SIAM J. OPTIM.
Vol. 23, No. 3, pp. 1480–1509

SPARSITY CONSTRAINED NONLINEAR OPTIMIZATION: OPTIMALITY CONDITIONS AND ALGORITHMS*

AMIR BECK† AND YONINA C. ELDAR‡

Abstract. This paper treats the problem of minimizing a general continuously differentiable function subject to sparsity constraints. We present and analyze several different optimality criteria which are based on the notions of stationarity and coordinatewise optimality. These conditions are then used to derive three numerical algorithms aimed at finding points satisfying the resulting optimality criteria: the iterative hard thresholding method and the greedy and partial sparse-simplex methods. The first algorithm is essentially a gradient projection method, while the remaining two algorithms are of a coordinate descent type. The theoretical convergence of these techniques and their relations to the derived optimality conditions are studied. The algorithms and results are illustrated by several numerical examples.

Key words. optimality conditions, sparsity constrained problems, stationarity, numerical methods, compressed sensing

AMS subject classifications. 90C26, 90C46, 90C90

DOI. 10.1137/120869778

1. Introduction.

Sparsity has long been exploited in signal processing, applied mathematics, statistics, and computer science for tasks such as compression, denoising, model selection, image processing, and more [10, 11, 17, 20, 22, 25, 26]. Recent years have witnessed a growing interest in sparsity-based processing methods and algorithms for sparse recovery [3, 2, 1, 28]. Despite the great interest in exploiting sparsity in various applications, most of the work to date has focused on recovering sparse data represented by a vector $\mathbf{x} \in \mathbb{R}^n$ from linear measurements of the form $\mathbf{b} = \mathbf{A}\mathbf{x}$. For example, the rapidly growing field of compressed sensing [12, 8, 19] considers recovery of a sparse $\mathbf{x}$ from a small set of linear measurements $\mathbf{b} \in \mathbb{R}^m$, where m is usually much smaller than n. Since in practice the measurements are contaminated by noise, a typical approach to recover $\mathbf{x}$ is to seek a sparse vector $\mathbf{x}$ that minimizes the quadratic function $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$.

In this paper we study the more general problem of minimizing a continuously differentiable objective function subject to a sparsity constraint. More specifically, we consider the problem

$$(\mathrm{P}): \quad \begin{array}{ll} \min & f(\mathbf{x}) \\ \text{s.t.} & \|\mathbf{x}\|_0 \le s, \end{array}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a continuously differentiable function, $s > 0$ is an integer smaller than n, and $\|\mathbf{x}\|_0$ is the ℓ_0 norm of $\mathbf{x}$, which counts the number of nonzero components

*Received by the editors March 12, 2012; accepted for publication (in revised form) May 13, 2013; published electronically July 23, 2013.
http://www.siam.org/journals/siopt/23-3/86977.html

†Faculty of Industrial Engineering and Management, Technion–Israel Institute of Technology, Haifa 32000, Israel (becka@ie.technion.ac.il). This author's research was partially supported by the Israel Science Foundation under grant 253/12.

‡Faculty of Electrical Engineering, Technion–Israel Institute of Technology, Haifa 32000, Israel (yonina@ee.technion.ac.il). This author's research was partially supported by the Intel Collaborative Research Institute for Computational Intelligence (ICRI-CI), by the Ollendorf Foundation, by the Israel Science Foundation under grant 170/10, and by the Semiconductor Research Corporation (SRC) through the Texas Analog Center of Excellence (TxACE).

in $\mathbf{x}$. We do not assume that f is a convex function. This, together with the fact that the constraint function is nonconvex, and in fact is not even continuous, renders the problem quite difficult. Our goal in this paper is to study necessary optimality conditions for problem (P) and to develop algorithms that find points satisying these conditions for general choices of f.

Two instances of problem (P) that have been considered in previous literature and will serve as prototype models throughout the paper are described in the following two examples.

Example 1.1 (compressive sensing). As mentioned above, compressed sensing is concerned with recovery of a sparse vector $\mathbf{x}$ from linear measurements $\mathbf{Ax} = \mathbf{b}$, where $\mathbf{A} \in \mathbb{R}^{m\times n}, \mathbf{b} \in \mathbb{R}^m$, and m is usually much smaller than n. It is well known that under suitable conditions on $\mathbf{A}$, only the order of $s \log n$ measurements are needed to recover $\mathbf{x}$ [29]. When noise is present in the measurements, it is natural to consider the corresponding optimization problem (P) with the objective function given by

$$f_{\mathrm{LI}}(\mathbf{x}) \equiv \|\mathbf{Ax} - \mathbf{b}\|^2.$$

A variety of algorithms have been proposed in order to approximate the solution to this problem [27, 28]. One popular approach is to replace the ℓ_0 norm with the convex ℓ_1 norm, which results in a convex problem when the objective function f is convex. A variety of different greedy methods have also been proposed, such as the matching pursuit (MP) [21] and orthogonal MP (OMP) [9] algorithms. We will relate our methods to these approaches in section 3.2.1. Another method that was proposed recently and is related to our approach below is the iterative hard thresholding (IHT) algorithm [6], also referred to as the M-sparse method. In [6] the authors consider a majorization-minimization approach to solve (P) with $f = f_{\mathrm{LI}}$ and show that the resulting method converges to a local minimum of (P) as long as the spectral norm of $\mathbf{A}$ satisfies $\|\mathbf{A}\| < 1$. This algorithm is essentially a gradient projection method with stepsize 1. In section 3.1 we will revisit the IHT method and show how it can be applied to the general formulation (P) as well as discuss the quality of the limit points of the sequence generated by the algorithm.

Although linear measurements are the most popular in the literature, recently, attention has been given to quadratic measurements. Sparse recovery problems from quadratic measurements arise in a variety of problems in optics, as we discuss in the next example.

Example 1.2. Recovery of sparse vectors from quadratic measurements has been treated recently in the context of subwavelength optical imaging [14, 24]. In these problems the goal is to recover a sparse image from its far-field measurements, where due to the laws of physics the relationship between the (clean) measurement and the unknown image is quadratic. In [24] the quadratic relationship is a result of using partially incoherent light. The quadratic behavior of the measurements in [14] is a result of coherent diffractive imaging in which the image is recovered from its intensity pattern. Under an appropriate experimental setup, this problem amounts to reconstruction of a sparse signal from the magnitude of its Fourier transform.

Mathematically, both problems can be described as follows. Given m symmetric matrices $\mathbf{A}_1, \ldots, \mathbf{A}_m \in \mathbb{R}^{n\times n}$, find a vector $\mathbf{x}$ satisfying

$$\begin{aligned} \mathbf{x}^T \mathbf{A}_i \mathbf{x} &\approx c_i, \quad i = 1, \ldots, m, \\ \|\mathbf{x}\|_0 &\le s. \end{aligned}$$

This problem can be written in the form of problem (P) with

$$f_{\mathrm{QU}}(\mathbf{x}) \equiv \sum_{i=1}^{m} \left(\mathbf{x}^T \mathbf{A}_i \mathbf{x} - c_i\right)^2.$$

In this case, the objective function is nonconvex and quartic.

Quadratic measurements appear more generally in phase retrieval problems, in which a signal $\mathbf{x}$ is to be recovered from the magnitude of its measurements $y_i = |\mathbf{d}_i^* \mathbf{x}|$, where each measurement is a linear transform of the input $\mathbf{x} \in \mathbb{R}^n$. Note that $\mathbf{d}_i$ are complex-valued, that is, $\mathbf{d}_i \in \mathbb{C}^n$. Denoting by b_i the corresponding noisy measurements, and assuming a sparse input, our goal is to minimize $\sum_{i=1}^m (b_i^2 - |\mathbf{d}_i^* \mathbf{x}|^2)^2$ subject to the constraint that $\|\mathbf{x}\|_0 \leq s$ for some s, where m is the number of measurements. The objective function has the same structure as f_{QU} with $\mathbf{A}_i = \Re(\mathbf{d}_i)\Re(\mathbf{d}_i)^T + \Im(\mathbf{d}_i)\Im(\mathbf{d}_i)^T$. In [24], an algorithm was developed to treat such problems based on a semidefinite relaxation and low-rank matrix recovery. However, for large-scale problems, the method is not efficient and is difficult to implement. An alternative algorithm was designed in [14] based on a greedy search. This approach requires solving a nonconvex optimization program in each internal iteration.

To conclude this example, we note that the problem of recovering a signal from the magnitude of its Fourier transform has been studied extensively in the literature. Many methods have been developed for phase recovery [18] which often rely on prior information about the signal, such as positivity or support constraints. One of the most popular techniques is based on alternating projections, where the current signal estimate is transformed back and forth between the object and the Fourier domains. The prior information and observations are used in each domain in order to form the next estimate. Two of the main approaches of this type are those of Gerchberg and Saxton [16] and Fienup [15]. In general, these methods are not guaranteed to converge and often require careful parameter selection and sufficient signal constraints in order to provide a reasonable result.

In this paper we present a uniform approach to treating problems of the form (P). Necessary optimality conditions for problems consisting of minimizing differentiable (possibly nonconvex) objective functions over convex feasibility sets are well known [4]. These conditions are also very often the basis for efficient algorithms for solving the respective optimization problems. However, classical results on nonconvex optimization do not cover the case of sparsity constraints, which are neither convex nor continuous. In section 2 we derive three classes of necessary optimality conditions for problem (P): basic feasibility, L-stationarity, and coordinatewise (CW) optimality. We then show that CW optimality implies L-stationarity for suitable values of L, and they both imply the basic feasibility property. In section 3 we present two classes of algorithms for solving (P). The first technique is a generalization of the IHT method and is based on the notion of L-stationarity. Under appropriate conditions we show that the limit points of the algorithm are L-stationary points. The second class of methods are based on the concept of CW optimality. These are basically coordinate descent type algorithms which update the support at each iteration by one or two variables. Due to their resemblance to the celebrated simplex method for linear programming, we refer to these methods as "sparse-simplex" algorithms. As we show, these techniques are as simple as the IHT method while obtaining stronger optimality guarantees. We prove the convergence results of the various algorithms, establishing that the limit points of each of the methods

satisfy the respective necessary optimality conditions. A MATLAB implementation of the sparse-simplex approaches, as well as documentation, can be found at http://iew3.technion.ac.il/~becka/papers/sparse_simplex_package.zip.

2. Necessary optimality conditions.

2.1. Notation and assumptions. For a given vector $\mathbf{x} \in \mathbb{R}^n$ and an index set $R \subseteq \{1, \ldots, n\}$, we denote by $\mathbf{x}_R$ the subvector of $\mathbf{x}$ corresponding to the indices in R. For example, if $\mathbf{x} = (4, 5, 2, 1)^T$ and $R = \{1, 3\}$, then $\mathbf{x}_R = (4, 2)^T$. The support set of $\mathbf{x}$ is defined by

$$I_1(\mathbf{x}) \equiv \{i : x_i \neq 0\},$$

and its complement is

$$I_0(\mathbf{x}) \equiv \{i : x_i = 0\}.$$

We denote by C_s the set of vectors $\mathbf{x}$ that are at most s-sparse:

$$C_s = \{\mathbf{x} : \|\mathbf{x}\|_0 \leq s\}.$$

For a vector $\mathbf{x} \in \mathbb{R}^n$ and $i \in \{1, 2, \ldots, n\}$, the ith largest absolute value component in $\mathbf{x}$ is denoted by $M_i(\mathbf{x})$, so that in particular

$$M_1(\mathbf{x}) \geq M_2(\mathbf{x}) \geq \cdots \geq M_n(\mathbf{x}).$$

Also, $M_1(\mathbf{x}) = \max_{i=1,\ldots,n} |x_i|$ and $M_n(\mathbf{x}) = \min_{i=1,\ldots,n} |x_i|$.

Throughout the paper we make the following assumption. The vector $\mathbf{e}_i \in \mathbb{R}^n$ is the n-length column vector whose ith component is one while all the others are zeros.

Assumption 1. The objective function f is lower bounded. That is, there exists $\gamma \in \mathbb{R}$ such that $f(\mathbf{x}) \geq \gamma$ for all $\mathbf{x} \in \mathbb{R}^n$.

2.2. Basic feasibility. Optimality conditions have an important theoretical role in the study of optimization problems. From a practical point of view, they are the basis for most numerical solution methods. Therefore, as a first step in studying problem (P), we would like to consider its optimality conditions and then use them to generate algorithms. However, since (P) is nonconvex, it does not seem to possess necessary and sufficient conditions for optimality. Therefore, below we derive several necessary conditions and analyze the relationship between them. We will then show in section 3 how these conditions lead to algorithms that are guaranteed to generate a point satisfying the respective conditions.

For unconstrained differentiable problems, a necessary optimality condition is that the gradient is zero. It is therefore natural to expect that a similar necessary condition will be true over the support $I_1(\mathbf{x}^*)$ of an optimal point $\mathbf{x}^*$. Inspired by linear programming terminology, we will call a vector satisfying this property a *basic feasible* (BF) vector.

DEFINITION 2.1. *A vector* $\mathbf{x}^* \in C_s$ *is a BF vector of* (P) *if*

1. *when* $\|\mathbf{x}^*\|_0 < s$, $\nabla f(\mathbf{x}^*) = 0$;
2. *when* $\|\mathbf{x}^*\|_0 = s$, $\nabla_i f(\mathbf{x}^*) = 0$ *for all* $i \in I_1(\mathbf{x}^*)$.

We will also say that a vector satisfies the "basic feasibility property" if it is a BF vector. Theorem 2.1 establishes the fact that any optimal solution of (P) is also a BF vector.

THEOREM 2.1. *Let* $\mathbf{x}^*$ *be an optimal solution of* (P). *Then* $\mathbf{x}^*$ *is a BF vector.*

Proof. If $\|\mathbf{x}^*\|_0 < s$, then for any $i \in \{1, 2, \ldots, n\}$

$$0 \in \operatorname{argmin}\{g(t) \equiv f(\mathbf{x}^* + t\mathbf{e}_i)\}.$$

Otherwise there would exist a t_0 for which $f(\mathbf{x}^* + t_0\mathbf{e}_i) < f(\mathbf{x}^*)$, which is a contradiction to the optimality of $\mathbf{x}^*$. Therefore, we have $\nabla_i f(\mathbf{x}^*) = g'(0) = 0$. If $\|\mathbf{x}^*\|_0 = s$, then the same argument holds for any $i \in I_1(\mathbf{x}^*)$. □

We conclude that a necessary condition for optimality is basic feasibility. It turns out that this condition is quite weak, namely, there are many BF points that are not optimal points. In the following two subsections we will consider stricter necessary optimality conditions.

