PROXIMAL MAPPING FOR SYMMETRIC PENALTY AND SPARSITY

AMIR BECK [*] AND NADAV HALLAK [†]

Abstract. This paper studies a class of problems consisting of minimizing a continuously differentiable function penalized with the so-called ℓ_0-norm over a symmetric set. These problems are hard to solve, yet prominent in many fields and applications. We first study the proximal mapping with respect to the ℓ_0-norm over symmetric sets, and provide an efficient method to attain it. The method is then improved for symmetric sets satisfying a sub-modularity-like property, which we call "second order monotonicity" (SOM). It is shown that many important symmetric sets, such as the $\ell_1, \ell_2, \ell_\infty$-balls, the simplex and the full-simplex, satisfy this SOM property. We then develop, under the validity of the SOM property, necessary optimality conditions, and corresponding algorithms that are guaranteed to converge to points satisfying the aforementioned optimality conditions. We prove the existence of a hierarchy between the optimality conditions, and consequently between the corresponding algorithms.

1. Introduction.

Sparsity plays an important role in countless applications in various fields, especially in the emerging fields of compressed sensing and image and signal processing (see the in-depths reviews [15, 18, 19, 27]). Optimization problems involving sparsity are on the border between continuous and combinatorial optimization, and the vast majority of them are considered to be very hard (see for example [23]). This difficulty is usually treated either by relaxing the sparsity term ([4, 12, 14, 27, 29]) or by assuming restrictive assumptions ([13, 16, 17]). Since in general it is impossible to attain an optimal solution, or even to verify if a feasible solution to a sparse optimization problem is optimal, necessary optimality conditions and methods to obtain good solutions are imperative.

Sparse optimization problems usually belong to one of two classes of problems: problems with a sparsity constraint, or problems with sparsity penalty term. The literature on sparsity constrained problems is rather rich, in particular, immense literature has been accumulated on the problem of sparsest solution to a system of linear equations (see for example the review paper [11] and references therein). Several studies in recent years have expanded and generalized these results to include more general functions and additional constraints. In particular, several works have been done in the area of compressed sensing with nonlinear measurements, see for example [7] and the recent studies on sparse phase retrieval [24, 25, 28].

Several studies were conducted on the study of general optimality conditions and algorithms on sparsity-constrained problem. The work [2] presented optimality conditions and their hierarchy for sparsity constrained problems, as well as methods that guarantee to converge in some sense to points satisfying the derived optimality conditions. Later on, the work [3] studied the generalization where the feasible set is a *sparse symmetric set*, meaning a set which is an intersection of a closed convex symmetric set and the set of all k-sparse vectors (for some positive integer k). It was shown in [3] how to compute the orthogonal projection onto sparse symmetric sets efficiently, and used this result to define optimality conditions and algorithms, as well as establish a hierarchy between the different methods and conditions.

The recent work [5] studied the sparse PCA problem, devising a coordinate-wise optimality condition that was superior (i.e. more restrictive) than other frequently used stationary-based optimality conditions. The superiority (i.e. restrictiveness) of coordinate-wise based conditions over stationary-based conditions was proved and illustrated in all three papers [2, 3, 5], demonstrating the importance of studying optimality conditions in the sparse optimization setting.

Much less is known on problems incorporating a sparsity term as a penalty term. In particular, in this

[*] Faculty of Industrial Engineering and Management, Technion - Israel Institute of Technology, Haifa 32000, Israel. Email: becka@ie.technion.ac.il

[†] Faculty of Industrial Engineering and Management, Technion - Israel Institute of Technology, Haifa 32000, Israel. Email: ndvhllk@campus.technion.ac.il

paper we study the following optimization problem:

$$\min\{f(\mathbf{x}) + \lambda\|\mathbf{x}\|_0 : \mathbf{x} \in B\}, \tag{1.1}$$

where $B \subseteq \mathbb{R}^n$ satisfies some symmetry assumptions. Problem (1.1) for the specific case where $f(\mathbf{x}) \equiv \|\mathbf{Ax} - \mathbf{b}\|_2$ ($\mathbf{A} \in \mathbb{R}^{m\times n}, \mathbf{b} \in \mathbb{R}^m$) was studied in [26] where a stationarity-based optimality condition was derived. The work [8] studied the convergence of a gradient descent thresholding method to points satisfying the aforementioned optimality condition. The paper [22] studied the penalized problem over component-wise separable sets, and provided a stationarity-based optimality condition.

