assertion by contradiction. Assume to the contrary that the statement is not true. Then, for any $\delta > 0$, there exists $x_\delta \in \mathcal{B}(\hat{x}, \delta)$ such that $x_\delta \in \Omega_{\ell-1}$. By Item (ii) of Proposition 2.4, the set $\Omega_{\ell-1}$ is closed. This implies that $\hat{x} \in \Omega_{\ell-1}$, which contradicts the assumption that $\hat{x} \in A_\ell$. Therefore, there exists $\delta > 0$ such that for all $x \in \mathcal{B}(\hat{x}, \delta)$, $x \in A_j$, for some $j \geq \ell$. This further implies that $\|x\|_0 \geq \|\hat{x}\|_0$.

(ii) By Lemma 2.6, we may choose $\delta_0 > 0$ such that

$$\operatorname{dist}(\hat{x}, A_\ell \backslash \partial \mathcal{C}_{S(\hat{x})}) > \delta_0.$$

By Item (i) of this proposition, there exists a δ with $\delta_0 > \delta > 0$ such that for all $x \in \mathcal{B}(\hat{x}, \delta)$, we have that $x \in A_j$, for some $j \geq \ell$. Hence, for this positive number δ, there holds

$$(\mathcal{B}(\hat{x}, \delta) \backslash \partial \mathcal{C}_{S(\hat{x})}) \cap A_\ell = \emptyset.$$

It follows that for all $x \in \mathcal{B}(\hat{x}, \delta) \backslash \mathcal{C}_{S(\hat{x})}$, there holds $x \in A_j$, for some $j \geq \ell + 1$. Consequently, for all $x \in \mathcal{B}(\hat{x}, \delta) \backslash \mathcal{C}_{S(\hat{x})}$, we have that $\|x\|_0 \geq \|\hat{x}\|_0 + 1$. □

3. Sparsity Regularization Under a Transform

In this section, we consider the minimization problem of sparsity regularization under a transform. The rationale for considering the regularization problem (1.2) with the function f having the form (1.1) is that g has a global minimizer but it may not be sparse under the transform. We then impose the regularization term. By choosing the parameter λ appropriately, we seek a global minimizer of f having sparsity of a prescribed level and close to the global minimizer of g. Specifically, we intend to understand how choices of the regularization parameter λ lead to sparsity

(under the transform M) of a global minimizer of the function f defined by (1.1) when M is a real $d \times m$ matrix of full rank.

We first comment on a connection between the sets A_ℓ defined by Eq. (2.1) and the ℓ_0 norm. By the definition of the ℓ_0 norm, for any $x \in \mathbb{R}^d$ we have that

$$\|x\|_0 = \ell, \quad \text{if } x \in A_\ell, \quad \text{for some } \ell \in \mathbb{Z}_{d+1}. \tag{3.1}$$

Formula (3.1) can simplify the function f defined by (1.1) on each set B_ℓ and provides a key to understand the solution of the related regularization problem (1.2). In fact, by employing formula (3.1) and the partition B_j, $j \in \mathbb{Z}_{d+1}$, of $\mathbb{R}^m$, connected with the sets A_ℓ, $\ell \in \mathbb{Z}_{d+1}$ via Proposition 2.2, we have an alternative representation of function f defined by (1.1). Namely,

$$f(x) = g(x) + \lambda\ell, \quad \text{for all } x \in B_\ell,\ \ell \in \mathbb{Z}_{d+1}. \tag{3.2}$$

Geometric interpretation of the function f defined by (1.1) provides insights to sparsity of a global minimizer of f under the transform M. By adding the regularization term $\lambda\|M \cdot\|_0$ to g results in lifting the graph of g according to the sparsity in the range of M. In other words, the regularization term $\lambda\|M \cdot\|_0$ terraces the graph of function g. Specifically, the values of function g that stay unchanged are $g(x)$ for all $x \in B_0$ (that is, in the null space of M) and every other value $g(x)$ is lifted according to which set B_j the points x belong to. For example, for all $x \in B_1$, the values $g(x)$ are lifted to $g(x) + \lambda$. In general, for all $x \in B_j$, the values $g(x)$ are lifted to $g(x) + \lambda j$, for $j \in \mathbb{Z}_{d+1}$. On the highest level of the terraces are $g(x) + \lambda d$, for all $x \in B_d$, where all components of Mx are nonzero. Hence, by changing the value of the parameter λ, the landscape of the graph of the associated function f is changed and accordingly the sparsity of the global minimizer of f is changed. For instance, if the most sparse global minimizer is desired (that is, a point in the null space of M as a global minimizer of f), then a value of λ is chosen so that the function values $g(x) + \lambda j$, for all $x \in B_j$, $j \in \mathbb{N}_d$, are greater than the value $g(\hat{x})$, where $\hat{x}$ is in the null space of the matrix M. This understanding is a key to guide for choices of the parameter λ.

When a global minimizer of f that are most sparse is desired, we have the parameter choice strategy described in the next theorem.

Theorem 3.1. *Let $x^* \in \mathbb{R}^m$ be a global minimizer of g and $x_0 \in B_0$ be a minimizer of g on B_0. If the parameter λ is chosen to satisfy*

$$\lambda \geq g(x_0) - g(x^*), \tag{3.3}$$

then x_0 is a global minimizer of f in the space $\mathbb{R}^m$, Mx_0 has sparsity of level 0 and the global minimum value of f is given by $g(x_0)$.

Proof. Let $x \in \mathbb{R}^m$ be an arbitrarily fixed vector. We make use of the partition B_j, $j \in \mathbb{Z}_{d+1}$, of $\mathbb{R}^m$ to conclude that there exists $j' \in \mathbb{Z}_{d+1}$ such that $x \in B_{j'}$. We consider two cases: $j' = 0$ and $j' \in \mathbb{N}_d$.

In the case when $j' = 0$, we have that $x \in B_0$ and $Mx = 0$. In this case, by employing Eq. (3.2), from the assumption that $x_0 \in B_0$ is a minimizer of g on B_0 we get that

$$f(x) = g(x) \geq g(x_0) = f(x_0).$$

Next, we consider the case when $j' \in \mathbb{N}_d$. In this case, we have that $x \in B_{j'}$, that is, $Mx \in A_{j'}$. Once again, we employ Eq. (3.2) to obtain that

$$f(x) = g(x) + \lambda j' \geq g(x) + \lambda.$$

Combining this inequality with the assumption on x^* and condition (3.3), we obtain that

$$f(x) \geq g(x) + \lambda \geq g(x^*) + \lambda \geq g(x_0) = f(x_0).$$

In both of the cases, we have shown that $f(x) \geq f(x_0)$ for all $x \in \mathbb{R}^m$. Therefore, x_0 is a global minimizer of f on the space $\mathbb{R}^m$. □

Theorem 3.1 also provides the error bound between the regularized global minimum value $f(x_0)$ and the original global minimum value $g(x^*)$ when the parameter λ is chosen according to (3.3). Indeed, since $f(x_0) = g(x_0)$ and x^* is a global minimizer of g, we observe that $f(x_0) - g(x^*) = g(x_0) - g(x^*) \geq 0$. Hence, by (3.3), we obtain that

$$0 \leq f(x_0) - g(x^*) \leq \lambda.$$

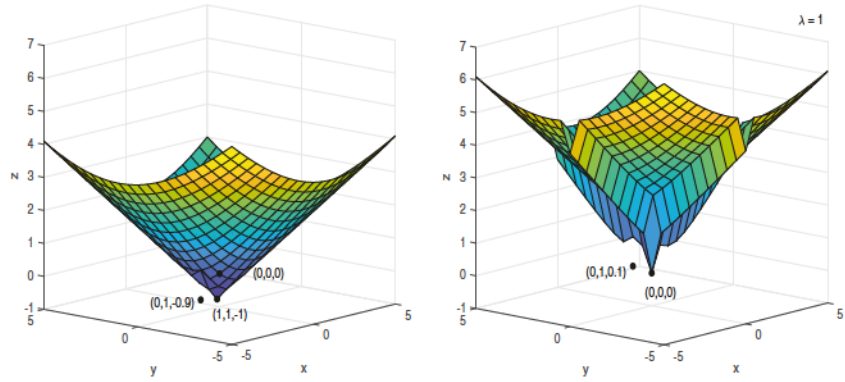
In general, the error is not equal to zero unless the global minimizer x^* of g is in B_0, in which case sparse regularization is not necessary.