Before concluding this section we consider in more detail the special case of $f(\mathbf{x}) \equiv f_{\mathrm{LI}}(\mathbf{x}) \equiv \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2$. We now show that under a suitable condition on $\mathbf{A}$, which we refer to as *s-regularity*, there are only a finite number of BF points. This implies that there are only a finite number of points suspected to be optimal solutions.

DEFINITION 2.2 (s-regularity). *A matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ is called s-regular if for every index set $I \subseteq \{1, 2, \ldots, n\}$ with $|I| = s$, the columns of $\mathbf{A}$ associated with the index set I are linearly independent.*

Remark 2.1. s-regularity can also be expressed in terms of the Kruskal rank of $\mathbf{A}$. The Kruskal rank of a matrix $\mathbf{A}$ is equal to the largest s satisfying the property that every s columns of $\mathbf{A}$ are linearly independent. Another way to express this property is via the spark—spark($\mathbf{A}$) is the minimum number of linearly dependent columns (see [13]). Thus, $\mathbf{A}$ is s-regular if and only if spark($\mathbf{A}$) $\geq$ s + 1.

When $s \leq m$, the s-regularity property is rather mild in the sense that if the components of $\mathbf{A}$ are independently randomly generated from a continuous distribution, then the s-regularity property will be satisfied with probability one.

It is interesting to note that in the compressed sensing literature, it is typically assumed that $\mathbf{A}$ is $2s$-regular. This condition is necessary in order to ensure uniqueness of the solution to $\mathbf{b} = \mathbf{A}\mathbf{x}$ for any $\mathbf{x}$ satisfying $\|\mathbf{x}\|_0 \leq s$. Here we are only requiring s-regularity, which is a milder requirement.

The next lemma shows that when the s-regularity property holds, the number of BF points is finite.

LEMMA 2.1. *Let $f(\mathbf{x}) \equiv f_{\mathrm{LI}}(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2$, where $\mathbf{A} \in \mathbb{R}^{m\times n}$ is an s-regular matrix and $\mathbf{b} \in \mathbb{R}^m$. Then the number of BF points of problem* (P) *is finite.*

Proof. Any BF vector $\mathbf{x}$ satisfies

$$\|\mathbf{x}\|_0 \leq s \text{ and } \nabla_i f_{\mathrm{LI}}(\mathbf{x}) = \mathbf{0}, \quad i \in I_1(\mathbf{x}).$$

Denote the support set of $\mathbf{x}$ by $S = I_1(\mathbf{x})$. Then $|S| \leq s$ and from the derivative condition,

$$\mathbf{A}_S^T(\mathbf{A}_S\mathbf{x}_S - \mathbf{b}) = \mathbf{0},$$

where $\mathbf{A}_S$ is the submatrix of $\mathbf{A}$ made up of the columns corresponding to the set S. Here we used the fact that $\mathbf{A}\mathbf{x} = \mathbf{A}_S\mathbf{x}_S$ for any $\mathbf{x}$ with support S. By the s-regularity assumption it follows that the matrix $\mathbf{A}_S^T\mathbf{A}_S$ is nonsingular. Thus,

$$\mathbf{x}_S = (\mathbf{A}_S^T\mathbf{A}_S)^{-1}\mathbf{A}_S^T\mathbf{b}.$$

To summarize, for each set of indices S satisfying $|S| \leq s$, there is at most one candidate for a BF vector with support S. Since the number of subsets of $\{1, 2, \ldots, n\}$ is finite, the result follows. □

2.3. *L*-stationarity. As we will see in the examples below, the basic feasibility property is a rather weak necessary optimality condition. Therefore, stronger necessary conditions are needed in order to obtain higher quality solutions. In this section we consider the L-stationarity property which is an extension of the concept of stationarity for convex constrained problems. In the next section we discuss CW optimality, which leads to stronger optimality results.

We begin by recalling some well-known elementary concepts on optimality conditions for convex constrained differentiable problems. (For more details see, e.g., [4].) Consider a problem of the form

$$\text{(C)}: \quad \min\{g(\mathbf{x}) : \mathbf{x} \in C\}, \tag{2.1}$$

where C is a closed convex set and g is a continuously differentiable function, which is possibly nonconvex. A vector $\mathbf{x}^* \in C$ is called stationary if

$$\langle \nabla g(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle \geq 0 \text{ for all } \mathbf{x} \in C. \tag{2.2}$$

If $\mathbf{x}^*$ is an optimal solution of (P), then it is also stationary. Therefore, stationarity is a necessary condition for optimality. Many optimization methods devised for solving nonconvex problems of the form (C) are only guaranteed to converge to stationary points. (Occasionally it is only shown that all limit points of the generated sequence are stationary.)

It is often useful to use the property that *for any* $L > 0$, a vector $\mathbf{x}^*$ is a stationary point if and only if

$$\mathbf{x}^* = P_C\left(\mathbf{x}^* - \frac{1}{L}\nabla g(\mathbf{x}^*)\right), \tag{2.3}$$

where for a closed subset $D \subseteq \mathbb{R}^n$ the operator $P_D(\cdot)$ denotes the orthogonal projection onto D, that is,

$$P_D(\mathbf{y}) \equiv \operatorname*{argmin}_{\mathbf{x} \in D} \|\mathbf{x} - \mathbf{y}\|^2.$$

It is interesting to note that condition (2.3)—although expressed in terms of the parameter L—does not actually depend on L by its equivalence to (2.2).

It is natural to try to extend (2.2) or (2.3) to the nonconvex (feasible set) setting. Condition (2.2) with $g = f$ and $C = C_s$ is actually not a necessary optimality condition so we do not pursue it further. To extend (2.3) to the sparsity constrained problem (P), we introduce the notion of L-stationarity.

DEFINITION 2.3. *A vector* $\mathbf{x}^* \in C_s$ *is called an* L*-stationary point of* (P) *if it satisfies the relation*

$$[\text{NC}_L] \quad \mathbf{x}^* \in P_{C_s}\left(\mathbf{x}^* - \frac{1}{L}\nabla f(\mathbf{x}^*)\right). \tag{2.4}$$

Note that since C_s is not a convex set, the orthogonal projection operator $P_{C_s}(\cdot)$ is not single-valued. Specifically, the orthogonal projection $P_{C_s}(\mathbf{x})$ is an n-length vector consisting of the s components of $\mathbf{x}$ with the largest absolute value. In general, there could be more than one choice to the s largest components. For example,

$$P_{C_2}((2,1,1)^T) = \left\{(2,1,0)^T, (2,0,1)^T\right\}.$$

Our results below do not depend on the specific choice.

Remark 2.2. It is interesting to note that a different notion of "stationarity" for problems with nonconvex feasibility sets was discussed in [23]. This work considers optimization problems over locally star-shaped feasible sets, of which the feasible set of (P) is a special case. The definition of stationarity in [23] is different from the one that will be described in what follows. More precisely, for problem (P) the condition reads as follows: $\|\mathbf{x}^*\|_0 \leq s$ and $\mathbf{d} = 0$ is a local minimizer of the problem

$$\min_{\mathbf{d}} \{\nabla f(\mathbf{x}^*)^T \mathbf{d} : \|\mathbf{x}^* + \mathbf{d}\|_0 \leq s\}.$$

A simple argument shows that this condition is in fact equivalent to the notion of basic feasibility.

Below we will show that under an appropriate Lipschitz condition, L-stationarity is a necessary condition for optimality. Before proving this result, we describe a more explicit representation of [NC$_L$].

LEMMA 2.2. *For any $L > 0$, $\mathbf{x}^*$ satisfies* [NC$_L$] *if and only if $\|\mathbf{x}^*\|_0 \leq s$ and*

$$|\nabla_i f(\mathbf{x}^*)| \begin{cases} \leq LM_s(\mathbf{x}^*) & \text{if } i \in I_0(\mathbf{x}^*), \\ = 0 & \text{if } i \in I_1(\mathbf{x}^*). \end{cases} \tag{2.5}$$

Proof. [NC$_L$] $\Rightarrow$ (2.5). Suppose that $\mathbf{x}^*$ satisfies [NC$_L$]. Note that for any index $j \in \{1, 2, \ldots, n\}$, the jth component of $P_{C_s}(\mathbf{x}^* - \frac{1}{L}\nabla f(\mathbf{x}^*))$ is either zero or equal to $x_j^* - \frac{1}{L}\nabla_j f(\mathbf{x}^*)$. Now, since $\mathbf{x}^* \in P_{C_s}(\mathbf{x}^* - \frac{1}{L}\nabla f(\mathbf{x}^*))$, it follows that if $i \in I_1(\mathbf{x}^*)$, then $x_i^* = x_i^* - \frac{1}{L}\nabla_i f(\mathbf{x}^*)$, so that $\nabla_i f(\mathbf{x}^*) = 0$. If $i \in I_0(\mathbf{x}^*)$, then $\left|x_i^* - \frac{1}{L}\nabla_i f(\mathbf{x}^*)\right| \leq M_s(\mathbf{x}^*)$, which combined with the fact that $x_i^* = 0$ implies that $|\nabla_i f(\mathbf{x}^*)| \leq LM_s(\mathbf{x}^*)$, and consequently (2.5) holds true.

(2.5) $\Rightarrow$ [NC$_L$]. Suppose that $\mathbf{x}^*$ satisfies (2.5). If $\|\mathbf{x}^*\|_0 < s$, then $M_s(\mathbf{x}^*) = 0$ and by (2.5) it follows that $\nabla f(\mathbf{x}^*) = 0$; therefore, in this case, $P_{C_s}\left(\mathbf{x}^* - \frac{1}{L}\nabla f(\mathbf{x}^*)\right) = P_{C_s}(\mathbf{x}^*)$ is the set $\{\mathbf{x}^*\}$. If $\|\mathbf{x}^*\|_0 = s$, then $M_s(\mathbf{x}^*) \neq 0$ and $|I_1(\mathbf{x}^*)| = s$. By (2.5)

$$|x_i^* - 1/L\nabla_i f(\mathbf{x}^*)| \begin{cases} = |x_i^*| & i \in I_1(\mathbf{x}^*), \\ \leq M_s(\mathbf{x}^*) & i \in I_0(\mathbf{x}^*). \end{cases}$$

Therefore, the vector $\mathbf{x}^* - \frac{1}{L}\nabla f(\mathbf{x}^*)$ contains the s components of $\mathbf{x}^*$ with the largest absolute value and all other components are smaller or equal to them, so that [NC$_L$] holds. ∎

A direct result of Lemma 2.2 is that any L-stationary point is a BF point.

COROLLARY 2.1. *Suppose that $\mathbf{x}^*$ is an L-stationary point for some $L > 0$. Then $\mathbf{x}^*$ is a BF point.*

Remark 2.3. By Lemma 2.2 it follows that the condition for L-stationarity depends on L. In particular, [NC$_L$] is stronger/more restrictive as L gets smaller. That is, if $\mathbf{x}^*$ is an L_1 stationary point, then it is also an L_2-stationary point for any $L_2 \geq L_1$. This is a different situation than the one described for problems with convex feasible sets where stationarity does not depend on any parameter. Based on this observation, it is natural to define the *stationarity level* of a BF vector $\mathbf{x}^* \in C_s$ as the smallest nonnegative L for which condition (2.5) holds. If a BF vector $\mathbf{x}^*$ satisfies $\|\mathbf{x}^*\|_0 < s$, then the stationarity level is zero. If $\|\mathbf{x}^*\|_0 = s$, then the stationarity level, denoted by $SL(\mathbf{x}^*)$, is given by

$$SL(\mathbf{x}^*) \equiv \max_{i \in I_0(\mathbf{x}^*)} \frac{|\nabla_i f(\mathbf{x}^*)|}{M_s(\mathbf{x}^*)}.$$

The role of stationarity level will become apparent when we discuss the proposed algorithms.

In general, L-stationarity is not a necessary optimality condition for problem (P). To establish such a result, we need to assume a Lipschitz continuity property of ∇f.

Assumption 2. The gradient of the objective function ∇f is Lipschitz with constant $L(f)$ over $\mathbb{R}^n$:

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L(f)\|\mathbf{x} - \mathbf{y}\| \quad \text{for every } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n.$$

This assumption holds for $f = f_{\mathrm{LI}}$ with $L(f) = 2\lambda_{\max}(\mathbf{A}^T\mathbf{A})$ but not for $f = f_{\mathrm{QU}}$. Assumption 2 will *not* be made throughout the paper and it will be stated explicitly when needed.

It is well known that a function satisfying Assumption 2 can be upper bounded by a quadratic function whose associated matrix is a multiple of the identity matrix. This result is known as *the descent lemma*.

LEMMA 2.3 (the descent lemma [4]). *Let f be a continuously differentiable function satisfying Assumption* 2. *Then for every $L \geq L(f)$*

$$f(\mathbf{x}) \leq h_L(\mathbf{x}, \mathbf{y}) \text{ for any } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n,$$

where

$$h_L(\mathbf{x}, \mathbf{y}) \equiv f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2, \quad \mathbf{x}, \mathbf{y} \in \mathbb{R}^n. \tag{2.6}$$

Based on the descent lemma, we can prove the following technical and useful result.