Paper layout. Fundamental mathematical preliminaries for sparse optimization over symmetric sets are defined and presented in Section 2. Section 3 studies the sparse proximal mapping operator over symmetric sets. Essential properties of the sparse prox are proved, and an efficient method for computing the sparse prox in the general case is derived. This method is then improved for sets that satisfy a sub-modularity like property named *second order monotonicity*. Many important sets satisfy this property, as shown in the appendix. In Section 4 we develop necessary optimality conditions and prove their hierarchy under the validity of the second order monotonicity property. The section is concluded with examples illustrating the strictness of the hierarchy. Finally, Section 5 presents methods that obtain points satisfying the new derived optimality conditions.

Notation. Matrices and vectors are denoted by boldface letters. The vector of all zeros is denoted by $\mathbf{0}$ and the vector of all ones by $\mathbf{e}$. The vector which has 1 in its ith component and zeros elsewhere is denoted by $\mathbf{e}_i$. For any $p \in [1, \infty]$ and $\alpha > 0$, the set $B_p[\mathbf{0}, \alpha] = \{\mathbf{x} : \|\mathbf{x}\|_p \leq \alpha\}$ is the ℓ_p ball with center $\mathbf{0}$ and radius α. For a vector $\mathbf{x} \in \mathbb{R}^n$, the vector $|\mathbf{x}|$ is the vector of absolute values of the components of $\mathbf{x}$, and the vector $\text{sign}(\mathbf{x})$ is defined by $\text{sign}(\mathbf{x})_i = 1$ when $x_i \geq 0$ and -1 otherwise. For any two vectors $\mathbf{x}, \mathbf{y}$ of the same dimension, $\mathbf{x} \odot \mathbf{y}$ denotes their component-wise product (a.k.a. Hadmard product). For a given set $S \subseteq \mathbb{R}^n$, the orthogonal projection of $\mathbf{x}$ onto S is defined as

$$P_S(\mathbf{x}) = \text{argmin}\{\|\mathbf{y} - \mathbf{x}\|_2^2 : \mathbf{y} \in S\}.$$

The indicator function of a given set $S \subseteq \mathbb{R}^n$ is denoted by δ_S and is given by $\delta_S(\mathbf{x}) = 0$ for $\mathbf{x} \in S$ and ∞ otherwise. The so-called ℓ_0-norm, $\|\mathbf{x}\|_0$, which counts the number of nonzero elements in $\mathbf{x}$ is defined by $\|\mathbf{x}\|_0 \equiv |\{i : x_i \neq 0\}|$. For a given integer $s \in \{1, \ldots, n\}$, the set C_s comprises all vectors with at most s non-zero elements:

$$C_s = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_0 \leq s\}.$$

In this context, s will be called "the sparsity level". The *support* set of a vector $\mathbf{x} \in \mathbb{R}^n$ is denoted by $I_1(\mathbf{x}) \equiv \{i \in \{1, \ldots, n\} : x_i \neq 0\}$, and the *off-support* is denoted by $I_0(\mathbf{x}) \equiv \{i \in \{1, \ldots, n\} : x_i = 0\}$. A vector is said to have a *full support* with respect to a given sparsity level s if $\|\mathbf{x}\|_0 = s$ and a *non-full support* if $\|\mathbf{x}\|_0 < s$. Given a vector $\mathbf{x} \in \mathbb{R}^n$, the subvector of $\mathbf{x}$ composed of the components of $\mathbf{x}$ whose indices are in a given subset $T \subseteq \{1, \ldots, n\}$ is denoted by $\mathbf{x}_T \in \mathbb{R}^{|T|}$. The matrix $\mathbf{U}_T$ denotes the submatrix of the n-dimensional identity matrix $\mathbf{I}_n$ constructed from the columns corresponding to the index set T. Obviously, if $T = \{i : x_i \neq 0\}$, then $\mathbf{x} = \mathbf{U}_T\mathbf{x}_T$. Given a set $B \subseteq \mathbb{R}^n$ and a set of indices $T \subseteq \{1, 2, \ldots, n\}$, the *restriction of the set* B *to the index set* T is defined by $B_T = \{\mathbf{x} \in \mathbb{R}^{|T|} : \mathbf{U}_T\mathbf{x} \in B\}$ whenever $T \neq \emptyset$, and in case of an empty index set as

$$B_\emptyset = \begin{cases} \{\mathbf{0}\}, & \mathbf{0} \in B, \\ \emptyset, & \mathbf{0} \notin B. \end{cases}$$