We illustrate Theorem 3.1 by a simple example in $\mathbb{R}^2$. To this end, we consider a non-convex function defined by

$$g(x) := \begin{cases} \dfrac{\sqrt{2}}{2}\|x - (1,1)\|_2 - 1, & x \neq (0,1), \\ -0.9, & x = (0,1). \end{cases} \tag{3.4}$$

Clearly, as shown in Fig. 1(a) g has a unique, *non-sparse* global minimizer $x^* = (1,1)$ and the minimum value $g(x^*) = -1$. In this example, we choose $M = I$. The value $g(x_0) = 0$, where $x_0 := (0,0)$. According to Theorem 3.1, we choose $\lambda = g(x_0) - g(x^*) = 1$. The regularized non-convex function f defined by (1.3) with g in the form (3.4) and $\lambda = 1$ is shown in Fig. 1(b). Note that $x_0 := (0,0)$ is the *sparse* global minimizer of f and the minimum value $f(x_0) = 0$. Also, the error between the regularized global minimizer and the original global minimizer is given by $f(x_0) - g(x^*) = 1 = \lambda$. These figures illustrate how the regularization term $\lambda\|\cdot\|_0$ terraces the graph of g.

In the case that we wish to reduce the error of the regularized global minimum value, we may choose not to demand the most sparsity (under the transform M).



(a) (b)

Fig. 1. (a) Graph of function g; (b) Graph of the regularized function f with $\lambda = 1$.

For this reason, we consider a choice of the parameter λ with which the function f defined by Eq. (1.1) has a global minimizer having sparsity (under the transform M) of a prescribed level. We present a parameter choice strategy in the next theorem. For this purpose, we find it convenient to define another sequence of sets. For all $\ell \in \mathbb{Z}_{d+1}$, we let

$$\Gamma_\ell := \bigcup_{j \in \mathbb{Z}_{\ell+1}} B_j. \tag{3.5}$$

It can be readily verified that for each $\ell \in \mathbb{Z}_{d+1}$, the set Γ_ℓ is the preimage set of Ω_ℓ under the transform M. Moreover, by (2.9), we have that

$$\Gamma_d = \mathbb{R}^m. \tag{3.6}$$

Theorem 3.2. *Let $x^* \in \mathbb{R}^m$ be a global minimizer of g, $x' \in \Gamma_\ell$ be a minimizer of g on Γ_ℓ for some $\ell \in \mathbb{N}_d$, and $x_j \in \Gamma_j$ be a minimizer of g on Γ_j, for all $j \in \mathbb{Z}_\ell$. Suppose that*

$$g(x') - g(x^*) \le \frac{1}{\ell - j}[g(x_j) - g(x')], \quad \text{for all } j \in \mathbb{Z}_\ell. \tag{3.7}$$

If the parameter λ is chosen to satisfy the conditions

$$g(x') - g(x^*) \le \lambda \le \frac{1}{\ell - j}[g(x_j) - g(x')], \quad \text{for all } j \in \mathbb{Z}_\ell, \tag{3.8}$$

then x' is a global minimizer of f on $\mathbb{R}^m$, Mx' has sparsity of level ℓ' with $\ell' \le \ell$ and the global minimum value of f on $\mathbb{R}^m$ is given by $f(x') = g(x') + \lambda\ell'$.

Proof. We shall verify that

$$f(x') \le f(x), \quad \text{for all } x \in \mathbb{R}^m. \tag{3.9}$$

Since $x' \in \Gamma_\ell$ is a minimizer of g on Γ_ℓ, we use the definition (3.5) of Γ_ℓ to consider cases when $x' \in B_j$ for all $j \in \mathbb{Z}_{\ell+1}$.

We consider the first case when $x' \in B_0$, that is, $Mx' \in A_0$. In this case, we shall show (3.9) with $Mx' = 0$ by verifying it for all $x \in B_j$, for all $j \in \mathbb{Z}_{d+1}$. To this end, we consider two subcases according to the index $j \in \mathbb{Z}_{d+1}$: (1) $j \le \ell$ and (2) $j > \ell$. In subcase (1), we consider $x \in B_j$, $j \in \mathbb{Z}_{\ell+1}$. By the definition of the sets B_j, we have that $Mx \in A_j$, which implies

$$\|Mx\|_0 = j, \quad \text{if } x \in B_j, \quad \text{for all } j \in \mathbb{Z}_{\ell+1}.$$

Since x' is a minimizer of g on Γ_ℓ and $\|Mx'\|_0 = 0$, we have for all $x \in B_j$, $j \in \mathbb{Z}_{\ell+1}$, that

$$f(x') = g(x') \le g(x) \le g(x) + \lambda j = f(x).$$

In subcase (2), we consider $x \in B_j$, $j = \ell+1, \ell+2, \ldots, d$, where we also have that $Mx \in A_j$ and thus, $\|Mx\|_0 = j$. By employing the first inequality of (3.8) and the fact that x^* is a global minimizer of g, for all $x \in B_j$, $j = \ell+1, \ell+2, \ldots, d$, we obtain that

$$f(x') = g(x') \le g(x^*) + \lambda \le g(x) + \lambda \le g(x) + \lambda j = f(x).$$

This proves the first case of (3.9).

We next consider the second case when $x' \in B_k$, for $k \in \mathbb{N}_\ell$ and show (3.9) with $\|Mx'\|_0 = k$ by verifying (3.9) for all $x \in B_j$, for $j \in \mathbb{Z}_{d+1}$. To this end, we consider three subcases according to the index $j \in \mathbb{Z}_{d+1}$: (1) $j < \ell$, (2) $j = \ell$, (3) $j > \ell$. In subcase (1) for which $j < \ell$, we consider $x \in B_j$, $j \in \mathbb{Z}_\ell$. Clearly, we have that $\|Mx\|_0 = j$. Thus, by using the second inequality of (3.8) and the hypothesis that $x_j \in \Gamma_j$ is a minimizer of g on Γ_j, we observe that

$$g(x') + \lambda\ell \le g(x_j) + \lambda j \le g(x) + \lambda j = f(x).$$

This together with the fact $k \le \ell$ implies that

$$f(x') = g(x') + \lambda k \le g(x') + \lambda\ell \le f(x), \quad \text{for all } x \in B_j, \quad \text{for all } j \in \mathbb{Z}_\ell.$$

In subcase (2) for which $j = \ell$, we consider $x \in B_\ell$. Since x' is a minimizer of g on Γ_ℓ, which contains B_ℓ as a subset, and $k \le \ell$, we find that

$$f(x') = g(x') + \lambda k \le g(x) + \lambda k \le g(x) + \lambda\ell = f(x), \quad \text{for all } x \in B_\ell.$$

In subcase (3) for which $j > \ell$, we consider $x \in B_j$, for $j = \ell+1, \ell+2, \ldots, d$. By using the first inequality of (3.8) and again the fact that x^* is a global minimizer of g, we derive for all $x \in B_j$, for all $j = \ell+1, \ell+2, \ldots, d$ that

$$f(x') = g(x') + \lambda k \le g(x^*) + \lambda(k+1) \le g(x) + \lambda j = f(x).$$

That is, (3.9) holds true for the second case.

Summarizing the above verification, we conclude that (3.9) holds true for all cases and thus, x' is a global minimizer of the function f on the space $\mathbb{R}^m$. □

A comment on the parameter choice (3.8) in Theorem 3.2 is in order. To be able to choose such a parameter, it requires that condition (3.7) is satisfied. This hypothesis is indeed needed to ensure that the choice (3.8) of the parameter λ is feasible. The hypothesis (3.7) is equivalent to the following conditions that

$$g(x') \leq \frac{g(x_j) + (\ell - j)g(x^*)}{\ell - j + 1}, \quad \text{for all } j \in \mathbb{Z}_\ell. \tag{3.10}$$

The right-hand side of the inequality (3.10) is a weighted average of $g(x_j)$ and $g(x^*)$. By the definition of x', x_j and x^*, it is clear that

$$g(x^*) \leq g(x') \leq g(x_j), \quad \text{for all } j \in \mathbb{Z}_\ell.$$

This shows that the hypothesis (3.7) of Theorem 3.2 is reasonable. Condition (3.8) also reveals that the error of the regularized global minimum value $f(x')$ approximating the original global minimum value $g(x^*)$ is bounded by the value of the regularization parameter so chosen. That is,

$$0 \leq f(x') - g(x^*) \leq \lambda.$$

We next illustrate the result of Theorem 3.2 by presenting a corollary of Theorem 3.2 for the special case when $\ell = 1$. The corollary gives a choice of the parameter λ which guarantees that a global minimizer of f has sparsity of level 1 under the transform M. That is, the corresponding function f has a global minimizer having at most one nonzero component under the transform M.