LEMMA 2.4. *Suppose that Assumption* 2 *holds and that $L > L(f)$. Then for any $\mathbf{x} \in C_s$ and $\mathbf{y} \in \mathbb{R}^n$ satisfying*

$$\mathbf{y} \in P_{C_s}\left(\mathbf{x} - \frac{1}{L}\nabla f(\mathbf{x})\right), \tag{2.7}$$

we have

$$f(\mathbf{x}) - f(\mathbf{y}) \geq \frac{L - L(f)}{2}\|\mathbf{x} - \mathbf{y}\|^2. \tag{2.8}$$

Proof. Note that (2.7) can be written as

$$\mathbf{y} \in \operatorname*{argmin}_{\mathbf{z} \in C_s} \left\| \mathbf{z} - \left(\mathbf{x} - \frac{1}{L}\nabla f(\mathbf{x})\right)\right\|^2. \tag{2.9}$$

Since

$$\begin{aligned} h_L(\mathbf{z}, \mathbf{x}) &= f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{z} - \mathbf{x} \rangle + \frac{L}{2}\|\mathbf{z} - \mathbf{x}\|^2 \\ &= \frac{L}{2}\left\| \mathbf{z} - \left(\mathbf{x} - \frac{1}{L}\nabla f(\mathbf{x})\right)\right\|^2 + \underbrace{f(\mathbf{x}) - \frac{1}{2L}\|\nabla f(\mathbf{x})\|^2}_{\text{constant w.r.t. } \mathbf{z}}, \end{aligned}$$

it follows that the minimization problem (2.9) is equivalent to

$$\mathbf{y} \in \operatorname*{argmin}_{\mathbf{z} \in C_s} h_L(\mathbf{z}, \mathbf{x}).$$

This implies that

$$h_L(\mathbf{y}, \mathbf{x}) \leq h_L(\mathbf{x}, \mathbf{x}) = f(\mathbf{x}). \tag{2.10}$$

Now, by the descent lemma we have

$$f(\mathbf{x}) - f(\mathbf{y}) \geq f(\mathbf{x}) - h_{L(f)}(\mathbf{y}, \mathbf{x}),$$

which combined with (2.10) and the identity

$$h_{L(f)}(\mathbf{x}, \mathbf{y}) = h_L(\mathbf{x}, \mathbf{y}) - \frac{L - L(f)}{2}\|\mathbf{x} - \mathbf{y}\|^2$$

yields (2.8). ∎

Under Assumption 2 we now show that an optimal solution of (P) is an L-stationary point for any $L > L(f)$.

THEOREM 2.2. *Suppose that Assumption* 2 *holds,* $L > L(f)$*, and let* $\mathbf{x}^*$ *be an optimal solution of* (P)*. Then*

(i) $\mathbf{x}^*$ *is an* L*-stationary point;*

(ii) *the set* $P_{C_s}\left(\mathbf{x}^* - \frac{1}{L}\nabla f(\mathbf{x}^*)\right)$ *is a singleton.*[1]

Proof. We will prove both parts simultaneously. Suppose to the contrary that there exists a vector

$$\mathbf{y} \in P_{C_s}\left(\mathbf{x}^* - \frac{1}{L}\nabla f(\mathbf{x}^*)\right), \tag{2.11}$$

which is different from $\mathbf{x}^*$ ($\mathbf{y} \neq \mathbf{x}^*$). Invoking Lemma 2.4 with $\mathbf{x} = \mathbf{x}^*$, we have

$$f(\mathbf{x}^*) - f(\mathbf{y}) \geq \frac{L - L(f)}{2}\|\mathbf{x}^* - \mathbf{y}\|^2,$$

contradicting the optimality of $\mathbf{x}^*$. We conclude that $\mathbf{x}^*$ is the only vector in the set $P_{C_s}\left(\mathbf{x}^* - \frac{1}{L}\nabla f(\mathbf{x}^*)\right)$. ∎

To summarize this section, we have shown that under a Lipschitz condition on ∇f, L-stationarity for any $L > L(f)$ is a necessary optimality condition, which also implies the basic feasibility property. In section 3.1 we will show how the IHT method for solving the general problem (P) can be used in order to find L-stationary points (for $L > L(f)$).

2.4. CW-minima. The L-stationarity necessary optimality condition has two major drawbacks: first, it requires the function's gradient to be Lipschitz continuous, and second, in order to validate it, we need to know a bound on the Lipschitz constant. We now consider a different and stronger necessary optimality condition that does not require such knowledge on the Lipschitz constant, and in fact does not even require Assumption 2 to hold.

For a general unconstrained optimization problem, a vector $\mathbf{x}^*$ is a CW-minimum if for every $i = 1, 2, \ldots, n$ the scalar x_i^* is a minimum of f with respect to the ith component x_i while keeping all other variables fixed:

$$x_i^* \in \underset{x_i}{\operatorname{argmin}} f(x_1^*, \ldots, x_{i-1}^*, x_i, x_{i+1}^*, \ldots, x_n^*).$$

Clearly, any optimal $\mathbf{x}^*$ is also a CW-minimum. It is therefore natural to extend this definition to problem (P) in order to obtain an alternative necessary condition.

DEFINITION 2.4. *Let* $\mathbf{x}^*$ *be a feasible solution of* (P)*. Then* $\mathbf{x}^*$ *is a CW-minimum of* (P) *if one of the following cases holds true:*

[1] A set is called a *singleton* if it contains exactly one element.

Case I. $\|\mathbf{x}^*\|_0 < s$ *and for every* $i = 1, 2, \ldots, n$ *one has*

$$f(\mathbf{x}^*) = \min_{t\in\mathbb{R}} f(\mathbf{x}^* + t\mathbf{e}_i). \tag{2.12}$$

Case II. $\|\mathbf{x}^*\|_0 = s$ *and for every* $i \in I_1(\mathbf{x}^*)$ *and* $j = 1, 2, \ldots, n$ *one has*

$$f(\mathbf{x}^*) \le \min_{t\in\mathbb{R}} f(\mathbf{x}^* - x_i^*\mathbf{e}_i + t\mathbf{e}_j). \tag{2.13}$$

Obviously, any optimal solution of (P) is also a CW-minimum. This is formally stated in the next theorem.

THEOREM 2.3. *Let* $\mathbf{x}^*$ *be an optimal solution of* (P). *Then* $\mathbf{x}^*$ *is a CW-minimum of* (P). Of course, a CW-minimum is not necessarily an optimal solution and in fact can be far from the global optimum.

It is easy to see that any CW-minimum is also a BF vector, as stated in the following lemma.

LEMMA 2.5. *Let* $\mathbf{x}^* \in C_s$ *be a CW-minimum of* (P). *Then* $\mathbf{x}^*$ *is also a BF vector.*

Proof. We first show that if a vector $\mathbf{x}^*$ satisfying $\|\mathbf{x}^*\|_0 = s$ is a CW-minimum of (P), then (2.12) is satisfied for any $i \in I_1(\mathbf{x}^*)$. Indeed, inequality (2.13) with $i \in I_1(\mathbf{x}^*)$ and $j = i$ becomes

$$f(\mathbf{x}^*) \le \min_{t\in\mathbb{R}} f(\mathbf{x}^* - x_i^*\mathbf{e}_i + t\mathbf{e}_i). \tag{2.14}$$

Since $f(\mathbf{x}^* - x_i^*\mathbf{e}_i + x_i^*\mathbf{e}_i) = f(\mathbf{x}^*)$, it follows that (2.14) is equivalent to

$$f(\mathbf{x}^*) = \min_{t\in\mathbb{R}} f(\mathbf{x}^* - x_i^*\mathbf{e}_i + t\mathbf{e}_i),$$

which letting $s = t - x_i^*$ becomes

$$f(\mathbf{x}^*) = \min_{s\in\mathbb{R}} f(\mathbf{x}^* + s\mathbf{e}_i).$$

We conclude that for any CW-minimum $\mathbf{x}^*$ of (P) we have

$$\nabla_i f(\mathbf{x}^*) = 0 \text{ for all } i \in I_1(\mathbf{x}^*). \tag{2.15}$$

In addition, in Case I of Definition 2.4, we obviously have that $\nabla f(\mathbf{x}^*) = 0$, which completes the proof. ∎

We have previously established under Assumption 2 in Theorem 2.2 that being an L-stationary point for $L > L(f)$ is a necessary condition for optimality. A natural question that arises is what is the relation between CW-minima and L-stationary points (for $L > L(f)$). We will show that being a CW-minimum is a stronger, i.e., more restrictive, condition than being an L-stationary point for any $L \ge L(f)$. In fact, a stronger result will be established: any CW-minimum is also an $\tilde{L}$-stationary point for an $\tilde{L}$ which is less than or equal to $L(f)$. In practice, $\tilde{L}$ can be much smaller than $L(f)$.

In order to precisely define $\tilde{L}$, we note that under Assumption 2, it follows immediately that for any $i \ne j$ there exists a constant $L_{i,j}(f)$ for which

$$\|\nabla_{i,j} f(\mathbf{x}) - \nabla_{i,j} f(\mathbf{x} + \mathbf{d})\| \le L_{i,j}(f)\|\mathbf{d}\| \tag{2.16}$$

for any $\mathbf{x} \in \mathbb{R}^n$ and any $\mathbf{d} \in \mathbb{R}^n$ which has at most two nonzero components. Here $\nabla_{i,j} f(\mathbf{x})$ denotes a vector of length 2 whose elements are the ith and jth elements of

$\nabla f(\mathbf{x})$. We will be especially interested in the following constant, which we call *the local Lipschitz constant*:

$$L_2(f) \equiv \max_{i \neq j} L_{i,j}(f).$$

Clearly (2.16) is satisfied when replacing $L_{i,j}(f)$ by $L(f)$. Therefore, in general,

$$L_2(f) \leq L(f).$$

In practice, $L_2(f)$ can be much smaller than $L(f)$, as the following example illustrates.

Example 2.1. Suppose that the objective function in (P) is $f(\mathbf{x}) = \mathbf{x}^T\mathbf{Q}\mathbf{x} + 2\mathbf{b}^T\mathbf{x}$, with $\mathbf{b}$ being a vector in $\mathbb{R}^n$ and

$$\mathbf{Q} = \mathbf{I}_n + \mathbf{J}_n,$$

where $\mathbf{I}_n$ is the $n \times n$ identity matrix and $\mathbf{J}_n$ is the $n \times n$ matrix of all ones. Then

$$L(f) = 2\lambda_{\max}(\mathbf{Q}) = 2\lambda_{\max}(\mathbf{I}_n + \mathbf{J}_n) = 2(n+1).$$

On the other hand, for any $i \neq j$ the constant $L_{i,j}(f)$ is twice the maximum eigenvalue of the submatrix of $\mathbf{Q}$ consisting of the ith and jth rows and columns. That is,

$$L_{i,j}(f) = 2\lambda_{\max}\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} = 6.$$

For large n, $L(f) = 2n + 2$ can be much larger than $L_2(f) = 6$. It is not difficult to see that the descent lemma (Lemma 2.3) can be refined to a suitable "local" version.

LEMMA 2.6 (local descent lemma). *Suppose that Assumption* 2 *holds. Then*

$$f(\mathbf{x} + \mathbf{d}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^T\mathbf{d} + \frac{L_2(f)}{2}\|\mathbf{d}\|^2$$

for any vector $\mathbf{d} \in \mathbb{R}^n$ *with at most two nonzero components.*

Using the local descent lemma we can now show that a CW-minimum is also an $L_2(f)$-stationary point.

THEOREM 2.4. *Suppose that Assumption* 2 *holds and let* $\mathbf{x}^*$ *be a CW-minimum of* (P). *Then*

$$|\nabla_i f(\mathbf{x}^*)| \begin{cases} \leq L_2(f)M_s(\mathbf{x}^*), & i \in I_0(\mathbf{x}^*), \\ = 0, & i \in I_1(\mathbf{x}^*), \end{cases} \tag{2.17}$$

that is, $\mathbf{x}^*$ *is an* $L_2(f)$*-stationary point.*

Proof. Since $\mathbf{x}^*$ is a CW-minimum, it follows by Lemma 2.5 that it is a BF vector. Thus, if $\|\mathbf{x}^*\|_0 < s$, we have $\nabla f(\mathbf{x}^*) = \mathbf{0}$, establishing the result for this case.

Suppose now that $\|\mathbf{x}^*\|_0 = s$. Let $i \in I_1(\mathbf{x}^*)$. Then again by Lemma 2.5 it follows that $\mathbf{x}^*$ is a BF vector and thus $\nabla_i f(\mathbf{x}^*) = 0$. Now let $i \in I_0(\mathbf{x}^*)$ and let m be an index for which $|x_m^*| = M_s(\mathbf{x}^*)$. Obviously, $m \in I_1(\mathbf{x}^*)$, and thus, since $\mathbf{x}^*$ is a CW-minimum, it follows in particular that

$$f(\mathbf{x}^*) \leq f(\mathbf{x}^* - x_m^*\mathbf{e}_m - \sigma x_m^*\mathbf{e}_i), \tag{2.18}$$

where $\sigma = \operatorname{sgn}(x_m^* \nabla_i f(\mathbf{x}^*))$. By the local descent lemma (Lemma 2.6) we have

$$
\begin{aligned}
(2.19) \qquad & f(\mathbf{x}^* - x_m^* \mathbf{e}_m - \sigma x_m^* \mathbf{e}_i) \\
& \leq f(\mathbf{x}^*) + \nabla f(\mathbf{x}^*)^T(-x_m^* \mathbf{e}_m - \sigma x_m^* \mathbf{e}_i) + \frac{L_2(f)}{2}\|x_m^* \mathbf{e}_m + \sigma x_m^* \mathbf{e}_i\|^2 \\
& = f(\mathbf{x}^*) - x_m^* \nabla_m f(\mathbf{x}^*) - \sigma x_m^* \nabla_i f(\mathbf{x}^*) + L_2(f)(x_m^*)^2 \\
& = f(\mathbf{x}^*) - \sigma x_m^* \nabla_i f(\mathbf{x}^*) + L_2(f)(x_m^*)^2,
\end{aligned}
$$

where the last equality follows from the fact that since $m \in I_1(\mathbf{x}^*)$, it follows by (2.15) that $\nabla_m f(\mathbf{x}^*) = 0$.

Combining (2.18) and (2.19) we obtain that

$$0 \leq -\sigma x_m^* \nabla_i f(\mathbf{x}^*) + L_2(f)(x_m^*)^2.$$

Recalling the definition of σ, we conclude that

$$|x_m^* \nabla_i f(\mathbf{x}^*)| \leq L_2(f)(x_m^*)^2,$$

which is equivalent to

$$|\nabla_i f(\mathbf{x}^*)| \leq L_2(f)|x_m^*| = L_2(f) M_s(\mathbf{x}^*),$$

concluding the proof. □

An immediate consequence of Theorem 2.4 is that under Assumption 2, any optimal solution of (P) is an $L_2(f)$-stationary point.

COROLLARY 2.2. *Suppose that Assumption* 2 *holds. Then any optimal solution of* (P) *is also an* $L_2(f)$*-stationary point of* (P).