2. Mathematical Preliminaries.

2.1. Permutations and Orders. The permutation group comprising all possible $n!$ permutations of $\{1, \ldots, n\}$ is denoted by Σ_n. Given a vector $\mathbf{x} \in \mathbb{R}^n$, the vector $\mathbf{x}^\sigma$ is the vector defined by

$$(\mathbf{x}^\sigma)_i = x_{\sigma(i)},$$

which is the reordering of $\mathbf{x}$ according to the permutation σ. For example, if $\mathbf{x} = (4,\ 5,\ 6)^T$, and σ is the permutation given by $\sigma(1) = 3, \sigma(2) = 2, \sigma(3) = 1$, then $\mathbf{x}^\sigma = (6, 5, 4)^T$, meaning that σ reorders the elements of $\mathbf{x}$ in a non-ascending order. Such permutations will be called *sorting permutations* and their formal definition follows.

Definition 2.1 (sorting permutations). *Let $\mathbf{x} \in \mathbb{R}^n$. Then a permutation which sorts the elements of $\mathbf{x}$ in a non-ascending order will be called a* ***sorting permutation*** *of $\mathbf{x}$. The set of all sorting permutations of $\mathbf{x}$ is a subset of Σ_n and is denoted by $\tilde{\Sigma}(\mathbf{x})$. Explicitly,*

$$\tilde{\Sigma}(\mathbf{x}) = \left\{\sigma \in \Sigma_n : x_{\sigma(1)} \geq x_{\sigma(2)} \geq \cdots \geq x_{\sigma(n-1)} \geq x_{\sigma(n)}\right\}.$$

Given $\sigma \in \Sigma_n$, the set S_j^σ is the set comprising the indices $\sigma(1), \sigma(2), \ldots, \sigma(j)$:

$$S_j^\sigma = \begin{cases} \{\sigma(1), \sigma(2), \ldots, \sigma(j)\}, & 1 \leq j \leq n, \\ \emptyset. & \text{otherwise.} \end{cases}$$

2.2. Symmetric Functions and Sets. Symmetry will play a vital role in the analysis to follow. We begin with the definition of symmetric functions.

Definition 2.2 (symmetric functions). *A function $h : \mathbb{R}^n \to [-\infty, \infty]$ is called* ***symmetric*** *if $h(\mathbf{x}) = h(\mathbf{x}^\sigma)$ for any $\mathbf{x} \in \mathbb{R}^n$ and $\sigma \in \Sigma_n$.*

In a similar way, we define symmetric sets.

Definition 2.3 (symmetric sets). *Let $D \subseteq \mathbb{R}^n$. Then D is a* ***symmetric set*** *if for any vector $\mathbf{x} \in D$ and $\sigma \in \Sigma_n$, it holds that $\mathbf{x}^\sigma \in D$.*

When a symmetric function is also invariant under sign changes, it is called an *absolutely symmetric function.*

Definition 2.4 (absolutely symmetric functions). *A function $h : \mathbb{R}^n \to [-\infty, \infty]$ is called* ***absolutely symmetric*** *if for any $\mathbf{x} \in \mathbb{R}^n$ and $\sigma \in \Sigma_n$, it holds that $h(\mathbf{x}) = h(|\mathbf{x}^\sigma|)$.*

Example 2.5. *The ℓ_1-norm function $h(\mathbf{z}) = \sum_{i=1}^n |z_i|$ is absolutely symmetric, while the sum function $h(\mathbf{z}) = \sum_{i=1}^n z_i$ is symmetric, but not absolutely symmetric.*

Correspondingly, sets that are symmetric under sign changes and permutations will be called *absolutely symmetric.*

Definition 2.6 (absolutely symmetric sets). *Let $D \subseteq \mathbb{R}^n$ be a symmetric set. Then D is an* ***absolutely symmetric set*** *if it satisfies that $\mathbf{x} \in D$ if and only if $|\mathbf{x}| \in D$.*

Obviously, $D \subseteq \mathbb{R}^n$ is a symmetric set if and only if δ_D is a symmetric function, and is an absolutely symmetric set if and only if δ_D is an absolutely symmetric function. Much of the properties of symmetric and absolutely symmetric functions were studied in the seminal papers of Lewis in [20, 21] and in the book of Lewis and Borwein [10, Section 5.2, etc.]. We note that the terminology used in this manuscript differs from the one used in [3] where "symmetric" and "absolutely symmetric" sets were referred to as "type-1" and "type-2" symmetric sets respectively.