Corollary 3.3. *Let $x^* \in \mathbb{R}^d$ be a global minimizer of g on the space $\mathbb{R}^m$, $x' \in \Gamma_1$ be a minimizer of g on Γ_1 and $x_0 \in B_0$ be a minimizer of g on B_0. If the parameter λ is chosen to satisfy the condition*

$$g(x') - g(x^*) \leq \lambda \leq g(x_0) - g(x'), \tag{3.11}$$

then x' is a global minimizer of f on the space $\mathbb{R}^m$.

Proof. This result is obtained by specializing Theorem 3.2 to the special case when $\ell = 1$. □

Clearly, as we have discussed earlier, for the choice (3.11) of the parameter λ to be feasible, we need to require that

$$g(x') \leq \frac{1}{2}[g(x^*) + g(x_0)]. \tag{3.12}$$

That is, $g(x')$ is less than or equal to the average of $g(x^*)$ and $g(x_0)$. The function g defined by (3.4) on $\mathbb{R}^2$ clearly satisfies condition (3.12).

The next theorem connects a global minimizer of g with a global (or local) minimizer of f. To this end, we recall the definition of a local minimizer of a non-convex function. A vector $x^* \in \mathbb{R}^d$ is called a local minimizer of f, if there exists a $\delta > 0$ such that

$$f(x^*) \le f(x), \quad \text{for all } x \in \mathcal{B}(x^*, \delta).$$

Theorem 3.4. *Let $x^* \in \mathbb{R}^m$ be a global minimizer of g.*

(i) *If $Mx^* = 0$, then x^* is a global minimizer of f on $\mathbb{R}^m$.*
(ii) *If for some $\ell \in \mathbb{N}_d$, $x^* \in B_\ell$ and if for a minimizer $x_j \in B_j$ of g on B_j for all $j \in \mathbb{Z}_\ell$, the parameter λ is chosen to satisfy*

$$0 \le \lambda \le \frac{1}{\ell - j}[g(x_j) - g(x^*)], \quad \text{for all } j \in \mathbb{Z}_\ell, \tag{3.13}$$

then x^* is a global minimizer of f on $\mathbb{R}^m$.
(iii) *If $x^* \in B_d$, then x^* is a local minimizer of f.*
(iv) *If $x^* \in B_d$, and for some $j \in \mathbb{Z}_{d+1}$ and for some $\tilde{x} \in B_j$,*

$$g(x^*) + \lambda(d - j) > g(\tilde{x}), \tag{3.14}$$

then x^* is a local minimizer of f but not a global minimizer of f on $\mathbb{R}^m$, and global minimizers (if exist) of f on $\mathbb{R}^m$ have sparsity of level at least $d - 1$.

Proof. Since $x^* \in \mathbb{R}^d$ is a global minimizer of g, we have that

$$g(x^*) \le g(x), \quad \text{for all } x \in \mathbb{R}^m. \tag{3.15}$$

For both Items (i) and (ii), we shall show that

$$f(x^*) \le f(x), \quad \text{for all } x \in \mathbb{R}^m. \tag{3.16}$$

(i) If $Mx^* = 0$, by the definition of $\|\cdot\|_0$, we have that $\|Mx^*\|_0 = 0$ and thus,

$$f(x^*) = g(x^*) + \|Mx^*\|_0 = g(x^*).$$

Consequently, according to condition (3.15), we obtain that

$$f(x^*) = g(x^*) \le g(x) \le g(x) + \lambda\ell = f(x), \quad \text{for } x \in B_\ell, \quad \text{for all } \ell \in \mathbb{Z}_{d+1}.$$

This confirms that (3.16) holds true.

(ii) Since for some $\ell \in \mathbb{N}_d$, $x^* \in B_\ell$ and satisfies condition (3.13), and since $x_j \in B_j$ is a minimizer of g on B_j for all $j \in \mathbb{Z}_\ell$, we have that

$$f(x^*) = g(x^*) + \lambda\ell \le g(x_j) + \lambda j \le g(x) + \lambda j = f(x), \quad \text{for all } x \in B_j,\ j \in \mathbb{Z}_\ell.$$

Moreover, we have that

$$f(x^*) = g(x^*) + \lambda\ell \le g(x) + \lambda j = f(x), \quad \text{for all } x \in B_j,\ j = \ell, \ell+1, \dots, d.$$

Hence, (3.16) is satisfied.

(iii) Since $x^* \in B_d$ is a global minimizer of g on $\mathbb{R}^m$, we have for all $x \in B_d$ that

$$f(x^*) = g(x^*) + \lambda d \le g(x) + \lambda d = f(x).$$

That is,

$$f(x^*) \le f(x), \quad \text{for all } x \in B_d. \tag{3.17}$$

According to Proposition 2.5, B_d is an open set. Thus, x^* is an interior point of B_d. This ensures that there exists a $\delta > 0$ such that $\mathcal{B}(x^*, \delta)$ is contained in B_d. Therefore, from inequality (3.17) we conclude that $f(x^*) \le f(x)$ for all $x \in \mathcal{B}(x^*, \delta)$, and thus, x^* is a local minimizer of f on $\mathbb{R}^m$.

(iv) By (iii), we have known that in this case, $x^* \in B_d$ is a local minimizer of f. We next show that x^* is not a global minimizer of f on $\mathbb{R}^m$. By (3.14), we have that for some $j \in \mathbb{Z}_{d+1}$ and for some $\tilde{x} \in B_j$,

$$g(x^*) + \lambda d > g(\tilde{x}) + \lambda j.$$

This together with the fact $x^* \in B_d$ ensures that

$$f(x^*) = g(x^*) + \lambda d > g(\tilde{x}) + \lambda j = f(\tilde{x}).$$

This implies that x^* is not a global minimizer of f.

Finally, we prove that there is a global minimizer of f on $\mathbb{R}^m$ having sparsity of level at least $d-1$. From (3.17), we know that x^* is a minimizer of f on B_d. Note that by (3.6), there holds $\mathbb{R}^m \backslash B_d = \Gamma_{d-1}$. Hence, the fact established earlier that x^* is not a global minimizer of f implies that global minimizers (if exist) of f must occur at a point in Γ_{d-1}. By the definition of Γ_{d-1} such a global minimizer has sparsity of level at least $d-1$. □

In the next theorem, we prove necessary conditions of a global minimizer of f.

Theorem 3.5. *Let $x^* \in \mathbb{R}^m$ be a global minimizer of f on $\mathbb{R}^m$.*

(i) *If $x^* \in B_\ell$ for some $\ell \in \mathbb{Z}_{d+1}$, then x^* is a minimizer of g on Γ_ℓ.*
(ii) *If $x^* \in \mathbb{R}^m$ is not a global minimizer of g on $\mathbb{R}^m$, then $x^* \in \Gamma_{d-1}$.*

Proof. (i) Since $x^* \in B_\ell$ for some $\ell \in \mathbb{Z}_{d+1}$ and it is a global minimizer of f on $\mathbb{R}^m$, we have that

$$g(x^*) + \lambda\ell = f(x^*) \le f(x) = g(x) + \lambda j, \quad \text{for all } x \in B_j,\ j \in \mathbb{Z}_{\ell+1}.$$

It follows that

$$g(x^*) + \lambda(\ell - j) \le g(x), \quad \text{for all } x \in B_j,\ j \in \mathbb{Z}_{\ell+1}.$$

Using this inequality and noting that $\lambda(\ell - j) \ge 0$, for all $j \in \mathbb{Z}_{\ell+1}$, we have that

$$g(x^*) \le g(x^*) + \lambda(\ell - j) \le g(x), \quad \text{for all } x \in B_j,\ j \in \mathbb{Z}_{\ell+1}.$$

The above inequality together with the definition (3.5) of the set Γ_ℓ ensures that $x^* \in B_\ell$ is a minimizer of g on Γ_ℓ.