To summarize our discussion on optimality conditions we have shown that without Assumption 2 we have the following relations:

$$
\begin{array}{rc}
 & \text{optimal solution of (P)} \\
\text{Theorem 2.3} & \Downarrow \\
 & \text{CW-minimum of (P)} \\
\text{Lemma 2.5} & \Downarrow \\
 & \text{BF vector of (P)}
\end{array}
$$

Under Assumption 2, we have

$$
\begin{array}{rc}
 & \text{optimal solution of (P)} \\
\text{Theorem 2.3} & \Downarrow \\
 & \text{CW-minimum of (P)} \\
\text{Theorem 2.4} & \Downarrow \\
 & L_2(f) - \text{stationary} \\
\text{Corrolary 2.1} & \Downarrow \\
 & \text{BF vector of (P)}
\end{array}
$$

To illustrate these relationships we consider a detailed example.

Example 2.2. Consider problem (P) with $s = 2, n = 5$, and

$$f(\mathbf{x}) = \mathbf{x}^T \mathbf{Q} \mathbf{x} + 2\mathbf{b}^T \mathbf{x},$$

where $\mathbf{Q} = \mathbf{I}_5 + \mathbf{J}_5$ as in Example 2.1 and $\mathbf{b} = -(3, 2, 3, 12, 5)^T$. In Lemma 2.1 we showed how to compute the BF vectors of problem (P) with a quadratic objective.

TABLE 2.1
Function values and stationarity levels of the 10 *BF vectors.*

BF vector number	1	2	3	4	5	6	7	8	9	10
Function value	-4.66	-6.00	-78	-12.66	-4.66	-82.66	-12.66	-78	-12.66	-72.66
Stationarity level	62	20	3	56	62	1.25	58	3	56	11

Using this method it is easy to see that in our case there are 10 BF vectors given by (each corresponding to a different choice of two variables out of 5):

$$\begin{aligned}
\mathbf{x}_1 &= (1.3333, 0.3333, 0, 0, 0)^T, \\
\mathbf{x}_2 &= (1.0000, 0, 1.0000, 0, 0)^T, \\
\mathbf{x}_3 &= (-2.0000, 0, 0, 7.0000, 0)^T, \\
\mathbf{x}_4 &= (0.3333, 0, 0, 0, 2.3333)^T, \\
\mathbf{x}_5 &= (0, 0.3333, 1.3333, 0, 0)^T, \\
\mathbf{x}_6 &= (0, -2.6667, 0, 7.3333, 0)^T, \\
\mathbf{x}_7 &= (0, -0.3333, 0, 0, 2.6667)^T, \\
\mathbf{x}_8 &= (0, 0, -2.0000, 7.0000, 0)^T, \\
\mathbf{x}_9 &= (0, 0, 0.3333, 0, 2.3333)^T, \\
\mathbf{x}_{10} &= (0, 0, 0, 6.3333, -0.6667)^T.
\end{aligned}$$

The stationarity levels (see Remark 2.3) and function values up to two digits of accuracy of each of the BF vectors is given in Table 2.1.

Since in this case $L_2(f) = 6$ (see Example 2.1), it follows by Corollary 2.2 that any optimal solution is a 6-stationary point, implying that only the three BF vectors $\mathbf{x}_3, \mathbf{x}_6, \mathbf{x}_8$ are candidates for being optimal solutions. In addition, by Theorem 2.4, only these three BF vectors may be CW-minima. By direct calculation we found that only $\mathbf{x}_6$—the optimal solution of the problem—is a CW-minimum. Therefore, in this case, the only CW-minimum is the global optimal solution. Note, however, that there could of course be examples in which there exist CW-minima which are not optimal.

3. Numerical algorithms. We now develop two classes of algorithms that achieve the necessary conditions defined in the previous section:

- IHT. The first algorithm results from using the L-stationary condition. For the case $f \equiv f_{\mathrm{LI}}$, and under the assumption that $\|\mathbf{A}\|_2 < 1$, it coincides with the IHT method [6]. Our approach extends this algorithm to the general case under Assumption 2, and it will be referred to as the IHT method in our general setting as well. We note that a generalization of IHT to the nonlinear case can also be found in [5]. We will prove that the limit points of the algorithm are $L(f)$-stationary points. As we show, this method is well defined only when Assumption 2 holds and relies on knowledge of the Lipschitz constant.
- Sparse-simplex methods. The other two algorithms we suggest are essentially coordinate descent methods that optimize the objective function at each iteration with respect to either one or two decision variables.
  The first algorithm in this class seeks the coordinate or coordinates that lead to the largest decrease and optimizes with respect to them. Since the support of the iterates changes by at most *one* index, it has some resemblance to the celebrated simplex method for linear programming and will thus be

referred to as the greedy sparse-simplex method. We show that any limit point of the sequence generated by this approach is a CW-minimum, which as shown in Theorem 2.4 is a stronger notion than L-stationarity for any $L \geq L_2(f)$. An additional advantage of this approach is that it is well defined even when Assumption 2 is not valid, and it does not require any knowledge of the Lipschitz constant even when one exists. The disadvantage of the greedy sparse-simplex method is that it does not have a selection strategy for choosing the indices of the variables to be optimized but rather explores all possible choices. Depending on the objective, this may be a very costly step. To overcome this drawback, we suggest a second coordinate descent algorithm with an extremely simple index selection rule; this rule discards the need to perform an exhaustive search for the relevant indices on which the optimization will be performed. This approach will be referred to as the partial sparse-simplex method. Under Assumption 2 we show that it is guaranteed to converge to $L_2(f)$-stationary points.

In the ensuing subsections we consider each of the algorithms above.

3.1. The IHT method. One approach for solving problem (P) is to employ the following fixed point method in order to enforce the L-stationary condition (2.4):

$$\mathbf{x}^{k+1} \in P_{C_s}\left(\mathbf{x}^k - \frac{1}{L}\nabla f(\mathbf{x}^k)\right), \quad k = 0, 1, 2, \ldots. \tag{3.1}$$

Convergence results on this method can be obtained when Assumption 2 holds; we will therefore make this assumption throughout this subsection. The iterations defined by (3.1) were studied in [6] for the special case in which $f \equiv f_{\mathrm{LI}}$ and $\|\mathbf{A}\|_2 < 1$ and were referred to as the M-sparse algorithm. Later on, in [7], the authors referred to this approach as the IHT method (again, for $f = f_{\mathrm{LI}}$) and analyzed a version with an adaptive stepsize which avoids the need for the normalization property $\|\mathbf{A}\|_2 < 1$. Similarly, we refer to this approach for more general objective functions as the IHT method:

The IHT method
Input: a constant $L > L(f)$.
• **Initialization:** Choose $\mathbf{x}_0 \in C_s$.
• **General step :** $\mathbf{x}^{k+1} \in P_{C_s}\left(\mathbf{x}^k - \frac{1}{L}\nabla f(\mathbf{x}^k)\right), \quad (k = 0, 1, 2, \ldots)$

Remark 3.1. We adopt the convention in the literature that does not specify the strategy for choosing a vector from the set $D_k \equiv P_{C_s}\left(\mathbf{x}^k - \frac{1}{L}\nabla f(\mathbf{x}^k)\right)$. The strategy of choosing an element in D_k is a matter of implementation. For example, one can implement the method so that in case of ties the smallest indices are chosen. The convergence theorems are independent of the specific strategy for choosing the vector from the set D_k and are valid for any sequence satisfying the inclusion relation in the general step of the IHT method. It can be shown that the general step of the IHT method is equivalent to the relation

$$\mathbf{x}^{k+1} \in \operatorname*{argmin}_{\mathbf{x} \in C_s} h_L(\mathbf{x}, \mathbf{x}^k), \tag{3.2}$$

where $h_L(\mathbf{x}, \mathbf{y})$ is defined by (2.6). (See also the proof of Theorem 2.2.)

Several basic properties of the IHT method are summarized in the following lemma.

LEMMA 3.1. *Let $\{\mathbf{x}^k\}_{k\geq 0}$ be the sequence generated by the IHT method with a constant stepsize $\frac{1}{L}$ where $L > L(f)$. Then*

1. $f(\mathbf{x}^k) - f(\mathbf{x}^{k+1}) \geq \frac{L-L(f)}{2}\|\mathbf{x}^k - \mathbf{x}^{k+1}\|^2$;
2. $\{f(\mathbf{x}^k)\}_{k\geq 0}$ *is a nonincreasing sequence;*
3. $\|\mathbf{x}^k - \mathbf{x}^{k+1}\| \to 0$;
4. *for every $k = 0, 1, 2, \ldots$, if $\mathbf{x}^k \neq \mathbf{x}^{k+1}$, then $f(\mathbf{x}^{k+1}) < f(\mathbf{x}^k)$.*

Proof. Part 1 follows by substituting $\mathbf{x} = \mathbf{x}^k, \mathbf{y} = \mathbf{x}^{k+1}$ in (2.8). Parts 2, 3, and 4 follow immediately from part 1. ∎

A direct consequence of Lemma 3.1 is the convergence of the sequence of function values.

COROLLARY 3.1. *Let $\{\mathbf{x}^k\}_{k\geq 0}$ be the sequence generated by the IHT method with a constant stepsize $\frac{1}{L}$ where $L > L(f)$. Then the sequence $\{f(\mathbf{x}^k)\}_{k\geq 0}$ converges.*

As we have seen, the IHT algorithm can be viewed as a fixed point method for solving the condition for L-stationarity. The following theorem states that all accumulation points of the sequence generated by the IHT method with constant stepsize $\frac{1}{L}$ are indeed L-stationary points.

THEOREM 3.1. *Let $\{\mathbf{x}^k\}_{k\geq 0}$ be the sequence generated by the IHT method with stepsize $\frac{1}{L}$ where $L > L(f)$. Then any accumulation point of $\{\mathbf{x}^k\}_{k\geq 0}$ is an L-stationary point.*

Proof. Suppose that $\mathbf{x}^*$ is an accumulation point of the sequence. Then there exists a subsequence $\{\mathbf{x}^{k_n}\}_{n\geq 0}$ that converges to $\mathbf{x}^*$. By Lemma 3.1

$$f(\mathbf{x}^{k_n}) - f(\mathbf{x}^{k_n+1}) \geq \frac{L - L(f)}{2}\|\mathbf{x}^{k_n} - \mathbf{x}^{k_n+1}\|^2. \tag{3.3}$$

Since $\{f(\mathbf{x}^{k_n})\}_{n\geq 0}$ and $\{f(\mathbf{x}^{k_n+1})\}_{n\geq 0}$ both converge to the same limit f^*, it follows that $f(\mathbf{x}^{k_n}) - f(\mathbf{x}^{k_n+1}) \to 0$ as $n \to \infty$, which combined with (3.3) yields that

$$\mathbf{x}^{k_n+1} \to \mathbf{x}^* \text{ as } n \to \infty.$$

Recall that for all $n \geq 0$

$$x^{k_n+1} \in P_{C_s}\left(\mathbf{x}^{k_n} - \frac{1}{L}\nabla f(\mathbf{x}^{k_n})\right).$$

Let $i \in I_1(\mathbf{x}^*)$. By the convergence of $\mathbf{x}^{k_n}$ and $\mathbf{x}^{k_n+1}$ to $\mathbf{x}^*$, it follows that there exists N such that

$$x_i^{k_n}, x_i^{k_n+1} \neq 0 \text{ for all } n > N,$$

and therefore, for $n > N$,

$$x_i^{k_n+1} = x_i^{k_n} - \frac{1}{L}\nabla_i f(\mathbf{x}^{k_n}).$$

Taking n to ∞ we obtain that

$$\nabla_i f(\mathbf{x}^*) = 0.$$

Now let $i \in I_0(\mathbf{x}^*)$. If there exists an infinite number of indices k_n for which $x_i^{k_n+1} \neq 0$, then as in the previous case we obtain that $x_i^{k_n+1} = x_i^{k_n} - \frac{1}{L}\nabla_i f(\mathbf{x}^{k_n})$ for these indices, implying (by taking the limit) that $\nabla_i f(\mathbf{x}^*) = 0$. In particular,

$|\nabla_i f(\mathbf{x}^*)| \leq LM_s(\mathbf{x}^*)$. On the other hand, if there exists an $M > 0$ such that for all $n > M$ $x_i^{k_n+1} = 0$, then

$$\left| x_i^{k_n} - \frac{1}{L}\nabla_i f(\mathbf{x}^{k_n}) \right| \leq M_s\left(\mathbf{x}^{k_n} - \frac{1}{L}\nabla f(\mathbf{x}^{k_n})\right) = M_s(\mathbf{x}^{k_n+1}).$$

Thus, taking n to infinity while exploiting the continuity of the function M_s, we obtain that

$$|\nabla_i f(\mathbf{x}^*)| \leq LM_s\left(\mathbf{x}^*\right),$$

establishing the desired result. □

3.1.1. The Case $f = f_{\mathrm{LI}}$. When $f(\mathbf{x}) \equiv f_{\mathrm{LI}}(\mathbf{x}) \equiv \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2$, and under the assumption of s-regularity of (P), we know by Lemma 2.1 that the number of BF vectors is finite. Utilizing this fact we can now show convergence of the whole sequence generated by the IHT method when $f = f_{\mathrm{LI}}$. This result is stronger than the one of Theorem 3.1, which only shows that all accumulation points are L-stationary points.

THEOREM 3.2. *Let $f(\mathbf{x}) \equiv f_{\mathrm{LI}}(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2$. Suppose that the s-regularity property holds for the matrix $\mathbf{A}$. Then the sequence generated by the IHT method with stepsize $\frac{1}{L}$ where $L > L(f)$ converges to an L-stationary point.*

Proof. Let $\{\mathbf{x}^k\}_{k\geq 0}$ be the sequence generated by the IHT method. We begin by showing that the sequence is bounded. By the descent property of the sequence of function values (see Lemma 3.1), it follows that the sequence $\{\mathbf{x}^k\}_{k\geq 0}$ is contained in the level set

$$T = \{\mathbf{x} \in \mathbb{R}^n : f_{\mathrm{LI}}(\mathbf{x}) \leq f_{\mathrm{LI}}(\mathbf{x}^0)\}.$$

We now show that T is bounded. To this end, note that the number of subsets of $\{1, 2, \ldots, n\}$ whose size is no larger than s is equal to

$$p = \sum_{k=0}^{s} \binom{n}{k}.$$

By denoting these p subsets as $I_1, I_2, \ldots, I_p$, we can represent the set T as the union

$$T = \bigcup_{j=1}^{p} T_j,$$

where

$$T_j = \left\{\mathbf{x} \in \mathbb{R}^n : f_{\mathrm{LI}}(\mathbf{x}) \leq f_{\mathrm{LI}}(\mathbf{x}^0), x_i = 0 \quad \text{for all } i \notin I_j\right\}.$$

In this notation, we can rewrite T_j as

$$T_j = \left\{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{A}_{T_j}\mathbf{x}_{T_j} - \mathbf{b}\|^2 \leq f_{\mathrm{LI}}(\mathbf{x}^0), \mathbf{x}_{\overline{T_j}} = \mathbf{0}\right\}.$$

The set T_j is bounded since the s-regularity of $\mathbf{A}$ implies that the matrix $\mathbf{A}_{T_j}^T\mathbf{A}_{T_j}$ is positive definite. This implies the boundedness of T.