Another important property of sets is non-negativity.

Definition 2.7 (nonnegative sets). *A set $D \subseteq \mathbb{R}^n$ is* ***nonnegative*** *if $\mathbf{x} \geq \mathbf{0}$ for any $\mathbf{x} \in D$.*

Examples of nonnegative symmetric sets are the nonnegative orthant, the unit-simplex $\Delta_n \equiv \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x} \geq \mathbf{0}, \mathbf{e}^T\mathbf{x} = 1\}$, and nonnegative boxes of the form $[l, u]^n$ $(u \geq l \geq 0)$. Examples of absolutely symmetric sets are ℓ_p-balls $(p > 0)$, $\mathbb{R}^n$, as well as any box set of the form $[-l, l]^n$ for some $l \geq 0$.

2.3. Review of Results for the Projection onto Sparse Symmetric Sets.

Let $\mathbf{x} \in \mathbb{R}^n$, $s \in \{1, 2, \ldots, n\}$ and let $B \subseteq \mathbb{R}^n$ be a nonempty closed and convex set. A vector in $P_{C_s \cap B}(\mathbf{x})$ is called an *s-sparse projection vector of* $\mathbf{x}$ *onto* B. Since C_s is a nonconvex set, computing a member in $P_{C_s \cap B}(\mathbf{x})$ is in general a hard task; however, in [3] it was shown that whenever B is either nonnegative symmetric or absolutely symmetric, it is possible to find a sparse projection vector onto B efficiently.[1]

We will mostly focus on sets that are *simple symmetric sets*, a term that we define explicitly below.

Definition 2.8 (simple symmetric sets). *A set $B \subseteq \mathbb{R}^n$ is called* **a simple symmetric set** *if (i) it is either nonnegative symmetric or absolutely symmetric and (ii) it is nonempty closed and convex.*

To simplify and unify the analysis, in cases where B is either an absolutely symmetric or a nonnegative symmetric set, we will use the following *symmetry function* $p_B : \mathbb{R}^n \to \mathbb{R}^n$ that distinguishes between the two types of possible symmetries B might posses:

$$p_B(\mathbf{x}) \equiv \begin{cases} \mathbf{x}, & B \text{ is nonnegative symmetric,} \\ |\mathbf{x}|, & B \text{ is absolutely symmetric.} \end{cases} \tag{2.1}$$

The following theorem recalls how to compute a sparse projection vector onto simple symmetric sets.

Theorem 2.9 ([3, Theorem 4.4]). *Suppose that $B \subseteq \mathbb{R}^n$ is a simple symmetric set. Let $s \in \{1, 2, \ldots, n\}$, $\mathbf{x} \in \mathbb{R}^n$ and $\sigma \in \tilde{\Sigma}(p_B(\mathbf{x}))$. Then $\mathbf{U}_T P_{B_T}(\mathbf{x}_T) \in P_{C_s \cap B}(\mathbf{x})$, where $T = S_s^\sigma$.*

Theorem 2.9 essentially states that a member in $P_{C_s \cap B}(\mathbf{x})$ can be constructed by projecting the s largest components of $\mathbf{x}$ (in value or in absolute value depending on the type of symmetry) onto the restriction of B to these components and plugging zeros elsewhere. Note that the choice of the s largest components is not necessarily unique, and therefore there can be more than one member in $P_{C_s \cap B}(\mathbf{x})$.

The following two properties of the orthogonal projection onto symmetric sets from [3] will be useful in our analysis.