(ii) We prove this assertion by contradiction. Assume to the contrary that $x^* \notin \Gamma_{d-1}$. Since $x^* \in \mathbb{R}^m$,

$$\mathbb{R}^m = \Gamma_{d-1} \cup B_d \quad \text{and} \quad \Gamma_{d-1} \cap B_d = \emptyset,$$

we must have that $x^* \in B_d$. By Statement (i) of this theorem with $\ell = d$, we conclude that x^* is a minimizer of g on Γ_d. Noting that $\Gamma_d = \mathbb{R}^m$, we confirm that x^* is a global minimizer of g on $\mathbb{R}^m$. This contradicts the hypothesis that x^* is not a global minimizer of g on $\mathbb{R}^m$. This contradiction ensures that $x^* \in \Gamma_{d-1}$. □

In Theorem 3.5(ii), we provide sufficient conditions which guarantee that a global minimizer of f is sparse under the transform M.

The next result follows immediately from Theorem 3.5(ii).

Corollary 3.6. *If $x^* \in \mathbb{R}^m$ is a global minimizer of f on $\mathbb{R}^m$, then, either $x^* \in \Gamma_{d-1}$ or x^* is a global minimizer of g on $\mathbb{R}^m$.*

4. Sparse Regularization in the Spacial Domain

This section is devoted to presentation of special results for regularization problem

$$\min\{f(x, y) : (x, y) \in \mathbb{R}^d \times \mathbb{R}^{d'}\}, \tag{4.1}$$

where f is defined by (1.4). In this model, we seek sparsity for the variable x only.

A typical example of optimization problem (4.1) is approximately sparse regularization. In such cases, the function f may take the following form

$$f(x, y) := \phi(y) + \mu\|x - Dy\|_2^2 + \lambda\|x\|_0, \quad (x, y) \in \mathbb{R}^d \times \mathbb{R}^{d'}, \tag{4.2}$$

or

$$f(x, y) := \phi(y) + \mu\|x - Dy\|_1 + \lambda\|x\|_0, \quad (x, y) \in \mathbb{R}^d \times \mathbb{R}^{d'}, \tag{4.3}$$

where ϕ is a convex function and D is $d \times d'$ matrix. Form (4.2) relates to regularization by the envelope of the ℓ_0 norm and form (4.3) relates to regularization by the capped ℓ_1 norm [11]. For specific examples of f, see [33] for inverting incomplete Fourier transform, [34, 35] for image/signal processing, [37] for medical image reconstruction and machine learning [16, 17, 21, 29].

Employing the partition A_j, $j \in \mathbb{Z}_{d+1}$, of $\mathbb{R}^d$ and the definition of $\|\cdot\|_0$, we have an alternative representation of function f:

$$f(x,y) = g(x,y) + \lambda\ell, \quad \text{for all } x \in A_\ell,\ \ell \in \mathbb{Z}_{d+1} \quad \text{and for all } y \in \mathbb{R}^{d'}. \tag{4.4}$$

Clearly, adding the function $\lambda\|x\|_0$ to $g(x,y)$ results in lifting the graph of $g(x,y)$ according to the sparsity partition of $\mathbb{R}^d$ with respect to the variable x. In other words, the ℓ_0 norm terraces the graph of function $g(x,y)$. Specifically, the only value that stays unchanged is $g(0,y)$ and every other value of $g(x,y)$ is lifted according to which set A_j the point x belongs to. For example, for $x \in A_1$, $g(x,y)$ is lifted to $g(x,y)+\lambda$. In general, for $x \in A_j$, $g(x,y)$ is lifted to $g(x,y)+\lambda j$, for $j \in \mathbb{Z}_{d+1}$. On the highest level of the terraces is $g(x,y)+\lambda d$, for all $x \in A_d$, where all components of x are nonzero. Hence, by changing the value of the parameter λ, the landscape of the graph of the corresponding function f is changed. Accordingly the sparsity of the global minimizer of f is changed.

We first consider a choice of the parameter λ with which the function f defined by (1.4) has the most sparse minimizer in variable x. To this end, we assume that the function $g(x,y)$ has a global minimizer $(x^*,y^*) \in \mathbb{R}^d \times \mathbb{R}^{d'}$.

Theorem 4.1. *Let $(x^*,y^*) \in \mathbb{R}^d \times \mathbb{R}^{d'}$ be a global minimizer of g and $y_0 \in \mathbb{R}^{d'}$ is a minimizer of $g(0,\cdot)$ on $\mathbb{R}^{d'}$. If the parameter λ is chosen to satisfy*

$$\lambda \geq g(0,y_0) - g(x^*,y^*), \tag{4.5}$$

then the pair $(0,y_0) \in \mathbb{R}^d \times \mathbb{R}^{d'}$ is a global minimizer of f on $\mathbb{R}^d \times \mathbb{R}^{d'}$ and $g(0,y_0)$ is the minimum value of f on $\mathbb{R}^d \times \mathbb{R}^{d'}$. Moreover, if the inequality (4.5) becomes strict, then the pair $(0,y_0) \in \mathbb{R}^d \times \mathbb{R}^{d'}$ is the unique global minimizer of f on $\mathbb{R}^d \times \mathbb{R}^{d'}$.

Proof. We consider an arbitrary $x \in \mathbb{R}^d\backslash\{0\}$ and use the sparsity partition of $\mathbb{R}^d$. There exists $j \in \mathbb{N}_d$ such that $x \in A_j$. We employ Eq. (4.4) to get

$$f(x,y) = g(x,y) + \lambda j \geq g(x,y) + \lambda, \quad \text{for all } y \in \mathbb{R}^{d'}.$$

Combining this inequality with condition (4.5), we obtain that

$$f(x,y) \geq g(x,y) + \lambda \geq g(x^*,y^*) + \lambda \geq g(0,y_0), \quad \text{for all } x \in \mathbb{R}^d\backslash\{0\} \text{ and all } y \in \mathbb{R}^{d'}. \tag{4.6}$$

Moreover, by the definition of f, we have that

$$f(0,y_0) = g(0,y_0) + \lambda\|0\|_0 = g(0,y_0). \tag{4.7}$$

Combining inequality (4.6) and Eq. (4.7) yields that $f(x,y) \geq f(0,y_0)$ for all $x \in \mathbb{R}^d\backslash\{0\}$ and for all $y \in \mathbb{R}^{d'}$. Furthermore, we have that $f(0,y) = g(0,y) \geq g(0,y_0) = f(0,y_0)$ for all $y \in \mathbb{R}^{d'}$. Therefore, the pair $(0,y_0) \in \mathbb{R}^d \times \mathbb{R}^{d'}$ is a global minimizer of f. The uniqueness of the global minimizer of f is guaranteed if a strict inequality for λ is imposed. □

We next consider a choice of the parameter λ with which the function $f(x,y)$ defined by (1.4) has a global minimizer with sparsity of a general level for the variable x.

Theorem 4.2. *Let $(x^*,y^*)\in\mathbb{R}^d\times\mathbb{R}^{d'}$ be a global minimizer of g, for some $\ell\in\mathbb{N}_d$, $(x',y')\in\Omega_\ell\times\mathbb{R}^{d'}$ be a minimizer of g on $\Omega_\ell\times\mathbb{R}^{d'}$ and $(x_j,y_j)\in\Omega_j\times\mathbb{R}^{d'}$ be a minimizer of g on $\Omega_j\times\mathbb{R}^{d'}$, for all $j\in\mathbb{Z}_\ell$. Suppose that*

$$g(x',y')-g(x^*,y^*)\le\frac{1}{\ell-j}[g(x_j,y_j)-g(x',y')],\quad \text{for all } j\in\mathbb{Z}_\ell. \tag{4.8}$$

If the parameter λ is chosen to satisfy the conditions

$$g(x',y')-g(x^*,y^*)\le\lambda\le\frac{1}{\ell-j}[g(x_j,y_j)-g(x',y')],\quad \text{for all } j\in\mathbb{Z}_\ell, \tag{4.9}$$

then (x',y') is a global minimizer of f on $\mathbb{R}^d\times\mathbb{R}^{d'}$.

Proof. It suffices to verify that

$$f(x',y')\le f(x,y),\quad \text{for all } (x,y)\in\mathbb{R}^d\times\mathbb{R}^{d'}. \tag{4.10}$$

Since $(x',y')\in\Omega_\ell\times\mathbb{R}^{d'}$ is a minimizer of g on $\Omega_\ell\times\mathbb{R}^{d'}$, we use the definition of Ω_ℓ to consider cases when $x'\in A_j$ for $j\in\mathbb{Z}_{\ell+1}$.