We conclude that the sequence $\{\mathbf{x}^k\}_{k\geq 0}$ is bounded and therefore, in particular, there exists a subsequence $\{\mathbf{x}^{k_n}\}_{n\geq 0}$ which converges to an accumulation point $\mathbf{x}^*$

which is an L-stationary point and hence also a BF vector. By Lemma 2.1, the number of BF vectors is finite, which implies that there exists an $\varepsilon > 0$ smaller than the minimal distance between all the pairs of the BF vectors. To show the convergence of the entire sequence to $\mathbf{x}^*$, suppose in contradiction that this is not the case. We will assume without loss of generality that the subsequence $\{\mathbf{x}^{k_n}\}_{n\geq 0}$ satisfies $\|\mathbf{x}^{k_n} - \mathbf{x}^*\| \leq \varepsilon$ for every $n \geq 0$. Since we assumed that the sequence is not convergent, the index t_n given by

$$t_n = \max\{l : \|\mathbf{x}^i - \mathbf{x}^*\| \leq \varepsilon, i = k_n, k_n + 1, \ldots, l\}$$

is well defined. We have thus constructed a subsequence $\{\mathbf{x}^{t_n}\}_{n\geq 0}$ for which

$$\|\mathbf{x}^{t_n} - \mathbf{x}^*\| \leq \varepsilon, \|\mathbf{x}^{t_n+1} - \mathbf{x}^*\| > \varepsilon, \quad n = 0, 1, \ldots.$$

It follows that $\mathbf{x}^{t_n}$ converges to $\mathbf{x}^*$, and in particular there exists an $N > 0$ such that for all $n > N$, $\|\mathbf{x}^{t_n} - \mathbf{x}^*\| \leq \varepsilon/2$. Thus, for all $n > N$,

$$\|\mathbf{x}^{t_n} - \mathbf{x}^{t_n+1}\| > \frac{\varepsilon}{2},$$

contradicting part 3 of Lemma 3.1. ∎

Remark 3.2. As we noted previously, the IHT method in the case $f = f_{\mathrm{LI}}$ with fixed stepsize set to 1 was proposed in [6]. It was shown in [6] that if $\mathbf{A}$ satisfies the s-regularity property and $\|\mathbf{A}\|_2 < 1$, then the algorithm converges to a local minimum. This result is consistent with Theorem 3.2 since when $\|\mathbf{A}\|_2 < 1$, the Lipschitz constant satisfies $L(f) < 1$, and we can therefore ensure convergence by Theorem 3.2 with stepsize equal to 1. In [7] the authors note that the IHT method with stepsize 1 might diverge when $\|\mathbf{A}\|_2 > 1$. To overcome this limitation, they propose an adaptive stepsize for which they show the same type of convergence results. Our result here shows that a fixed stepsize which depends on the Lipschitz constant can also be used.

3.1.2. Examples.

Example 3.1. Consider the problem

$$\min\left\{f(x_1, x_2) \equiv 12x_1^2 + 20x_1x_2 + 16x_2^2 + 2x_1 + 18x_2 : \left\|(x_1; x_2)^T\right\|_0 \leq 1\right\}. \tag{3.4}$$

The objective function is convex quadratic and the Lipschitz constant of its gradient is given by

$$L(f) = 2\lambda_{\max}\begin{pmatrix} 12 & 10 \\ 10 & 16 \end{pmatrix} = 48.3961.$$

It can be easily seen that there are only two BF vectors to this problem: $(0, -9/16)^T$, $(-1/12, 0)^T$ (constructed by taking one variable to be zero and the other to satisfy that the corresponding partial derivative is zero). The optimal solution of the problem is the first BF vector $(0, -9/16)^T$ with objective function value of $-81/16$. This point is an L-stationary point for any $L \geq L(f)$. The second point $(-1/12, 0)^T$ is not an optimal solution (its objective function value is $-1/12$). Since $\nabla_2 f((-1/12, 0)^T) = 49/3$, it follows by Lemma 2.2 that it is an L-stationary point for $L \geq \frac{49/3}{1/12} = 196$. Therefore, for any $L \in [L(f), 196)$, only the optimal solution $(0, -9/16)^T$ is an L-stationary point and the IHT method is guaranteed to converge to the global

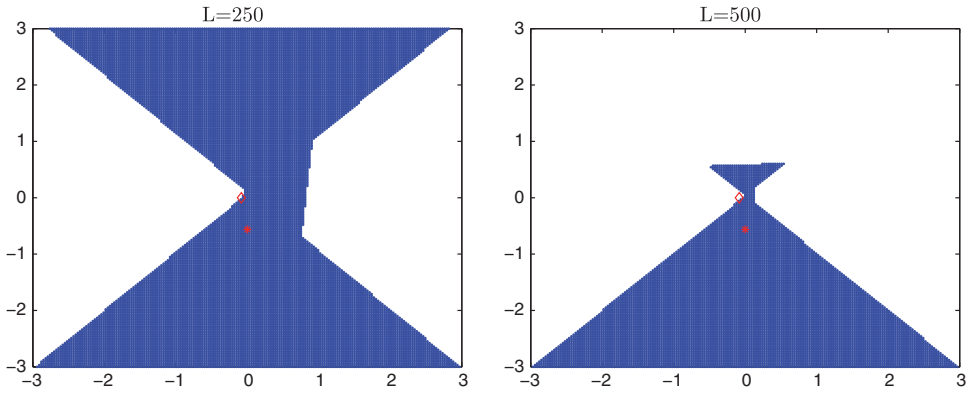

FIG. 3.1. *The optimal solution* $(0, -9/16)^T$ *is denoted by a red asterisk and the additional BF vector* $(-1/12, 0)^T$ *is denoted by a red diamond. The region of convergence to the optimal solution is the blue region and the points in the white region converged to the nonoptimal point* $(-1/12, 0)^T$*. The left image describes the convergence region when the IHT method was invoked with* $L = 250$*, while the right image describes the same for* $L = 500$*. When* L *gets larger, the chances to converge to the nonoptimal* L*-stationary point are higher.*

optimal solution. However, if the upper bound is chosen to satisfy $L \geq 196$, then $(-1/12, 0)^T$ is also an L-stationary point and the IHT method might converge to it. This is illustrated in Figure 3.1.

Example 3.1 illustrates the fact that although convergence of the sequence is always guaranteed by Theorem 3.2, the likelihood that the convergence will be to the global optimum decreases as L gets larger (i.e., the stepsize gets smaller).

Example 3.2. For any two positive number $a < b$, consider the problem

$$\min\{f(x_1, x_2) \equiv a(x_1 - 1)^2 + b(x_2 - 1)^2 : \|(x_1, x_2)^T\|_0 \leq 1\}.$$

Obviously the optimal solution of the problem is $(x_1, x_2) = (0, 1)$. An additional BF vector is $\tilde{\mathbf{x}} = (1, 0)^T$. Note that here $L(f) = 2b$. Therefore, since $\nabla f(\tilde{\mathbf{x}}) = (0, -2b)^T$ and $M_1(\tilde{\mathbf{x}}) = 1$, it follows that

$$|\nabla_2 f(\tilde{\mathbf{x}})| \leq L(f) M_1(\tilde{\mathbf{x}}),$$

and hence $\tilde{\mathbf{x}}$ will also be an L-stationary point *for any* $L \geq L(f)$. Therefore, in this problem, regardless of the value of L, there is always a chance to converge to a nonoptimal solution.

3.2. The greedy sparse-simplex method. The IHT algorithm is able to find L-stationary points for any $L > L(f)$ under Assumption 2. However, by Corollary 2.2, any optimal solution is also an $L_2(f)$-stationary point, and $L_2(f)$ can be significantly smaller than $L(f)$. It is therefore natural to seek a method that is able to generate such points. An even better approach would be to derive an algorithm that converges to a CW-minimum, which by Theorem 2.4 is a stronger notion than L-stationarity. An additional drawback of IHT is that it requires the validity of Assumption 2 and the knowledge of the Lipschitz constant $L(f)$.

Below we present *the greedy sparse-simplex method* which overcomes the faults of IHT alluded to above: its limit points are CW-minima and it does not require the validity of Assumption 2, but if the assumption does hold, then its limit points

are $L_2(f)$-stationary points (without the need to know any information on Lipschitz constants).

The Greedy Sparse-Simplex Method

- **Initialization:** Choose $\mathbf{x}_0 \in C_s$.
- **General step :** $(k = 0, 1, \ldots)$
  - If $\|\mathbf{x}^k\|_0 < s$, then compute for every $i = 1, 2, \ldots, n$

$$t_i \in \operatorname*{argmin}_{t\in\mathbb{R}} f(\mathbf{x}^k + t\mathbf{e}_i), \tag{3.5}$$
$$f_i = f(\mathbf{x}^k + t_i\mathbf{e}_i).$$

  Let $i_k \in \operatorname*{argmin}_{i=1,\ldots,n} f_i$. If $f_{i_k} < f(\mathbf{x}^k)$, then set

$$\mathbf{x}^{k+1} = \mathbf{x}^k + t_{i_k}\mathbf{e}_{i_k}.$$

  Otherwise, STOP.
  - If $\|\mathbf{x}^k\|_0 = s$, then for every $i \in I_1(\mathbf{x}^k)$ and $j = 1, \ldots, n$ compute

$$t_{i,j} \in \operatorname*{argmin}_{t\in\mathbb{R}} f(\mathbf{x}^k - x_i^k\mathbf{e}_i + t\mathbf{e}_j), \tag{3.6}$$
$$f_{i,j} = f(\mathbf{x}^k - x_i^k\mathbf{e}_i + t_{i,j}\mathbf{e}_j).$$

  Let $(i_k, j_k) \in \operatorname{argmin}\{f_{i,j} : i \in I_1(\mathbf{x}^k), j = 1, \ldots, n\}$. If $f_{i_k,j_k} < f(\mathbf{x}^k)$, then set

$$\mathbf{x}^{k+1} = \mathbf{x}^k - x_{i_k}^k\mathbf{e}_{i_k} + t_{i_k,j_k}\mathbf{e}_{j_k}.$$

  Otherwise, STOP.

Remark 3.3. One advantage of the greedy sparse-simplex method is that it can be easily implemented for the case $f \equiv f_{\mathrm{QU}}$, that is, the case when the objective function is quartic. In this case the minimization steps (3.5) and (3.6) consist of finding the minimum of a scalar quartic (though nonconvex) function, which is an easy task since the minimizer is one of at most three roots of the cubic polynomial derivative.

By its definition, the greedy sparse-simplex method generates a nonincreasing sequence of function values and gets stuck only at CW-minima.

LEMMA 3.2. *Let $\{\mathbf{x}^k\}$ be the sequence generated by the greedy sparse-simplex method. Then $f(\mathbf{x}^{k+1}) \le f(\mathbf{x}^k)$ for every $k \ge 0$ and equality holds if and only if $\mathbf{x}^k = \mathbf{x}^{k+1}$ and $\mathbf{x}^k$ is a CW-minimum.*

Theorem 3.3 establishes the main convergence result for the greedy simplex-sparse algorithm, namely, that its accumulation points are CW-minima.

THEOREM 3.3. *Let $\{\mathbf{x}^k\}$ be the sequence generated by the greedy sparse-simplex method. Then any accumulation point of $\{\mathbf{x}^k\}$ is a CW-minimum of* (P).

Proof. By Lemma 3.2 the sequence of function values $\{f(\mathbf{x}^k)\}$ is nonincreasing and by Assumption 1 is also bounded below. Therefore, $\{f(\mathbf{x}^k)\}$ converges. Suppose that $\mathbf{x}^*$ is an accumulation point of $\{\mathbf{x}^k\}$. Then there exists a subsequence $\{\mathbf{x}^{p_n}\}_{n\ge0}$ that converges to $\mathbf{x}^*$. Suppose that $\|\mathbf{x}^*\|_0 = s$. Then the convergence of $\{\mathbf{x}^{k_n}\}$ to $\mathbf{x}^*$ implies that there exists a subsequence of $\{\mathbf{x}^{p_n}\}_{n\ge0}$, which we will denote by $\{\mathbf{x}^{k_n}\}$, such that $I_1(\mathbf{x}^{k_n}) = I_1(\mathbf{x}^*)$ for all n. Let $i \in I_1(\mathbf{x}^*), j \in \{1, 2, \ldots, n\}$, and $t \in \mathbb{R}$. By definition of the method it follows that

$$f(\mathbf{x}^{k_n}) - f(\mathbf{x}^{k_n+1}) \geq f(\mathbf{x}^{k_n}) - f(\mathbf{x}^{k_n} - x_i^{k_n}\mathbf{e}_i + t\mathbf{e}_j) \text{ for all } n.$$

The convergence of $\{f(\mathbf{x}^{k_n})\}$ implies that when taking the limit $n \to \infty$ in the latter inequality, we obtain

$$0 \geq f(\mathbf{x}^*) - f(\mathbf{x}^* - x_i^*\mathbf{e}_i + t\mathbf{e}_j).$$

That is, $f(\mathbf{x}^*) \leq f(\mathbf{x}^* - x_i^*\mathbf{e}_i + t\mathbf{e}_j)$ for all $i \in I_1(\mathbf{x}^*), j \in \{1, 2, \ldots, n\}$, and $t \in \mathbb{R}$, meaning that

$$f(\mathbf{x}^*) \leq \min_{t\in\mathbb{R}} f(\mathbf{x}^* - x_i^*\mathbf{e}_i + t\mathbf{e}_j)$$

for all $i \in I_1(\mathbf{x}^*)$ and $j \in \{1, 2, \ldots, n\}$, thus showing that $\mathbf{x}^*$ is a CW-minimum.