Lemma 2.10 (properties of projection onto symmetric sets). *Let $D \subseteq \mathbb{R}^n$ be a symmetric set, $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathrm{P}_D(\mathbf{x})$. Then*

(a) for any permutation $\sigma \in \Sigma_n$ and any $i, j \in \{1, 2, \ldots, n\}$, it holds that $(y_i - y_j)(x_i - x_j) \geq 0$; ([3, Lemma 3.1])

(b) if D is absolutely symmetric, then $|\mathbf{y}| \in P_{D \cap \mathbb{R}^n_+}(|\mathbf{x}|)$; ([3, Corollary 3.1])

(c) if D is absolutely symmetric, then sign$(\mathbf{x}) \odot \mathbf{y} \in P_{D \cap \mathbb{R}^n_+}(|\mathbf{x}|)$. ([3, Lemma 3.3])

We will now derive a key result that states that if $\mathbf{y} \in P_D(\mathbf{x})$ for some symmetric set D, then $P_D(\mathbf{x})$ and $\mathbf{y}$ can be ordered simultaneously.

[1] Given that it is possible to efficiently find orthogonal projections onto nonempty closed convex sets of the form B_T (T being an index set).

THEOREM 2.11. *Suppose that $D \subseteq \mathbb{R}^n$ is either an absolutely symmetric set or a nonnegative symmetric set. Let $\mathbf{x} \in \mathbb{R}^n$. Then for any $\mathbf{y} \in P_D(\mathbf{x})$ there exists*

$$\sigma \in \tilde{\Sigma}(p_D(\mathbf{x})) \cap \tilde{\Sigma}(p_D(\mathbf{y})). \tag{2.2}$$

Proof. We will begin by showing that for any $\mathbf{y} \in P_D(\mathbf{x})$ there exists $\sigma \in \tilde{\Sigma}(\mathbf{y}) \cap \tilde{\Sigma}(\mathbf{x})$. Let $\mathbf{y} \in P_D(\mathbf{x})$, and $\sigma \in \tilde{\Sigma}(\mathbf{y})$; suppose that $\sigma \notin \tilde{\Sigma}(\mathbf{x})$. Then there exist indices $i_1 < i_2$ such that $x_{\sigma(i_1)} < x_{\sigma(i_2)}$. By Lemma 2.10(a), $(y_{\sigma(i_1)} - y_{\sigma(i_2)})(x_{\sigma(i_1)} - x_{\sigma(i_2)}) \geq 0$, which implies that $y_{\sigma(i_1)} \leq y_{\sigma(i_2)}$. Since $\sigma \in \tilde{\Sigma}(\mathbf{y})$, it follows that $y_{\sigma(i_1)} \geq y_{\sigma(i_2)}$, and hence $y_{\sigma(i_1)} = y_{\sigma(i_2)}$. Therefore, the permutation $\hat{\sigma}$ defined by $\hat{\sigma}(i_1) = i_2, \hat{\sigma}(i_2) = i_1$ and $\hat{\sigma}(k) = \sigma(k)$ otherwise, is also a sorting permutation of $\mathbf{y}$; we then set $\sigma \leftarrow \hat{\sigma}$. This procedure can be repeated as long as there are indices $i < j$ which violate the order $(x_{\sigma(i)} < x_{\sigma(j)})$. Since at each iteration of the procedure, the number of pairs of indices which violate the order of $\mathbf{x}$ is strictly reduced, the process is finite and ends with a sorting permutation $\sigma \in \tilde{\Sigma}(\mathbf{y})$ satisfying $\sigma \in \tilde{\Sigma}(\mathbf{x})$.

Let $\sigma \in \tilde{\Sigma}(\mathbf{y}) \cap \tilde{\Sigma}(\mathbf{x})$. If D is a nonnegative symmetric set, then (2.2) is trivially satisfied. If D is an absolutely symmetric set, then by Lemma 2.10(b) we have that $|\mathbf{y}| \in P_{D \cap \mathbb{R}^n_+}(|\mathbf{x}|)$. Hence, by the first part of this proof, there exists $\sigma \in \tilde{\Sigma}(|\mathbf{y}|) \cap \tilde{\Sigma}(|\mathbf{x}|)$, and the required is satisfied. □

3. Sparse Prox Vectors.

In this section we will focus on studying the properties, as well as computation methods, of vectors in $\operatorname{prox}_{\alpha g_B}$, where $\alpha > 0$ and $g_B(\mathbf{x}) \equiv \delta_B(\mathbf{x}) + \|\mathbf{x}\|_0$ with B being a simple symmetric set. Recall that by the definition of the prox operator

$$\operatorname{prox}_{\alpha g_B}(\mathbf{x}) = \operatorname*{argmin}_{\mathbf{u} \in B} \left\{ \alpha \|\mathbf{u}\|_0 + \frac{1}{2}\|\mathbf{u} - \mathbf{x}\|_2^2 \right\}.$$

An optimal solution of the above problem will be called a *sparse prox vector over B with constant α*, and in most cases we will just refer to the vector as a *sparse prox vector over B* without indication of the identity of the constant α. The results of this section will later be used in Section 4 to develop optimality conditions for problem (P) and in Section 5 in the developments of corresponding algorithms.