Step 1: We consider the case when $x'\in A_0$, that is, $x'=0$. We now show (4.10) with $x'=0$ by verifying it for all $x\in A_j$, for $j\in\mathbb{Z}_{d+1}$, and for all $y\in\mathbb{R}^{d'}$. The case for $j=0$ is trivial and we consider other cases. Since $(x',y')\in\Omega_\ell\times\mathbb{R}^{d'}$ is a minimizer of g on $\Omega_\ell\times\mathbb{R}^{d'}$ and $\|x'\|_0=0$, we have for all $x\in A_j$, $j=1,2,\dots,\ell$, and for all $y\in\mathbb{R}^{d'}$ that

$$f(x',y')=g(x',y')\le g(x,y)\le g(x,y)+\lambda j=f(x,y).$$

For all $x\in A_j$, $j=\ell+1,\ell+2,\dots,d$, and for all $y\in\mathbb{R}^{d'}$, by employing the first inequality of (4.9) and the assumption that (x^*,y^*) is a global minimizer of g on $\mathbb{R}^d\times\mathbb{R}^{d'}$, we obtain that

$$f(x',y')=g(x',y')\le g(x^*,y^*)+\lambda\le g(x,y)+\lambda\le g(x,y)+\lambda j=f(x,y).$$

We have shown (4.10) for the case when $x'=0$.

Step 2: We consider the case when $x'\in A_k$, for $k\in\mathbb{N}_\ell$. For all $j\in\mathbb{Z}_\ell$, the second inequality of (4.9) leads to $g(x',y')+\lambda\ell\le g(x_j,y_j)+\lambda j$. This together with the hypothesis that $(x_j,y_j)\in\Omega_j\times\mathbb{R}^{d'}$ is a minimizer of g on $\Omega_j\times\mathbb{R}^{d'}$, for all $j\in\mathbb{Z}_\ell$, ensures that for all $x\in A_j$, $j\in\mathbb{Z}_\ell$ and for all $y\in\mathbb{R}^{d'}$,

$$\begin{aligned}f(x',y')=g(x',y')+\lambda k&\le g(x',y')+\lambda\ell\\&\le g(x_j,y_j)+\lambda j\le g(x,y)+\lambda j=f(x,y).\end{aligned}$$

For all $x \in A_\ell$, since (x', y') is a minimizer of g on $\Omega_\ell \times \mathbb{R}^{d'}$ and $k \leq \ell$, we find that for all $x \in A_\ell$ and for all $y \in \mathbb{R}^{d'}$,

$$f(x', y') = g(x', y') + \lambda k \leq g(x, y) + \lambda k \leq g(x, y) + \lambda \ell = f(x, y).$$

For all $x \in A_j$, for $j = \ell + 1, \ell + 2, \ldots, d$, by using the first inequality of (4.9) and again the fact that (x^*, y^*) is a global minimizer of g, we derive that for all $x \in A_j$, for $j = \ell + 1, \ell + 2, \ldots, d$, and for all $y \in \mathbb{R}^{d'}$,

$$f(x', y') = g(x', y') + \lambda k \leq g(x^*, y^*) + \lambda(k + 1) \leq g(x, y) + \lambda j = f(x, y).$$

We have shown (4.10) for the cases $x' \in A_k$, for $k \in \mathbb{N}_\ell$.

Summarizing the above two steps of verification, we conclude that (4.10) holds true and thus, (x', y') is a global minimizer of f on $\mathbb{R}^d \times \mathbb{R}^{d'}$. □

Similarly to the comment made after the proof of Theorem 3.2, we now remark on condition (4.8). It is straightforward to verify that condition (4.8) is equivalent to

$$g(x', y') \leq \frac{g(x_j, y_j) + (\ell - j) g(x^*, y^*)}{\ell - j + 1}, \quad \text{for all } j \in \mathbb{Z}_\ell. \tag{4.11}$$

Inequality (4.11) shows that $g(x', y')$ should be bounded above by an weighted average of $g(x^*, y^*)$ and $g(x_j, y_j)$. By the definition of (x^*, y^*), (x', y') and $g(x_j, y_j)$, we derive that

$$g(x^*, y^*) \leq g(x', y') \leq g(x_j, y_j).$$

This shows that condition (4.8) is reasonable.

We illustrate Theorem 4.2 by presenting in the next corollary its special case when $\ell = 1$. The corollary gives a choice of the parameter λ which guarantees a global minimizer of f with sparsity of level 1 for the variable x.

Corollary 4.3. *Let $(x^*, y^*) \in \mathbb{R}^d \times \mathbb{R}^{d'}$ be a global minimizer of g, $(x', y') \in \Omega_1 \times \mathbb{R}^{d'}$ be a minimizer of g on $\Omega_1 \times \mathbb{R}^{d'}$, and $(0, y_0) \in A_0 \times \mathbb{R}^{d'}$ be a minimizer of g on $A_0 \times \mathbb{R}^{d'}$. Suppose that*

$$g(x', y') - g(x^*, y^*) \leq g(0, y_0) - g(x', y'). \tag{4.12}$$

If the parameter λ is chosen to satisfy the condition

$$g(x', y') - g(x^*, y^*) \leq \lambda \leq g(0, y_0) - g(x', y'), \tag{4.13}$$

then (x', y') is a global minimizer of f on $\mathbb{R}^d \times \mathbb{R}^{d'}$.

Proof. This result is obtained by specializing Theorem 4.2 to the special case where $\ell = 1$ and noticing that $A_0 = \{0\}$. In this case, we have that $x_0 = 0$. □

The next theorem connects a global minimizer of g with a global (or local) minimizer of f.

Theorem 4.4. *Let $(x^*, y^*) \in \mathbb{R}^d \times \mathbb{R}^{d'}$ be a global minimizer of g on $\mathbb{R}^d \times \mathbb{R}^{d'}$.*

(i) *If $x^* = 0$, then (x^*, y^*) is a global minimizer of f on $\mathbb{R}^d \times \mathbb{R}^{d'}$.*
(ii) *If for some $\ell \in \mathbb{N}_d$, $x^* \in A_\ell$ and*

$$g(x^*, y^*) + \lambda(\ell - j) \le g(x, y), \quad \text{for all } (x, y) \in A_j \times \mathbb{R}^{d'},\ j \in \mathbb{N}_\ell, \tag{4.14}$$

then (x^*, y^*) is a global minimizer of f on $\mathbb{R}^d \times \mathbb{R}^{d'}$.
(iii) *If $x^* \in A_d$, and for some $j \in \mathbb{Z}_{d+1}$ and for some $x' \in A_j$,*

$$g(x^*, y^*) + \lambda(d - j) > g(x', y^*), \tag{4.15}$$

then (x^*, y^*) is a local minimizer of f but not a global minimizer of f.

Proof. For both Items (i) and (ii), we shall show that

$$f(x^*, y^*) \le f(x, y), \quad \text{for all } (x, y) \in \mathbb{R}^d \times \mathbb{R}^{d'}. \tag{4.16}$$

Since $(x^*, y^*) \in \mathbb{R}^d \times \mathbb{R}^{d'}$ is a global minimizer of g, we have that

$$g(x^*, y^*) \le g(x, y), \quad \text{for all } (x, y) \in \mathbb{R}^d \times \mathbb{R}^{d'}. \tag{4.17}$$

(i) If $x^* = 0$, by the definition of $\|\cdot\|_0$, we have that $\|x^*\|_0 = 0$ and thus, $f(x^*, y^*) = g(x^*, y^*)$. Using this equation together with condition (4.17), we observe for all $(x, y) \in A_j \times \mathbb{R}^{d'}$, for $j \in \mathbb{Z}_{d+1}$ that

$$f(x^*, y^*) = g(x^*, y^*) \le g(x, y) \le g(x, y) + \lambda j = f(x, y).$$

This together with (2.2) ensures that (4.16) holds.

(ii) Since for some $\ell \in \mathbb{N}_d$, there holds $x^* \in A_\ell$ and since condition (4.14) holds, we have for all $(x, y) \in A_j \times \mathbb{R}^{d'}$, for $j \in \mathbb{Z}_\ell$ that

$$f(x^*, y^*) = g(x^*, y^*) + \lambda\ell = g(x^*, y^*) + \lambda(\ell - j) + \lambda j \le g(x, y) + \lambda j = f(x, y).$$

Moreover, by (4.17) we have for $(x, y) \in A_{\ell+j} \times \mathbb{R}^{d'}$, $j \in \mathbb{Z}_{d-\ell+1}$ that

$$f(x^*, y^*) = g(x^*, y^*) + \lambda\ell \le g(x, y) + \lambda(\ell + j) = f(x, y).$$

Again, according to (2.2), we find that (4.16) holds.