Suppose now that $\|\mathbf{x}^*\|_0 < s$. By the convergence of $\{\mathbf{x}^{k_n}\}$ to $\mathbf{x}^*$, it follows that there exists an N for which $I_1(\mathbf{x}^*) \subseteq I_1(\mathbf{x}^{k_n})$ for all $n > N$. Take $n > N$; if $i \in I_1(\mathbf{x}^*)$, than $i \in I_1(\mathbf{x}^{k_n})$, which in particular implies that

$$f(\mathbf{x}^{k_n}) - f(\mathbf{x}^{k_n+1}) \geq f(\mathbf{x}^{k_n}) - f(\mathbf{x}^{k_n} + t\mathbf{e}_i) \text{ for all } t \in \mathbb{R}.$$

Taking $n \to \infty$ in the last inequality yields the desired inequality

$$f(\mathbf{x}^*) \leq \min_{t\in\mathbb{R}} f(\mathbf{x}^* + t\mathbf{e}_i). \tag{3.7}$$

Now suppose that $i \in I_0(\mathbf{x}^*)$ and take $n > N$. If $\|\mathbf{x}^{k_n}\|_0 < s$, then by the definition of the greedy sparse-simplex method we have

$$f(\mathbf{x}^{k_n}) - f(\mathbf{x}^{k_n+1}) \geq f(\mathbf{x}^{k_n}) - f(\mathbf{x}^{k_n} + t\mathbf{e}_i). \tag{3.8}$$

On the other hand, if $\|\mathbf{x}^{k_n}\|_0 = s$, then the set $I_1(\mathbf{x}^{k_n}) \setminus I_1(\mathbf{x}^*)$ is nonempty, and we can pick an index $j_n \in I_1(\mathbf{x}^{k_n}) \setminus I_1(\mathbf{x}^*)$. By definition of the greedy sparse-simplex method we have

$$f(\mathbf{x}^{k_n}) - f(\mathbf{x}^{k_n+1}) \geq f(\mathbf{x}^{k_n}) - f(\mathbf{x}^{k_n} - x_{j_n}^{k_n}\mathbf{e}_{j_n} + t\mathbf{e}_i) \text{ for all } t \in \mathbb{R}. \tag{3.9}$$

Finally, combining (3.8) and (3.9) we arrive at the conclusion that

$$f(\mathbf{x}^{k_n}) - f(\mathbf{x}^{k_n+1}) \geq f(\mathbf{x}^{k_n}) - f(\mathbf{x}^{k_n} + \mathbf{d}_n + t\mathbf{e}_i), \tag{3.10}$$

where

$$\mathbf{d}_n = \begin{cases} 0, & \|\mathbf{x}^{k_n}\|_0 < s, \\ -x_{j_n}^{k_n}\mathbf{e}_{j_n}, & \|\mathbf{x}^{k_n}\|_0 = s. \end{cases}$$

Since $\mathbf{d}_n \to 0$ as n tends to ∞, it follows by taking the limit $n \to \infty$ in (3.10) that the inequality

$$f(\mathbf{x}^*) \leq f(\mathbf{x}^* + t\mathbf{e}_i)$$

holds for all $t \in \mathbb{R}$, showing that also in this case $\mathbf{x}^*$ is a CW-minimum. ∎

Combining Theorem 3.3 with Theorem 2.4 leads to the following corollary.

COROLLARY 3.2. *Suppose that Assumption 2 holds and let $\{\mathbf{x}^k\}$ be the sequence generated by the greedy sparse-simplex method. Then any accumulation point of $\{\mathbf{x}^k\}$ is an $L_2(f)$-stationary point of* (P).

3.2.1. The case $f = f_{\mathrm{LI}}$. We consider now the greedy sparse-simplex method when $f \equiv f_{\mathrm{LI}}$. At step (3.5) we perform the minimization $t_i = \arg\min f(\mathbf{x}^k + t\mathbf{e}_i)$. Since $f(\mathbf{x}^k + t\mathbf{e}_i) = \|\mathbf{A}\mathbf{x}^k - \mathbf{b} + t\mathbf{a}_i\|^2$ ($\mathbf{a}_i$ being the ith column of $\mathbf{A}$), we have immediately that

$$t_i = -\frac{\mathbf{a}_i^T \mathbf{r}_k}{\|\mathbf{a}_i\|^2},$$

where $\mathbf{r}_k = \mathbf{A}\mathbf{x}^k - \mathbf{b}$. We can then continue to compute

$$f_i = \left\| \mathbf{r}_k - \frac{\mathbf{a}_i^T \mathbf{r}_k}{\|\mathbf{a}_i\|^2} \mathbf{a}_i \right\|^2 = \|\mathbf{r}_k\|^2 - \frac{(\mathbf{a}_i^T \mathbf{r}_k)^2}{\|\mathbf{a}_i\|^2} \tag{3.11}$$

so that

$$i_k \in \operatorname*{argmin}_{i=1,\ldots,n} f_i = \operatorname*{argmax}_{i=1,\ldots,n} \frac{|\mathbf{a}_i^T \mathbf{r}_k|}{\|\mathbf{a}_i\|}.$$

The algorithm then proceeds as follows. For $\|\mathbf{x}^k\|_0 < s$ we choose

$$i_k \in \operatorname*{argmax}_{i=1,\ldots,n} \frac{|\mathbf{a}_i^T \mathbf{r}_k|}{\|\mathbf{a}_i\|}. \tag{3.12}$$

If $\mathbf{a}_{i_k}^T \mathbf{r}_k \neq 0$, then we set

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \frac{\mathbf{a}_{i_k}^T \mathbf{r}_k}{\|\mathbf{a}_{i_k}\|^2} \mathbf{e}_{i_k}.$$

In this case,

$$\mathbf{r}_{k+1} = \mathbf{A}\mathbf{x}^{k+1} - \mathbf{b} = \mathbf{r}_k - \frac{\mathbf{a}_{i_k}^T \mathbf{r}_k}{\|\mathbf{a}_{i_k}\|^2} \mathbf{a}_{i_k}.$$

Otherwise we stop. Note that if $\mathbf{A}$ has full row-rank, then $\mathbf{a}_{i_k}^T \mathbf{r}_k = 0$ only if $\mathbf{r}_k = 0$.

For $\|\mathbf{x}^k\|_0 = s$ we choose

$$(i_k, j_k) = \operatorname*{argmax}_{i \in I_1(\mathbf{x}^k), j \in \{1,2,\ldots,n\}} \frac{|\mathbf{a}_j^T \mathbf{r}_k^i|}{\|\mathbf{a}_j\|}$$

with $\mathbf{r}_k^i = \mathbf{A}\mathbf{x}^k - x_i^k \mathbf{a}_i - \mathbf{b}$. Let $f_{i_k,j_k} = f(\mathbf{x}^k - x_{i_k}^k \mathbf{e}_{i_k} + t\mathbf{e}_{j_k})$ with

$$t = -\frac{\mathbf{a}_{j_k}^T \mathbf{r}_k^{i_k}}{\|\mathbf{a}_{j_k}\|^2}.$$

If $f_{i_k,j_k} < f(\mathbf{x}^k)$, then we set

$$\mathbf{x}^{k+1} = \mathbf{x}^k - x_{i_k}^k \mathbf{e}_{i_k} - \frac{\mathbf{a}_{j_k}^T \mathbf{r}_k^{i_k}}{\|\mathbf{a}_{j_k}\|^2} \mathbf{e}_{j_k}.$$

Otherwise we stop.

It is interesting to compare the resulting iterations with the MP algorithm [21] designed to find a sparse solution to the system $\mathbf{A}\mathbf{x} = \mathbf{b}$. The MP method begins with an initial guess of $\mathbf{x}^0 = \mathbf{0}$ and $\mathbf{r}_0 = \mathbf{b}$. At each iteration, we add an element to the support by choosing

$$m \in \operatorname*{argmax}_{i=1,2,\ldots,n} \frac{|\mathbf{a}_i^T \mathbf{r}_k|}{\|\mathbf{a}_i\|}. \tag{3.13}$$

The current estimate of $\mathbf{x}$ is then updated as

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \frac{\mathbf{a}_m^T \mathbf{r}_k}{\|\mathbf{a}_m\|^2} \mathbf{e}_m, \tag{3.14}$$

and the residual is updated as

$$\mathbf{r}^{k+1} = \mathbf{A}\mathbf{x}^{k+1} - \mathbf{b} = \mathbf{r}_k - \frac{\mathbf{a}_m^T \mathbf{r}_k}{\|\mathbf{a}_m\|^2} \mathbf{a}_m. \tag{3.15}$$

The iterations continue until there are s elements in the support. Evidently, the MP method coincides with our method as long as the support is smaller than s. Our approach, however, has several advantages:

- We do not need to initialize it with a zero vector.
- In MP once an index m is added to the support it will not be removed unless in some iteration $\mathbf{a}_m^T \mathbf{r}_k = x_m \|\mathbf{a}_m\|^2$ and m maximizes $\mathbf{a}_i^T \mathbf{r}_k / \|\mathbf{a}_i\|$. On the other hand, our approach allows us to remove elements from the support under much broader conditions. Thus, there is an inherent "correction" scheme incorporated into our algorithm.
- In MP the algorithm stops once the maximal support is achieved. In contrast, in our approach further iterations are made by utilizing the correction mechanism.

We note that once our method converges to a fixed support set, it continues to update the values on the support. Ultimately, it converges to the least-squares solution on the support since in this situation the method is a simple coordinate descent method employed on a convex function. This is similar in spirit to the OMP approach [20]. The OMP proceeds similarly to the MP method; however, at each stage it updates the vector $\mathbf{x}^k$ as the least-squares solution on the current support. In our approach, we will converge to the least-squares solution on the final support; however, in choosing the support values we do not perform this orthogonalization. Instead, we allow for a correction stage which aids in correcting erroneous decisions.

3.2.2. Examples.

Example 3.3. Consider the sparse least-squares problem

$$(P_2) \quad \min\{\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 : \mathbf{x} \in C_2\},$$

where $\mathbf{A} \in \mathbb{R}^{4\times 5}$ and $\mathbf{b} \in \mathbb{R}^4$ were constructed as follows. First, the components were randomly and independently generated from a standard normal distribution, and then all the columns were normalized. The vector $\mathbf{b}$ was chosen as $\mathbf{b} \equiv \mathbf{A}\mathbf{x}_{\text{true}}$, where $\mathbf{x}_{\text{true}} = (1, -1, 0, 0, 0)^T$, so that $\mathbf{x}_{\text{true}}$ is the optimal solution of the problem. Specifically, the MATLAB code constructing $\mathbf{A}$ and $\mathbf{b}$ is

```
randn('seed',327);
A=randn(4,5);
A=A*diag(1./sqrt(sum(A.^2)));
b=A*[1;-1;0;0;0]
```

The problem has 10 BF vectors (corresponding to the 5-choose-2 options for the support of the solution) and they are denoted by $1, 2, \ldots, 10$, where the first solution is the optimal solution $\mathbf{x}_{\text{true}}$. The corresponding objective function values and stationarity levels (with two digits of accuracy) are given in Table 3.1.

TABLE 3.1
Function values and stationarity levels of the 10 *BF vectors of* (P_2).

BF vector number	1	2	3	4	5	6	7	8	9	10
Function value	0.00	0.81	0.90	0.43	0.43	0.93	0.30	1.08	0.81	2.30
Stationarity level	0.00	2.90	8.46	0.91	1.08	13.97	0.69	18.70	1.50	9.05

In this problem $L(f) = 4.78$ and $L_2(f) = 3.4972$. We compared three methods:

- the IHT method w.th $L_1 = 2L(f)$,
- the IHT method with $L_2 = 1.1L(f)$,
- the greedy sparse-simplex method.

Each of these methods was run 1000 times with different randomly generated starting points. All the runs converged to one of the 10 BF vectors. The number of times each method converged to each of the BF vectors is given in Table 3.2.

First note that when employing IHT with $L_1 = 2L(f) = 9.56$, the method never converged to the BF vectors 6, 8. The theoretical reason for this phenomena is simple: the stationarity levels of these two points are 13.97 and 18.70, and they are therefore *not* 9.56-stationary points. When employing IHT with $L_2 = 1.1 \cdot L(f) = 5.26$, there are two additional BF vectors—3 and 10—to which convergence is impossible, because their stationarity level is 8.46 and 9.05. This illustrates the fact that as L gets larger, there are more nonoptimal candidates to which IHT can converge. The greedy sparse-simplex algorithm exhibits the best results with more than 80% chance to converge to the true optimal solution. Note that this method will never converge to the BF vectors $3, 6, 8$, and 10 since they are not $L_2(f)$-stationary points. Moreover, there are only three possible BF vectors to which the greedy sparse-simplex algorithm converge: $1, 4$, and 7. The reason is that among the 10 BF vectors, there are only three CW-minima. This illustrates the fact that even though any CW-minimum is an $L_2(f)$-stationary point, the reverse claim is not true—there are $L_2(f)$-stationary points which are not CW-minima.

In Table 3.3 we describe the 11 first iterations of the greedy sparse-simplex method. Note that at the fourth iteration the algorithm "finds" the correct support, and the rest of the iterations are devoted to computing the exact values of the nonnegative components of the BF vector.

Example 3.4 (comparison with MP and OMP). To compare the performance of MP and OMP to that of the greedy sparse-simplex, we generated 1000 realizations of $\mathbf{A}$ and $\mathbf{b}$ exactly as described in Example 3.3. We ran both MP and OMP on these problems with $s = 2$. Each of these methods were considered "successful" if it found the correct support. (MP usually does not find the correct values.) The greedy sparse-simplex was run with an initial vector of all zero, so that the first two iterations were

TABLE 3.2
Distribution of limit points among the 10 *BF vector.* $N_1(i)$ $(N_2(i))$ *is the amount of runs for which the IHT method with* L_1 (L_2) *converged to the* i*th BF vector.* $N_3(i)$ *is the amount of runs for which the greedy sparse-simplex method converged to the* i*th BF vector. The exact definition of* $N_4(i)$ *will be made clear in section* 3.3.

BF vector (i)	1	2	3	4	5	6	7	8	9	10
$N_1(i)$	329	50	63	92	229	0	130	0	61	46
$N_2(i)$	340	59	0	89	256	0	187	0	69	0
$N_3(i)$	813	0	0	112	0	0	75	0	0	0
$N_4(i)$	772	0	0	92	0	0	93	0	43	0

TABLE 3.3
First 11 *iterations of the greedy sparse-simplex method with starting point* $(0, 1, 5, 0, 0, 0)^T$.