A fundamental tool in analyzing and finding sparse prox vectors when the underlying set is simple symmetric is the sparse projection sequence.

3.1. The Sparse Projection Sequence.

DEFINITION 3.1 (i-sparse projection). *Let $B \subseteq \mathbb{R}^n$ be a simple symmetric set and let $\mathbf{x} \in \mathbb{R}^n, \sigma \in \tilde{\Sigma}(p_B(\mathbf{x}))$. Then for any $i \in \{1, \ldots, n\}$, the* ***i-sparse projection on B with respect to σ****, denoted by $P_B^\sigma(\mathbf{x}; i)$, is defined by*

$$P_B^\sigma(\mathbf{x}; i) = \mathbf{U}_T P_{B_T}(\mathbf{x}_T) \text{ where } T = S_i^\sigma.$$

When $\mathbf{0} \in B$, we will artificially define $P_B^\sigma(\mathbf{x}; 0) = \mathbf{0}$. The following binary variable will be used to indicate whether $\mathbf{0}$ is in B or not:

$$\ell_B \equiv \begin{cases} 0, & \mathbf{0} \in B, \\ 1, & \mathbf{0} \notin B. \end{cases}$$

When B is nonempty and absolutely symmetric, then by its convexity, $\mathbf{0} \in B$, and hence $\ell_B = 0$. Therefore, the case $\ell_B = 1$ is only relevant for some examples of nonnegative symmetric sets (such as the unit simplex). The sequence of all the i-sparse projections w.r.t. the same sorting permutation $\sigma \in \tilde{\Sigma}(p_B(\mathbf{x}))$,

$$\{P_B^\sigma(\mathbf{x}; i)\}_{i=\ell_B}^n$$

will be called the *sparse projection sequence on B with respect to σ.*

REMARK 3.2. *By Theorem 2.9, for any $i \in \{\ell_B, \ell_B+1, \dots, n\}$, $P_B^\sigma(\mathbf{x}; i) \in P_{C_i \cap B}(\mathbf{x})$.*

The next lemma states that for any sparse projection vector there exists a sorting permutation and a corresponding sparse projection sequence to which it belongs.

LEMMA 3.3. *Let $B \subseteq \mathbb{R}^n$ be a simple symmetric set, and let $\mathbf{y} \in P_{C_i \cap B}(\mathbf{x})$ where $i = \|\mathbf{y}\|_0$. Then there exists $\sigma \in \tilde{\Sigma}(p_B(\mathbf{x})) \cap \tilde{\Sigma}(p_B(\mathbf{y}))$ for which $\mathbf{y} = P_B^\sigma(\mathbf{x}; i)$.*

Proof. Since $\mathbf{y} \in P_{C_i \cap B}(\mathbf{x})$, it obviously satisfies $\mathbf{y}_{I_1(\mathbf{y})} = P_{B_{I_1(\mathbf{y})}}(\mathbf{x}_{I_1(\mathbf{y})})$, and therefore $\mathbf{y} = \mathbf{U}_T P_{B_T}(\mathbf{x}_T)$ where $T = I_1(\mathbf{y})$. By Theorem 2.11, since $\mathbf{y} \in P_{C_i \cap B}(\mathbf{x})$, there exists a sorting permutation $\sigma \in \tilde{\Sigma}(p_B(\mathbf{x})) \cap \tilde{\Sigma}(p_B(\mathbf{y}))$. For this permutation, $T = I_1(\mathbf{y}) = S_i^\sigma$, and hence $P_B^\sigma(\mathbf{x}; i) = \mathbf{U}_T P_{B_T}(\mathbf{x}_T) = \mathbf{y}$. ∎

The following technical lemma will be required.