(iii) Since $x^* \in A_d$ and (x^*, y^*) is a global minimizer of g on $\mathbb{R}^d \times \mathbb{R}^{d'}$, we have for all $(x, y) \in A_d \times \mathbb{R}^{d'}$ that

$$f(x^*, y^*) = g(x^*, y^*) + \lambda d \le g(x, y) + \lambda d = f(x, y).$$

That is,

$$f(x^*, y^*) \leq f(x, y), \quad \text{for all } (x, y) \in A_d \times \mathbb{R}^{d'}. \tag{4.18}$$

According to Proposition 2.1(iv), we know that A_d is an open set. Thus, $A_d \times \mathbb{R}^{d'}$ is an open set. Therefore, inequality (4.18) ensures that the pair (x^*, y^*) is a local minimizer of f.

It remains to show that (x^*, y^*) is not a global minimizer of f. By condition (4.15), we have for some $j \in \mathbb{Z}_{d+1}$ and for some $x' \in A_j$ that $g(x^*, y^*) + \lambda d > g(x', y^*) + \lambda j$. This together with the fact $x^* \in A_d$ and $x' \in A_j$ ensures that

$$f(x^*, y^*) = g(x^*, y^*) + \lambda d > g(x', y^*) + \lambda j = f(x', y^*).$$

This implies that (x^*, y^*) is not a global minimizer of f. □

In the next theorem, we provide properties of a global minimizer of f.

Theorem 4.5. *Let $(x^*, y^*) \in \mathbb{R}^d \times \mathbb{R}^{d'}$ be a global minimizer of f on $\mathbb{R}^d \times \mathbb{R}^{d'}$.*

(i) *If for some $\ell \in \mathbb{Z}_{d+1}$, $x^* \in A_\ell$, then (x^*, y^*) is a minimizer of g on $\Omega_\ell \times \mathbb{R}^{d'}$.*
(ii) *If $(x^*, y^*) \in \mathbb{R}^d \times \mathbb{R}^{d'}$ is not a global minimizer of g, then $x^* \in \Omega_{d-1}$.*

Proof. (i) Since for some $\ell \in \mathbb{Z}_{d+1}$, $(x^*, y^*) \in A_\ell \times \mathbb{R}^{d'}$ is a global minimizer of f, by (4.16), we have that

$$g(x^*, y^*) + \lambda \ell = f(x^*, y^*) \leq f(x, y) = g(x, y) + \lambda j,$$
$$\text{for all } (x, y) \in A_j \times \mathbb{R}^{d'},\ j \in \mathbb{Z}_{\ell+1}.$$

It follows that

$$g(x^*, y^*) + \lambda(\ell - j) \leq g(x, y), \quad \text{for all } (x, y) \in A_j \times \mathbb{R}^{d'},\ j \in \mathbb{Z}_{\ell+1}.$$

Using this inequality and noting that $\lambda(\ell - j) \geq 0$, for $j \in \mathbb{Z}_{\ell+1}$, we have that

$$g(x^*, y^*) \leq g(x^*, y^*) + \lambda(\ell - j) \leq g(x, y), \quad \text{for all } (x, y) \in A_j \times \mathbb{R}^{d'},\ \ell \in \mathbb{Z}_{\ell+1}.$$

This ensures that (x^*, y^*) is a minimizer of g on $\Omega_\ell \times \mathbb{R}^{d'}$.

(ii) We prove this assertion by contradiction. Assume to the contrary that $x^* \notin \Omega_{d-1}$. Since $x^* \in \mathbb{R}^d$, $\mathbb{R}^d = \Omega_{d-1} \cup A_d$ and $\Omega_{d-1} \cap A_d = \emptyset$, we must have that $x^* \in A_d$. By Item (i) of this theorem with $\ell = d$, we conclude that (x^*, y^*) is a minimizer of g on $\Omega_d \times \mathbb{R}^{d'}$. Noting that $\Omega_d = \mathbb{R}^d$, we confirm that (x^*, y^*) is a global minimizer of g on $\mathbb{R}^d \times \mathbb{R}^{d'}$, a contradiction. This contradiction ensures that $x^* \in \Omega_{d-1}$. □

Theorem 4.5(ii) provides a sufficient condition which guarantees that a global minimizer of f is sparse. The next corollary follows immediately from Theorem 4.5(ii).

Corollary 4.6. *If $(x^*, y^*) \in \mathbb{R}^d \times \mathbb{R}^{d'}$ is a global minimizer of f, then, either $x^* \in \Omega_{d-1}$ or (x^*, y^*) is a global minimizer of g on $\mathbb{R}^d \times \mathbb{R}^{d'}$.*

We next present an understanding of the relation between the local minimizers of minimization problem (4.1) for a fixed parameter $\lambda > 0$ and the constrained minimization problem without the term involving the ℓ_0-norm. We now define precisely the constrained minimization problem. For a given index set $\mathcal{I}$, we introduce a minimization problem on $\mathcal{C}_\mathcal{I} \times \mathbb{R}^{d'}$ by

$$\min\{g(x, y) : (x, y) \in \mathcal{C}_\mathcal{I} \times \mathbb{R}^{d'}\}. \tag{4.19}$$

We need a technical lemma to compare the support of a given vector with that of vectors in its close neighborhood.

Lemma 4.7. *If $x^* \in \mathbb{R}^d$ is given, then there exists $\delta_0 > 0$ such that*

(i) *for all $x \in \mathcal{B}(x^*, \delta_0)$, there holds $S(x^*) \subseteq S(x)$;*
(ii) *for all $x \in \mathcal{B}(x^*, \delta_0) \cap \mathcal{C}_\mathcal{I}$ with $\mathcal{I} := S(x^*)$, there holds $S(x^*) = S(x)$.*

Proof. For a fixed number $0 < \mu \le 1/2$, we let $\delta_0 := \min\{\mu|x_j^*| : j \in S(x^*)\}$. Clearly, $\delta_0 > 0$. Suppose that $i \in S(x^*)$. For all $x \in \mathcal{B}(x^*, \delta_0)$, we have that

$$|x_i^* - x_i| \le \|x - x^*\|_2 \le \delta_0 \quad \text{and} \quad |x_i| \ge |x_i^*| - |x_i^* - x_i| \ge \delta_0 > 0.$$

This implies that $i \in S(x)$. Thus, $S(x^*) \subseteq S(x)$, which proves Item (i).

To show Item (ii), we note that when $x \in \mathcal{C}_\mathcal{I}$, by the definition of $\mathcal{C}_\mathcal{I}$, there holds $S(x) \subseteq S(x^*)$. This together with Item (i) yields $S(x^*) = S(x)$. □

A pair $(x^*, y^*) \in \mathbb{R}^d \times \mathbb{R}^{d'}$ is called a local minimizer of the minimization problem (4.1), if there exists a $\delta > 0$ such that

$$f(x^*, y^*) \le f(x, y), \quad \text{for all } x \in \mathcal{B}(x^*, \delta),\ y \in \mathcal{B}(y^*, \delta).$$

Here comes the theorem concerning the relation between local minimizers of minimization problems (4.1) and (4.19).