Iteration number	x_1	x_2	x_3	x_4	x_5
0	0	1	5	0	0
1	0	1.0000	1.5608	0	0
2	0	0	1.5608	0	-0.6674
3	1.6431	0	0	0	-0.6674
4	1.6431	-0.8634	0	0	0
5	1.0290	-0.8634	0	0	0
6	1.0290	-0.9938	0	0	0
7	1.0013	-0.9938	0	0	0
8	1.0013	-0.9997	0	0	0
9	1.0001	-0.9997	0	0	0
10	1.0001	-1.0000	0	0	0
11	1.0000	-1.0000	0	0	0

identical to MP. The results were the following: of the 1000 realizations, both MP and OMP found the correct support in 452 cases. The greedy sparse-simplex method, which adds "correcting" steps to MP, was able to recover the correct support in 652 instances.

An additional advantage of greedy sparse-simplex is that it is capable of running from various starting points. We therefore added the following experiment: for each realization of $\mathbf{A}$ and $\mathbf{b}$, we ran the greedy sparse-simplex method from five different initial vectors generated in the same way as in Example 3.3 (and not the all-zeros vector). If at least one of these five runs detected the correct support, then the experiment is considered to be a success. In this case the correct support was found 952 times out of the 1000 realizations.

The example above illustrates an important feature of the greedy sparse-simplex algorithm: since it can be initialized with varying starting points, it is possible to improve its performance by using several starting points and obtaining several possible sparse solutions. The final solution can then be taken as the one with minimal objective function value. This feature provides additional flexibility over the MP and OMP methods.

3.3. The partial sparse-simplex method. The greedy sparse-simplex algorithm, as illustrated in Example 3.3, has several advantages over the IHT method: first, its limit points satisfy stronger optimality conditions, and as a result it is more likely to converge to the optimal solution; second, it does not require knowledge of a Lipschitz constant. On the other hand, the computational effort per iteration of the greedy sparse-simplex is larger than the one required by IHT. Indeed, in the worst case it requires the call for $O(s \cdot n)$ one-dimensional minimization procedures; this computational burden is caused by the fact that the method has no index selection strategy. That is, instead of deciding a priori according to some policy on the index or indices on which the optimization will be performed, the algorithm invokes an optimization procedure for all possible choices and then picks the index resulting in the minimal objective function value.

The *partial sparse-simplex method* described below has an extremely simple way to choose the index or indices on which the optimization will be performed. The only difference from the greedy sparse-simplex algorithm is in the case when $\|\mathbf{x}^k\|_0 = s$, where there are two options: Either perform a minimization with respect to the variable in the support of $\mathbf{x}^k$ which causes the maximum decrease in function value, or replace the variable in the support with the smallest absolute value (that is,

substituting zero instead of the current value) with the nonsupport variable corresponding to the largest absolute value of the partial derivative—the value of the new nonzero variable is set by performing a minimization procedure with respect to it. Finally, the best of the two choices (in terms of objective function value) is selected. Since the method is no longer "greedy" and only considers part of the choices for the pair of indices, we will call it *the partial sparse-simplex method*.

The Partial Sparse-Simplex Method

- **Initialization:** $\mathbf{x}^0 \in C_s$.
- **General Step** ($k = 0, 1, 2, \ldots$)**:**
  - If $\|\mathbf{x}^k\|_0 < s$, then compute for every $i = 1, 2, \ldots, n$

$$t_i \in \operatorname*{argmin}_{t\in\mathbb{R}} f(\mathbf{x}^k + t\mathbf{e}_i),$$
$$f_i = f(\mathbf{x}^k + t_i\mathbf{e}_i).$$

    Let $i_k \in \operatorname*{argmin}_{i=1,\ldots,n} f_i$. If $f_{i_k} < f(\mathbf{x}^k)$, then set

$$\mathbf{x}^{k+1} = \mathbf{x}^k + t_{i_k}\mathbf{e}_{i_k}.$$

    Otherwise, STOP.
  - If $\|\mathbf{x}^k\|_0 = s$, then compute for every $i \in I_1(\mathbf{x}^k)$

$$t_i \in \operatorname*{argmin}_{t\in\mathbb{R}} f(\mathbf{x}^k + t\mathbf{e}_i),$$
$$f_i = f(\mathbf{x}^k + t_i\mathbf{e}_i).$$

    Let

$$i_k^1 \in \operatorname{argmin}\{f_i : i \in I_1(\mathbf{x}^k)\},$$
$$i_k^2 \in \operatorname{argmax}\{|\nabla_i f(\mathbf{x}^k)| : i \in I_0(\mathbf{x}^k)\},$$
$$m_k \in \operatorname{argmin}\{|x_i^k| : i \in I_1(\mathbf{x}^k)\},$$

    and let

$$D_k^1 = \min_{t\in\mathbb{R}} f(\mathbf{x}^k + t\mathbf{e}_{i_k^1}), \qquad T_k^1 \in \operatorname*{argmin}_{t\in\mathbb{R}} f(\mathbf{x}^k + t\mathbf{e}_{i_k^1})$$
$$D_k^2 = \min_{t\in\mathbb{R}} f(\mathbf{x}^k - x_{m_k}^k\mathbf{e}_{m_k} + t\mathbf{e}_{i_k^2}), \quad T_k^2 \in \operatorname*{argmin}_{t\in\mathbb{R}} f(\mathbf{x}^k - x_{m_k}^k\mathbf{e}_{m_k} + t\mathbf{e}_{i_k^2})$$

    If $D_k^1 < D_k^2$, then set

$$\mathbf{x}^{k+1} = \mathbf{x}^k + T_k^1\mathbf{e}_{i_k^1}.$$

    Else

$$\mathbf{x}^{k+1} = \mathbf{x}^k - x_{m_k}^k\mathbf{e}_{m_k} + T_k^2\mathbf{e}_{i_k^2}.$$

Remark 3.4. The partial sparse-simplex coincides with the greedy sparse-simplex when $\|\mathbf{x}^k\|_0 < s$. Therefore, when $f \equiv f_{\mathrm{LI}}$, the partial sparse-simplex method coincides with MP for the first s steps and when the initial vector is the vector of all zeros.

The basic property of the partial sparse-simplex method is that it generates a nonincreasing sequence of function values and that all its limit points are BF vectors.

LEMMA 3.3. *Let $\{\mathbf{x}^k\}$ be the sequence generated by the partial sparse-simplex method. Then any accumulation point of $\{\mathbf{x}^k\}$ is a BF vector.*

Proof. The proof of Theorem 3.3 until (3.7) is still valid for the partial sparse-simplex method, so that for any $i \in I_1(\mathbf{x}^*)$ and any $t \in \mathbb{R}$,

$$f(\mathbf{x}^*) \leq f(\mathbf{x}^* + t\mathbf{e}_i),$$

which in particular means that $0 \in \operatorname{argmin}\{g_i(t) \equiv f(\mathbf{x}^* + t\mathbf{e}_i)\}$ and thus $\nabla_i f(\mathbf{x}^*) = g_i'(0) = 0$. ∎

The limit points of the partial sparse-simplex are not necessarily CW-minima. However, when Assumption 2 holds, they are $L_2(f)$-stationary points, which is a better result than the one known for IHT.

THEOREM 3.4. *Suppose that Assumption 2 holds and let $\{\mathbf{x}^k\}$ be the sequence generated by the partial sparse-simplex method. Then any accumulation point of $\{\mathbf{x}^k\}$ is an $L_2(f)$-stationary point.*

The proof of the theorem relies on the following lemma.

LEMMA 3.4. *Suppose that Assumption 2 holds and let $\{\mathbf{x}^k\}$ be the sequence generated by the partial sparse-simplex method. Then for any k for which $\|\mathbf{x}^k\|_0 < s$ it holds that*

$$f(\mathbf{x}^k) - f(\mathbf{x}^{k+1}) \geq \frac{1}{2L_2(f)} \max_{i=1,2,\ldots,n} (\nabla_i f(\mathbf{x}^k))^2. \tag{3.16}$$

For any k with $\|\mathbf{x}^k\|_0 = s$, the inequality

$$f(\mathbf{x}^k) - f(\mathbf{x}^{k+1}) \geq A(\mathbf{x}^k) \tag{3.17}$$

holds true with

$$\begin{aligned} A(\mathbf{x}) \equiv \max \Bigg\{ & \frac{1}{2L_2(f)} \max_{i \in I_1(\mathbf{x})} (\nabla_i f(\mathbf{x}))^2, M_s(\mathbf{x}) \\ & \times \left[\max_{i \in I_0(\mathbf{x})} |\nabla_i f(\mathbf{x})| - \max_{i \in I_1(\mathbf{x})} |\nabla_i f(\mathbf{x})| - L_2(f) M_s(\mathbf{x}) \right] \Bigg\}. \end{aligned} \tag{3.18}$$

Proof. Suppose that $\|\mathbf{x}^k\|_0 < s$. Then by the definition of the method we have for all $i = 1, 2, \ldots, n$,

$$f(\mathbf{x}^{k+1}) \leq f\left(\mathbf{x}^k - \frac{1}{L_2(f)} \nabla_i f(\mathbf{x}^k) \mathbf{e}_i \right). \tag{3.19}$$

On the other hand, for any $i = 1, 2, \ldots, n$,

$$\begin{aligned} f\left(\mathbf{x}^k - \frac{1}{L_2(f)} \nabla_i f(\mathbf{x}^k) \mathbf{e}_i \right) &\leq f(\mathbf{x}^k) - \frac{1}{L_2(f)} (\nabla_i f(\mathbf{x}^k))^2 + \frac{1}{2L_2(f)} (\nabla_i f(\mathbf{x}^k))^2 \\ &\qquad \text{(Lemma 2.6)} \\ &= f(\mathbf{x}^k) - \frac{1}{2L_2(f)} (\nabla_i f(\mathbf{x}^k))^2, \end{aligned}$$

which combined with (3.19) implies that

$$f(\mathbf{x}^k) - f(\mathbf{x}^{k+1}) \geq \frac{1}{2L_2(f)} \max_{i=1,2,\ldots,n} (\nabla_i f(\mathbf{x}^k))^2,$$

establishing (3.16).

Next, suppose that $\|\mathbf{x}^k\|_0 = s$. A similar argument to the one just invoked shows that

$$f(\mathbf{x}^k) - f(\mathbf{x}^{k+1}) \geq \frac{1}{2L_2(f)} \max_{i \in I_1(\mathbf{x}^k)} (\nabla_i f(\mathbf{x}^k))^2. \tag{3.20}$$

By the definition of the greedy sparse-simplex method, it follows that

$$f(\mathbf{x}^k) - f(\mathbf{x}^{k+1}) \geq f(\mathbf{x}^k) - f(\mathbf{x}^k - x^k_{m_k}\mathbf{e}_{m_k} + T_k^2 \mathbf{e}_{i_k^2}) \geq f(\mathbf{x}^k) - f(\mathbf{x}^k - x^k_{m_k}\mathbf{e}_{m_k} - \sigma x^k_{m_k}\mathbf{e}_{i_k^2}), \tag{3.21}$$

where $\sigma = \operatorname{sgn}(x^k_{m_k} \nabla_{i_k^2} f(\mathbf{x}^k))$. Using the local descent lemma (Lemma 2.6) once more, we obtain that

$$\begin{aligned} &f(\mathbf{x}^k - x^k_{m_k}\mathbf{e}_{m_k} - \sigma x^k_{m_k}\mathbf{e}_{i_k^2}) \\ &\quad \leq f(\mathbf{x}^k) + \nabla f(\mathbf{x}^k)^T(-x^k_{m_k}\mathbf{e}_{m_k} - \sigma x^k_{m_k}\mathbf{e}_{i_k^2}) + \frac{L_2(f)}{2}\left\|-x^k_{m_k}\mathbf{e}_{m_k} - \sigma x^k_{m_k}\mathbf{e}_{i_k^2}\right\|^2 \\ &\quad = f(\mathbf{x}^k) - x^k_{m_k}\nabla_{m_k} f(\mathbf{x}^k) - \sigma x^k_{m_k}\nabla_{i_k^2} f(\mathbf{x}^k) + L_2(f)(x^k_{m_k})^2 \\ &\quad = f(\mathbf{x}^k) + M_s(\mathbf{x}^k)\left[L_2(f)M_s(\mathbf{x}^k) - |\nabla_{i_k^2} f(\mathbf{x}^k)|\right] - x^k_{m_k}\nabla_{m_k} f(\mathbf{x}^k). \end{aligned} \tag{3.22}$$

Combining (3.21) and (3.22) we obtain that

$$f(\mathbf{x}^k) - f(\mathbf{x}^{k+1}) \geq M_s(\mathbf{x}^k)\left[\max_{i \in I_0(\mathbf{x}^k)} |\nabla_i f(\mathbf{x}^k)| - L_2(f)M_s(\mathbf{x}^k)\right] + x^k_{m_k}\nabla_{m_k} f(\mathbf{x}^k). \tag{3.23}$$

Finally, (3.20) and (3.23) along with the fact that

$$x^k_{m_k}\nabla_{m_k} f(\mathbf{x}^k) \geq -M_s(\mathbf{x}^k) \max_{i \in I_1(\mathbf{x}^k)} |\nabla_i f(\mathbf{x}^k)|$$

readily imply the inequality (3.17). □

We now turn to prove Theorem 3.4.