LEMMA 3.4. *Suppose that $B \subseteq \mathbb{R}^n$ is a simple symmetric set. Let $\mathbf{x} \in \mathbb{R}^n$, $i \in \{1, \dots, n\}$, and $\sigma \in \tilde{\Sigma}(p_B(\mathbf{x}))$. Let i, j be two integers satisfying $\ell_B \le j \le i \le n$. If*

$$\|\mathbf{x} - P_B^\sigma(\mathbf{x}; j)\|_2^2 = \|\mathbf{x} - P_B^\sigma(\mathbf{x}; i)\|_2^2, \tag{3.1}$$

then $P_B^\sigma(\mathbf{x}; j) = P_B^\sigma(\mathbf{x}; i)$.

Proof. By splitting the sums in each of the sides of (3.1) to indices in S_i^σ and in $(S_i^\sigma)^c$, we obtain[2]

$$\|P_B^\sigma(\mathbf{x}; i)_{S_i^\sigma} - \mathbf{x}_{S_i^\sigma}\|_2^2 + \|\mathbf{x}_{(S_i^\sigma)^c}\|_2^2 = \|P_B^\sigma(\mathbf{x}; j)_{S_i^\sigma} - \mathbf{x}_{S_i^\sigma}\|_2^2 + \|\mathbf{x}_{(S_i^\sigma)^c}\|_2^2,$$

and hence,

$$\|P_B^\sigma(\mathbf{x}; i)_{S_i^\sigma} - \mathbf{x}_{S_i^\sigma}\|_2^2 = \|P_B^\sigma(\mathbf{x}; j)_{S_i^\sigma} - \mathbf{x}_{S_i^\sigma}\|_2^2. \tag{3.2}$$

Note that $P_B^\sigma(\mathbf{x}; i)_{S_i^\sigma}$ is the orthogonal projection of $\mathbf{x}_{S_i^\sigma}$ onto $B_{S_i^\sigma}$ and that $P_B^\sigma(\mathbf{x}; j)_{S_i^\sigma} \in B_{S_i^\sigma}$. Therefore, by the uniqueness of the orthogonal projection of $\mathbf{x}_{S_i^\sigma}$ onto the nonempty closed and convex set $B_{S_i^\sigma}$, (3.2) implies that $P_B^\sigma(\mathbf{x}; j)_{S_i^\sigma} = P_B^\sigma(\mathbf{x}; i)_{S_i^\sigma}$, and hence, also that $P_B^\sigma(\mathbf{x}; j) = P_B^\sigma(\mathbf{x}; i)$. ∎

We can utilize Lemma 3.4 and obtain the following result stating that if the projection sequence comprises different vectors up to a certain point, then these vectors must have full support.

THEOREM 3.5. *Suppose that $B \subseteq \mathbb{R}^n$ is a simple symmetric set. Let $\mathbf{x} \in \mathbb{R}^n$, $i \in \{1, \dots, n\}$, and $\sigma \in \tilde{\Sigma}(p_B(\mathbf{x}))$. Let $k \in \{1, 2, \dots, n\}$. If $P_B^\sigma(\mathbf{x}; i) \neq P_B^\sigma(\mathbf{x}; i+1)$ for any $i = \ell_B, \ell_B+1, \dots, k-1$, then $\|P_B^\sigma(\mathbf{x}; i)\|_0 = i$ for any $i = \ell_B, \ell_B+1, \dots, k$.*

Proof. We will prove that $\|P_B^\sigma(\mathbf{x}; i)\|_0 = i$ by induction on i. For $i = \ell_B$, it is obvious that $\|P_B^\sigma(\mathbf{x}; \ell_B)\|_0 = \ell_B$. Now we assume that $\|P_B^\sigma(\mathbf{x}; j)\|_0 = j$ for all $j = \ell_B, \ell_B+1, \dots, i-1$ where $i \in \{\ell_B+1, \ell_B+2, \dots, k\}$, and will prove that $\|P_B^\sigma(\mathbf{x}; i)\|_0 = i$. Suppose by contradiction that $\|P_B^\sigma(\mathbf{x}; i)\|_0 \le i-1$. This, along with the fact that $P_B^\sigma(\mathbf{x}; i-1) \in \underset{\mathbf{y} \in C_{i-1} \cap B}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{x}\|_2^2$ implies that

$$\|P_B^\sigma(\mathbf{x}; i-1) - \mathbf{x}\|_2^2 \le \|P_B^\sigma(\mathbf{x}; i) - \mathbf{x}\|_2^2.$$

Since $P_B^\sigma(\mathbf{x}; i) \in \underset{\mathbf{y} \in C_i \cap B}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{x}\|_2^2$,

$$\|P_B^\sigma(\mathbf{x}; i) - \mathbf{x}\|_2^2 \le \|P_B^\sigma(\mathbf{x}; i-1) - \mathbf{x}\|_2^2.$$

[2] When $i = n$, the term $\|\mathbf{x}_{(S_i^\sigma)^c}\|_2^2$ is omitted from both sides.