Theorem 4.8. *Suppose that $(x^*, y^*) \in \mathbb{R}^d \times \mathbb{R}^{d'}$ is given. The pair (x^*, y^*) is a local minimizer of the minimization problem (4.1) with a fixed parameter $\lambda > 0$ if and only if (x^*, y^*) is a local minimizer of the constrained minmization problem (4.19) with $\mathcal{I} := S(x^*)$.*

Proof. Suppose that the pair (x^*, y^*) is a local minimizer of the minimization problem (4.1) with a fixed parameter $\lambda > 0$ and we show that the pair is a local minimizer of the constrained minimization problem (4.19) with $\mathcal{I} = S(x^*)$. We prove this by contradiction. Since $\mathcal{I} := S(x^*)$, we note that $x^* \in \mathcal{C}_\mathcal{I}$. Assume to the contrary that the pair (x^*, y^*) is not a local minimizer of the constrained minimization problem (4.19) with $\mathcal{I} := S(x^*)$. According to the definition of the

local minimizer of g, we observe that for any $\delta > 0$, there exist $x_\delta \in \mathcal{B}(x^*, \delta) \cap \mathcal{C}_\mathcal{I}$ and $y_\delta \in \mathcal{B}(y^*, \delta)$ such that $g(x^*, y^*) > g(x_\delta, y_\delta)$. Item (ii) of Lemma 4.7 ensures that there exists $\delta_0 > 0$ such that for all $x \in \mathcal{B}(x^*, \delta_0) \cap \mathcal{C}_\mathcal{I}$ with $\mathcal{I} := S(x^*)$, there holds $S(x^*) = S(x)$. Hence, for any $0 < \delta < \delta_0$, there exist $x_\delta \in \mathcal{B}(x^*, \delta) \cap \mathcal{C}_\mathcal{I}$ and $y_\delta \in \mathcal{B}(y^*, \delta)$ such that $S(x^*) = S(x_\delta)$, which implies $\|x^*\|_0 = \|x_\delta\|_0$, and $g(x^*, y^*) > g(x_\delta, y_\delta)$. This implies that for the fixed parameter $\lambda > 0$, there holds

$$f(x^*, y^*) = g(x^*, y^*) + \lambda\|x^*\|_0 > g(x_\delta, y_\delta) + \lambda\|x_\delta\|_0 = f(x_\delta, y_\delta).$$

This violates the assumption that (x^*, y^*) is a local minimizer of the minimization problem (4.1) with the parameter $\lambda > 0$.

Now, suppose that (x^*, y^*) is a local minimizer of the constrained minimization problem (4.19) with $\mathcal{I} := S(x^*)$ and we prove that (x^*, y^*) is a local minimizer of the minimization problem (4.1) with the fixed parameter $\lambda > 0$. We proceed the proof by considering two cases $x \in \mathcal{C}_\mathcal{I}$ and $x \notin \mathcal{C}_\mathcal{I}$ separately. We first consider the case when $x \in \mathcal{C}_\mathcal{I}$. The definition of the local minimizer ensures that there exists a $\delta_1 > 0$ such that

$$g(x^*, y^*) \leq g(x, y), \quad \text{for all } x \in \mathcal{B}(x^*, \delta_1) \cap \mathcal{C}_\mathcal{I}, \quad y \in \mathcal{B}(y^*, \delta_1). \tag{4.20}$$

By Item (ii) of Lemma 4.7, we have that there exists a $\delta_0 > 0$ such that for all $x \in \mathcal{B}(x^*, \delta_0) \cap \mathcal{C}_\mathcal{I}$, there holds $S(x^*) = S(x)$. This implies that for all $x \in \mathcal{B}(x^*, \delta_0) \cap \mathcal{C}_\mathcal{I}$, there holds $\|x^*\|_0 = \|x\|_0$. Choose $\delta := \min\{\delta_0, \delta_1\}$. Then, for all $x \in \mathcal{B}(x^*, \delta) \cap \mathcal{C}_\mathcal{I}$ and for all $y \in \mathcal{B}(y^*, \delta)$, by (4.20), there holds

$$f(x^*, y^*) = g(x^*, y^*) + \lambda\|x^*\|_0 \leq g(x, y) + \lambda\|x^*\|_0 = g(x, y) + \lambda\|x\|_0 = f(x, y).$$

This yields

$$f(x^*, y^*) \leq f(x, y), \quad \text{for all } x \in \mathcal{B}(x^*, \delta) \cap \mathcal{C}_\mathcal{I}, \quad y \in \mathcal{B}(y^*, \delta). \tag{4.21}$$

We next consider the case when $x \notin \mathcal{C}_\mathcal{I}$. By Item (ii) of Proposition 2.7, we conclude that there exists $\delta_2 > 0$ such that

$$\|x\|_0 \geq \|x^*\|_0 + 1, \quad \text{for all } x \in \mathcal{B}(x^*, \delta_2) \backslash \mathcal{C}_\mathcal{I}. \tag{4.22}$$

Since g is continuous, for $\lambda > 0$, there exists $\delta_3 > 0$ such that

$$g(x^*, y^*) \leq g(x, y) + \lambda, \quad \text{for all } x \in \mathcal{B}(x^*, \delta_3) \backslash \mathcal{C}_\mathcal{I}, \quad y \in \mathcal{B}(y^*, \delta_3). \tag{4.23}$$

Choose $\delta := \min\{\delta_j : j = 0, 1, 2, 3\}$. By employing inequality (4.23) and then inequality (4.22), we have for all $x \in \mathcal{B}(x^*, \delta) \backslash \mathcal{C}_\mathcal{I}$, $y \in \mathcal{B}(y^*, \delta)$, that

$$\begin{aligned} f(x^*, y^*) &= g(x^*, y^*) + \lambda\|x^*\|_0 \leq g(x, y) + \lambda\|x^*\|_0 + \lambda \\ &\leq g(x, y) + \lambda\|x\|_0 = f(x, y). \end{aligned}$$

That is,

$$f(x^*, y^*) \leq f(x, y), \quad \text{for all } x \in \mathcal{B}(x^*, \delta) \backslash \mathcal{C}_\mathcal{I}, \quad y \in \mathcal{B}(y^*, \delta). \tag{4.24}$$

Combining inequalities (4.21) and (4.24) leads to $f(x^*, y^*) \leq f(x, y)$, for all $x \in \mathcal{B}(x^*, \delta)$, $y \in \mathcal{B}(y^*, \delta)$, which implies that the pair (x^*, y^*) is a local minimizer of the minimization problem (4.1) with a fixed parameter $\lambda > 0$. □

Theorem 4.8 is useful in developing efficient numerical algorithms for solving non-convex minimization problems involved functions f having the form (4.2) or (4.3) and analyzing convergence of the algorithms, since in such cases Theorem 4.8 guarantees that the non-convex minimization problems are reduced to convex minimization problems on certain support sets. We next present two corollaries that specialize Theorem 4.8 to functions f having a special form (4.2) or (4.3).

Corollary 4.9. *Suppose that $(x^*, y^*) \in \mathbb{R}^d \times \mathbb{R}^{d'}$ is given. The pair (x^*, y^*) is a local minimizer of the minimization problem* (4.1) *with f being defined by* (4.2) *for a fixed parameter $\lambda > 0$ if and only if (x^*, y^*) is a global minimizer of the convex minimization problem* (4.19) *with $\mathcal{I} := S(x^*)$ and $g(x, y) := \phi(y) + \mu\|x - Dy\|_2^2$, where ϕ is the convex function and D is $d \times d'$ matrix appearing in* (4.2).

The sufficient condition for a pair (x^*, y^*) to be a local minimizer of the minimization problem (4.1) presented in Corollary 4.9 for a special example of convex function g was obtained in [34, Proposition 2.3].

Corollary 4.10. *Suppose that $(x^*, y^*) \in \mathbb{R}^d \times \mathbb{R}^{d'}$ is given. The pair (x^*, y^*) is a local minimizer of the minimization problem* (4.1) *with f being defined by* (4.3) *for a fixed parameter $\lambda > 0$ if and only if (x^*, y^*) is a global minimizer of the convex minimization problem* (4.19) *with $\mathcal{I} := S(x^*)$ and $g(x, y) := \phi(y) + \mu\|x - Dy\|_1$, where ϕ is the convex function and D is $d \times d'$ matrix appearing in* (4.3).

5. Final Remarks

We briefly discuss possible extension of the results presented in previous sections involving matrix M and comment on potential uses of the main results of this paper.

We first elaborate an extension of the results involving matrix M which has been assumed to satisfy hypothesise (2.8). We now suppose that the matrix M has an arbitrary rank r with $0 < r \leq \min\{d, m\}$. In this general case, the singular value decomposition of M can be used to remove hypothesise (2.8) on M. Clearly, matrix M has the singular value decomposition

$$M = U\Lambda V^*, \tag{5.1}$$

where U is a $d \times d$ unitary matrix, V is an $m \times m$ unitary matrix and Λ is a $d \times m$ diagonal matrix having the nonzero diagonal entries $\lambda_1 \geq \cdots \geq \lambda_r > 0$. We can first extend the results in Secs. 3 and 4 involving matrix M to the case when $M = \Lambda$. The regularization problem (1.2) with f having the form (1.1) for this special case is to impose the regularization only for the first r components of the variable x and leave its remaining $m - r$ components not regularized. Results for the general case

can be obtained by using appropriate changes of variables with the two unitary matrices U and V from the singular value decomposition (5.1). We would leave details of the extension to the interested reader.