Proof of Theorem 3.4. Let $\mathbf{x}^*$ be an accumulation point of the generated sequence. Then there exists a subsequence $\{\mathbf{x}^{k_n}\}_{n \geq 0}$ converging to $\mathbf{x}^*$. Suppose first that $\|\mathbf{x}^*\|_0 = s$. Then there exists an $N > 0$ such that $I_1(\mathbf{x}^{k_n}) = I_1(\mathbf{x}^*)$ for all $n > N$. Therefore, by (3.17) we have

$$f(\mathbf{x}^{k_n}) - f(\mathbf{x}^{k_n+1}) \geq A(\mathbf{x}^{k_n}) \tag{3.24}$$

for all $n > N$. Since $\{f(\mathbf{x}^k)\}$ is a nonincreasing and lower bounded sequence, it follows that the left-hand side of the inequality (3.24) tends to 0 as $n \to \infty$. Therefore, by the continuity of the operator A we have $A(\mathbf{x}^*) \leq 0$, from which it follows that

$$\frac{1}{2L_2(f)} \max_{i \in I_1(\mathbf{x}^*)} (\nabla_i f(\mathbf{x}^*))^2 = 0, \tag{3.25}$$

$$M_s(\mathbf{x}^*)\left[\max_{i \in I_0(\mathbf{x}^*)} |\nabla_i f(\mathbf{x}^*)| - \max_{i \in I_1(\mathbf{x}^*)} |\nabla_i f(\mathbf{x}^*)| - L_2(f)M_s(\mathbf{x}^*)\right] \leq 0. \tag{3.26}$$

By (3.25) it follows that $\nabla_i f(\mathbf{x}^*) = 0$ for all $i \in I_1(\mathbf{x}^*)$ and substituting this into (3.26) yields the inequality

$$\max_{i \in I_0(\mathbf{x}^*)} |\nabla_i f(\mathbf{x}^*)| \leq L_2(f)M_s(\mathbf{x}^*),$$

meaning that $\mathbf{x}^*$ is an $L_2(f)$-stationary point.

Now suppose that $\|\mathbf{x}^*\|_0 < s$. There are two cases. If there exists an infinite number of n-s for which $\|\mathbf{x}^{k_n}\|_0 < s$, then by Lemma 3.4 for each such n

$$f(\mathbf{x}^{k_n}) - f(\mathbf{x}^{k_n+1}) \geq \frac{1}{2L_2(f)} \max_{i=1,2,\ldots,n} \nabla_i f(\mathbf{x}^{k_n})^2,$$

and therefore by taking $n \to \infty$ along the n-s for which $\|\mathbf{x}^{k_n}\|_0 < s$, we obtain that $\nabla f(\mathbf{x}^*) = 0$. If, on the other hand, there exists an integer N such that the equality $\|\mathbf{x}^{k_n}\|_0 = s$ holds for all $n > N$, then by the definition of the method we have for all $n > N$

$$f(\mathbf{x}^{k_n}) - f(\mathbf{x}^{k_n+1}) \geq \frac{1}{2L_2(f)} \max_{i \in I_1(\mathbf{x}^{k_n})} (\nabla_i f(\mathbf{x}^{k_n}))^2 \tag{3.27}$$

and

$$\begin{aligned} f(\mathbf{x}^{k_n}) - f(\mathbf{x}^{k_n+1}) &\geq f(\mathbf{x}^{k_n}) - f(\mathbf{x}^{k_n} - x^k_{m_k}\mathbf{e}_{m_k} + T^2_k \mathbf{e}_{i^2_k}) \\ &= f(\mathbf{x}^{k_n}) - f(\mathbf{x}^{k_n} - x^k_{m_k}\mathbf{e}_{m_k}) + f(\mathbf{x}^{k_n} - x^k_{m_k}\mathbf{e}_{m_k}) \\ &\quad - f(\mathbf{x}^{k_n} - x^k_{m_k}\mathbf{e}_{m_k} + T^2_k \mathbf{e}_{i^2_k}). \end{aligned} \tag{3.28}$$

Since $T^2_k \in \operatorname{argmin}_{t\in\mathbb{R}} f(\mathbf{x}^{k_n} - x^k_{m_k}\mathbf{e}_{m_k} + t\mathbf{e}_{i^2_k})$, then

$$\begin{aligned} f(\mathbf{x}^{k_n} - x^k_{m_k}\mathbf{e}_{m_k}) - f(\mathbf{x}^{k_n} - x^k_{m_k}\mathbf{e}_{m_k} + T^2_k \mathbf{e}_{i^2_k}) &\geq \frac{1}{2L_2(f)} (\nabla_{i^2_k} f(\mathbf{x}^{k_n} - x^k_{m_k}\mathbf{e}_{m_k}))^2 \\ &= \frac{1}{2L_2(f)} \max_{i \in I_0(\mathbf{x}^k)} (\nabla_i f(\mathbf{x}^{k_n} - x^k_{m_k}\mathbf{e}_{m_k}))^2, \end{aligned}$$

which combined with (3.28) yields

$$\frac{1}{2L_2(f)} \max_{i \in I_0(\mathbf{x}^k)} (\nabla_i f(\mathbf{x}^{k_n} - x^k_{m_k}\mathbf{e}_{m_k}))^2 \leq f(\mathbf{x}^{k_n} - x^k_{m_k}\mathbf{e}_{m_k}) - f(\mathbf{x}^{k_n+1}). \tag{3.29}$$

In addition,

$$\begin{aligned} |\nabla_i f(\mathbf{x}^{k_n})| &\leq |\nabla_i f(\mathbf{x}^{k_n}) - \nabla_i f(\mathbf{x}^{k_n} - x^k_{m_k}\mathbf{e}_{m_k})| + |\nabla_i f(\mathbf{x}^{k_n} - x^k_{m_k}\mathbf{e}_{m_k})| \\ &\leq L_2(f)|x^k_{m_k}| + |\nabla_i f(\mathbf{x}^{k_n} - x^k_{m_k}\mathbf{e}_{m_k})| \\ &= L_2(f) M_s(\mathbf{x}^{k_n}) + |\nabla_i f(\mathbf{x}^{k_n} - x^k_{m_k}\mathbf{e}_{m_k})|, \end{aligned}$$

and thus (3.29) readily implies that

$$\max_{i \in I_0(\mathbf{x}^k)} |\nabla_i f(\mathbf{x}^{k_n})| \leq L_2(f) M_s(\mathbf{x}^{k_n}) + \sqrt{2L_2(f)[f(\mathbf{x}^{k_n} - x^k_{m_k}\mathbf{e}_{m_k}) - f(\mathbf{x}^{k_n+1})]},$$

which together with (3.27) yields that for all $i = 1, 2, \ldots, n$

$$\begin{aligned} |\nabla_i f(\mathbf{x}^{k_n})| \leq \min \Big\{ & L_2(f) M_s(\mathbf{x}^{k_n}) + \sqrt{2L_2(f)[f(\mathbf{x}^{k_n} - x^k_{m_k}\mathbf{e}_{m_k}) - f(\mathbf{x}^{k_n+1})]}, \\ & \sqrt{2L_2(f)[f(\mathbf{x}^{k_n}) - f(\mathbf{x}^{k_n+1})]} \Big\}. \end{aligned}$$

Since the right-hand side of the latter inequality converges to 0 as $n \to \infty$, it follows that the desired result $\nabla f(\mathbf{x}^*) = 0$ holds. ∎

We end this section by returning to Example 3.3, and adding a comparison to the partial sparse-simplex algorithm.

Example 3.4 (Example 3.3 continued). In Example 3.3 we added 1000 runs of the partial sparse-simplex method. The results can be found in Table 3.2 under $N_4(i)$, which is the amount of times in which the algorithm converged to the ith BF vector. As can be seen, the method performs very well, much better than IHT with either $L_1 = 1.1L(f)$ or $L_2 = 2L(f)$. It is only slightly inferior to the greedy sparse-simplex method since it has another BF vector to which it might converge (BF vector number 9). Thus, in this example the partial sparse-simplex is able to compare with the greedy sparse-simplex despite the fact that each iteration is much cheaper in terms of computational effort.

Example 3.5 (quadratic equations). We now consider an example of quadratic equations. Given m vectors $\mathbf{a}_1, \ldots, \mathbf{a}_m$, our problem is to find a vector $\mathbf{x} \in \mathbb{R}^n$ satisfying

$$(\mathbf{a}_i^T \mathbf{x})^2 = c_i, \quad i = 1, 2, \ldots, m, \tag{3.30}$$

$$\|\mathbf{x}\|_0 \leq s. \tag{3.31}$$

The problem of finding an $\mathbf{x} \in \mathbb{R}^n$ satisfying (3.30) and (3.31) is the same as finding an optimal solution to the optimization problem (P) with $f \equiv f_{\mathrm{QI}}$, where $\mathbf{A}_i = \mathbf{a}_i \mathbf{a}_i^T$. We compare the greedy and partial sparse-simplex algorithms on an example with $m = 80, n = 120$, and $s = 3, 4, \ldots, 10$. As noted in Remark 3.3, the greedy and partial sparse-simplex methods require solving several one-dimensional minimization problems of quartic equations at each iteration. Each component of the 80 vectors $\mathbf{a}_1, \ldots, \mathbf{a}_8 0$ was randomly and independently generated from a standard normal distribution. Then, the "true" vector $\mathbf{x}_{\mathrm{true}}$ was generated by choosing randomly the s nonzero components whose values were also randomly generated from a standard normal distribution. The vector $\mathbf{c}$ was then determined by $c_i = (\mathbf{a}_i^T \mathbf{x}_{\mathrm{true}})^2$. For each value of s ($s = 3, 4, \ldots, 10$), we ran the greedy and partial sparse-simplex algorithms from 100 different and randomly generated initial vectors. The numbers of runs out of 100 in which the methods found the correct solution are given in Table 3.4.

As can be clearly seen by the results in the table, the greedy sparse-simplex outperforms the partial sparse-simplex in terms of the success probability. In addition, the chances of obtaining the optimal solution decrease as s gets larger. Of course, we can easily increase the success probability of the partial sparse-simplex method by starting it from several initial vectors and taking the best result.

TABLE 3.4

The second (third) column contains the number of runs of 100 *for which the partial (greedy) sparse-simplex method converged.*

s	N_{PSS}	N_{GSS}
3	27	73
4	22	69
5	8	20
6	5	19
7	9	13
8	5	8
9	3	6
10	2	3

REFERENCES

[1] S. Bahmani, B. Raj, and P. Boufounos, *Greedy Sparsity-Constrained Optimization*, arxiv:1203.5483v2.pdf, 2012.

[2] A. Beck and M. Teboulle, *Gradient-based algorithms with applications to signal recovery problems*, in Convex Optimization in Signal Processing and Communications, Y. Eldar and D. Palomar, eds., Cambridge University Press, Cambridge, UK, 2010.

[3] E. V. D. Berg and M. P. Friedlander, *Sparse optimization with least-squares constraints*, SIAM J. Optim., 21 (2011), pp. 1201–1229.

[4] D. P. Bertsekas, *Nonlinear Programming*, 2nd ed., Athena Scientific, Belmont, MA, 1999.

[5] T. Blumensath, *Compressed Sensing with Nonlinear Observations*, http://users.fmrib.ox.ac.uk/˜tblumens/papers/B_Nonlinear.pdf (2010).

[6] T. Blumensath and M. E. Davies, *Iterative thresholding for sparse approximations*, J. Fourier Anal. Appl., 14 (2008), pp. 629–654.

[7] T. Blumensath and M. E. Davies, *Normalised iterative hard thresholding: Guaranteed stability and performance*, IEEE J. Selected Topics in Signal Processing, 4 (2010), pp. 298–309.

[8] E. Candès, J. Romberg, and T. Tao, *Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information*, IEEE Trans. Inform. Theory, 52 (2006), pp. 489–509.

[9] G. Davis, S. Mallat, and M. Avellaneda, *Adaptive greedy approximations*, Constr. Approx., 13 (1997), pp. 57–98.

[10] R. DeVore, *Nonlinear approximation*, Acta Numerica, 7 (1998), pp. 51–150.

[11] D. Donoho, *Denoising by soft-thresholding*, IEEE Trans. Inform. Theory, 41 (1995), pp. 613–627.

[12] D. L. Donoho, *Compressed sensing*, IEEE Trans. Inform. Theory, 52 (2006), pp. 1289–1306.

[13] D. L. Donoho and M. Elad, *Optimally sparse representation in general (non-orthogonal) dictionaries via l*1 *minimization*, in Proc. Natl. Acad. Sci. USA, 100 (2003), pp. 2197–2202.

[14] A. Szameit et al., *Sparsity-based single-shot sub-wavelength coherent diffractive imaging*, Nature Materials, 11 (2012), pp. 455–459.

[15] J. R. Fienup, *Phase retrieval algorithms: A comparison*, Appl. Optics, 21 (1982), pp. 2758–2769.

[16] R. W. Gerchberg and W. O. Saxton, *A practical algorithm for the determination of phase from image and diffraction plane pictures*, Optik, 35 (1972), pp. 237–246.

[17] I. F. Gorodnitsky and B. D. Rao, *Sparse signal reconstruction from limited data using FOCUSS: A re-weighted minimum norm algorithm*, IEEE Trans. Signal Process., 45 (1997), pp. 600–616.

[18] N. Hurt, *Phase Retrieval and Zero Crossings*, Kluwer Academic, Norwell, MA, 1989.

[19] M. A. Davenport, M. F. Duarte, Y. C. Eldar, and G. Kutyniok, *Introduction to Compressed Sensing*, in Compressed Sensing: Theory and Applications, Cambridge University Press, Cambridge, MA, 2012.

[20] S. Mallat, *A Wavelet Tour of Signal Processing: The Sparse Way*, Academic Press, New York, 2008.

[21] S. Mallat and Z. Zhang, *Matching pursuits with time-frequency dictionaries*, IEEE Trans. Signal Process., 41 (1993), pp. 3397–3415.

[22] B. Olshausen and D. Field, *Emergence of simple-cell receptive field properties by learning a sparse representation*, Nature, 381 (1996), pp. 607–609.

[23] S. Scholtes, *Nonconvex structures in nonlinear programming*, Oper. Res., 52 (2004), pp. 368–383.

[24] Y. Shechtman, Y. C. Eldar, A. Szameit, and M. Segev, *Sparsity-based sub-wavelength imaging with partially spatially incoherent light via quadratic compressed sensing*, Optics Express, 19 (2011), pp. 14807–14822.

[25] D. Taubman and M. Marcellin, *JPEG* 2000*: Image Compression Fundamentals, Standards and Practice*, Kluwer, Dordrecht, Netherlands, 2001.

[26] R. Tibshirani, *Regression shrinkage and selection via the lasso*, J. Royal Statist. Soc. B, 58 (1996), pp. 267–288.

[27] J. Tropp, *Greed is good: Algorithmic results for sparse approximation*, IEEE Trans. Inform. Theory, 50 (2004), pp. 2231–2242.

[28] J. Tropp and S. J. Wright, *Computational methods for sparse solution of linear inverse problems*, Proc. IEEE, 98 (2010), pp. 948–958.

[29] R. Vershynin, *Introduction to the Non-asymptotic Analysis of Random Matrices*, in Compressed Sensing: Theory and Applications, Cambridge University Press, Cambridge, UK, 2012.