We have proved rigorously that if the regularization parameter is chosen appropriately, the ℓ_0 norm regularization will lead to sparse solutions, a result previously validated empirically. Regularization parameter choice strategies presented in Secs. 3 and 4 are all theoretical since in general it is not realistic to know a global minimizer of function g. Nevertheless, these results provide insights into the connection between the choice of the regularization parameter with the locations of global minimizers of f. They can serve as a guidance for further designing practical parameter choice strategies. For example, one may estimate a global minimizer of g via certain means. In such a case, our parameter strategies may lead to practical uses. This requires further investigation.

Finally, we indicate that Theorem 4.8 and especially Corollaries 4.9 and 4.10 are useful in developing efficient algorithms for finding local minimizers of the regularized non-convex optimization problems. The essence of Corollaries 4.9 and 4.10 is that they identify a local minimizer of a non-convex optimization problem with that of a convex optimization problem. Hence, finding a local minimizer of a non-convex optimization problem can be done by finding a local minimizer of a convex optimization problem. In general, solving a convex optimization problem is much easier than solving a non-convex optimization problem.

Acknowledgments

This work is supported in part by US National Science Foundation under grant DMS-1912958 and by US National Institutes of Health under grant R21CA263876.

References

[1] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces* (Springer, New York, 2010).
[2] R. H. Chan, S. D. Riemenschneider, L. Shen and Z. Shen, Tight frame: An efficient way for highresolution image reconstruction, *Appl. Comput. Harmon. Anal.* **17** (2004) 91–115.
[3] Z. Chen, C. A. Micchelli and Y. Xu, *Multiscale Methods for Fredholm Integral Equations* (Cambridge University Press, Cambridge, 2015).
[4] J. Chen, S. Pereverzyev Jr. and Y. Xu, Aggregation of regularized solutions from multiple observation models, *Inverse Probl.* **31** (2015) 075005.
[5] Z. Chen, Y. Xu and H. Yang, Fast collocation methods for solving ill-posed integral equations of the first kind, *Inverse Probl.* **24** (2008) 065007.
[6] D. Q. Dai, L. Shen, Y. Xu and N. Zhang, Noisy 1-bit compressive sensing: Models and algorithms, *Appl. Comput. Harmon. Anal.* **40** (2016) 1–32.
[7] I. Daubechies, *Ten Lectures on Wavelets*, CBMS-NSF Regional Conference Series in Applied Mathematics, Vol. 61 (SIAM, Philadelphia, 1992).
[8] L. He and S. Schaefer, Mesh denoising via L_0 Minimization, *ACM Trans. Graph.* **32** (2013) 64.

[9] B. Hofmann and P. Mathé, Tikhonov regularization with oversmoothing penalty for non-linear ill-posed problems in Hilbert scales, *Inverse Probl.* **34** (2018) 015007.
[10] J. Huang, H. Hom, Y. Jiao, Y. Liu and X. Lu, A constructive approach to L_0 penalized regression, *J. Mach. Learn. Res.* **19** (2018) 1–37.
[11] W. Jiang, F. Nie and H. Huang, Robust dictionary learning with capped ℓ_1-norm, in *Proc. Twenty-Fourth Int. Joint Conf. Artificial Intelligence* (AAAI Press, 2015), pp. 3590–3596.
[12] R. Kress, *Linear Integral Equations* (Springer-Verlag, New York, 1989).
[13] Y. LeCun, Y. Bengio and G. Hinton, Deep learning, *Nature* **521** (2015) 436–444.
[14] Q. Lian, L. Shen, Y. Xu and L. Yang, Filters of wavelets on invariant sets for image denoising, *Appl. Anal.* **90** (2011) 1299–1322.
[15] J. Liu, P. C. Cosman and B. D. Rao, Robust Linear Regression via L_0 Regularization, *IEEE Trans. Signal Process.* **66** (2018) 698–713.
[16] J. López, K. De Brabanter, J. R. Dorronsoro and J. A. K. Suykens, Sparse LS-SVMs with L_0-norm minimization, in *ESANN 2011 Proc. European Symp. Artificial Neural Networks, Computational Intelligence and Machine Learning* (Bruges, Belgium, 2011), pp. 27–29.
[17] C. Louizos, M. Welling and D. P. Kingma, Learning sparse neural networks through L_0 regularization, *Proceedings of the 6th International Conference on Learning Representations*, Vancouver, Canada, 30 April–3 May 2018, pp. 1–13.
[18] S. Mallat, *A Wavelet Tour of Signal Processing*, 2nd edn. (Academic Press, 1999).
[19] C. A. Micchelli and Y. Xu, Using the matrix refinement equation for the construction of wavelets on invariant sets, *Appl. Comput. Harm. Anal.* **1** (1994) 391–401.
[20] C. A. Micchelli and Y. Xu, Reconstruction and decomposition algorithms for biorthogonal multiwavelets, *Multidimensional Systems and Signal Processing* **8** (1997) 31–69.
[21] J. Pan, J. Lim, Z. Su and M.-H. Yang, L_0-regularized object representation for visual tracking, *Proceedings of the 2014 British Machine Vision Conference* (Nottingham, England, BMVA Press, 2014), pp. 1–12.
[22] A. Ron and Z. Shen, Affine systems in $\ell_2(\mathbf{R}^d)$: The analysis of the analysis operator, *J. Funct. Anal.* **148** (1997) 408–447.
[23] L. Rudin, S. Osher and E. Fatemi, Nonlinear total variation based noise removal algorithms, *Physica D: Nonlinear Phenomena* **60** (1992) 259–268.
[24] L. Shen, Y. Xu and N. Zhang, An approximate sparsity model for inpainting, *Appl. Comput. Harmon. Anal.* **37** (2014) 171–184.
[25] L. Shen, Y. Xu and X. Zeng, Wavelet inpainting with the l0 sparse regularization, *Appl. Comput. Harmon. Anal.* **41** (2016) 26–53.
[26] G. Strang, The discrete cosine transform, *SIAM Rev.* **41** (1999) 135–147.
[27] Y. Sun, S. Schaefer and W. Wang, Denoising point sets via L_0 minimization, *Comput. Aided Geom. Des.* **35–36** (2015) 2–15.
[28] M. Unser, A representer theorem for deep neural networks, *J. Mach. Learn. Res.* **20** (2019) 1–30.
[29] Z. Wang, Q. Ling, T. S. Huang, Learning deep ℓ_0 encoders, in *Proc. Thirtieth AAAI Conf. Artificial Intelligence* (2016).
[30] W. Wang, S. Lu, B. Hofmann and J. Cheng, Tikhonov regularization with ℓ_0-term complementing a convex penalty: ℓ_1-convergence under sparsity constraints, *J. Inverse Ill-Posed Probl.* **27**(4) (2019) 575–590.
[31] W. Wang, S. Lu, H. Mao and J. Cheng, Multi-parameter Tikhonov regularization with the ℓ_0 sparsity constraint, *Inverse Probl.* **29** (2013) 065018.

[32] D. Wipf and B. Rao, ℓ_0-norm minimization for basis selection, in *Proceedings of the 17th International Conference on Neural Information Processing Systems* (*NIPS*) (2005), pp. 1513–1520.

[33] T. Wu and Y. Xu, Inverting incomplete Fourier transforms by a sparse regularization and applications in seismic wavefield modeling, *J. Sci. Comput.* **92**(2) (2022) 1–35.

[34] X. Zeng, L. Shen and Y. Xu, A convergent fixed-point proximity algorithm accelerated by FISTA for the ℓ_0 sparse recovery problem, in *Imaging, Vision and Learning Based on Optimization and PDEs: IVLOPDE*, eds. X.-C. Tai, G. Bae and M. Lysaker (Springer, 2018).

[35] X. Zeng, L. Shen, Y. Xu and J. Lu, Matrix completion via minimizing an approximate rank, *Anal. Appl.* **17** (2019) 689–713.

[36] Z. Zheng, Y. Fan and J. Lv, High dimensional thresholded regression and shrinkage effect, *J. Roy. Stat. Soc. Ser. B* (*Stat. Methodol.*) **76** (2014) 627–649.

[37] W. Zheng, S. Li, A. Krol, C. R. Schmidtlein, X. Zeng and Y. Xu, Sparsity promoting regularization for effective noise suppression in SPECT image reconstruction, *Inverse Probl.* **35** (2019) 115011.