Thus, $\|P_B^\sigma(\mathbf{x};i)-\mathbf{x}\|_2^2 = \|P_B^\sigma(\mathbf{x};i-1)-\mathbf{x}\|_2^2$. Invoking Lemma 3.4, we obtain that $P_B^\sigma(\mathbf{x};i-1)=P_B^\sigma(\mathbf{x};i)$, which is a contradiction to the underlying assumption that the projection sequence (up to the kth vector) comprises different vectors. □

3.2. Computing Sparse Prox Vectors. In this section we will show how to find a sparse prox vector over B under the assumption that B is a simple symmetric set. In particular, Theorem 3.6 below shows that a sparse prox vector can always be found among the sparse projection sequence (w.r.t. an arbitrary sorting permutation). We begin by defining the set of indices

$$D_\alpha(\mathbf{x}) = \underset{i\in\{\ell_B,\ell_B+1,\ldots,n\}}{\operatorname{argmin}} \left\{\alpha\|P_B^\sigma(\mathbf{x};i)\|_0 + \frac{1}{2}\|P_B^\sigma(\mathbf{x};i)-\mathbf{x}\|_2^2\right\}.$$

THEOREM 3.6 (sparse prox characterization). *Suppose that B is a simple symmetric set. Let $\mathbf{x}\in\mathbb{R}^n, \sigma\in\tilde{\Sigma}(p_B(\mathbf{x}))$ and $\alpha>0$. Then $P_B^\sigma(\mathbf{x};m)\in\operatorname{prox}_{\alpha g_B}(\mathbf{x})$ for any $m\in D_\alpha(\mathbf{x})$.*

Proof. Since $\operatorname{prox}_{\alpha g_B}(\mathbf{x})$ is nonempty, there exists

$$\mathbf{z}\in\operatorname{prox}_{\alpha g_B}(\mathbf{x}) = \underset{\mathbf{u}\in B}{\operatorname{argmin}}\left\{\alpha\|\mathbf{u}\|_0+\frac{1}{2}\|\mathbf{u}-\mathbf{x}\|_2^2\right\}. \tag{3.3}$$

Since $m\in D_\alpha(\mathbf{x})$,

$$\alpha\|P_B^\sigma(\mathbf{x};m)\|_0+\frac{1}{2}\|P_B^\sigma(\mathbf{x};m)-\mathbf{x}\|_2^2 \le \alpha\|P_B^\sigma(\mathbf{x};\|\mathbf{z}\|_0)\|_0+\frac{1}{2}\|P_B^\sigma(\mathbf{x};\|\mathbf{z}\|_0)-\mathbf{x}\|_2^2. \tag{3.4}$$

By the definition of the projection sequence (see also Remark 3.2), $P_B^\sigma(\mathbf{x};\|\mathbf{z}\|_0)\in P_{C_{\|\mathbf{z}\|_0}\cap B}(\mathbf{x})$, and hence, $\|P_B^\sigma(\mathbf{x};\|\mathbf{z}\|_0)-\mathbf{x}\|_2^2\le\|\mathbf{z}-\mathbf{x}\|_2^2$, which combined with the obvious inequality $\|P_B^\sigma(\mathbf{x};\|\mathbf{z}\|_0)\|_0\le\|\mathbf{z}\|_0$, yields

$$\alpha\|P_B^\sigma(\mathbf{x};\|\mathbf{z}\|_0)\|_0+\frac{1}{2}\|P_B^\sigma(\mathbf{x};\|\mathbf{z}\|_0)-\mathbf{x}\|_2^2\le\alpha\|\mathbf{z}\|_0+\frac{1}{2}\|\mathbf{z}-\mathbf{x}\|_2^2. \tag{3.5}$$

Combining (3.4) and (3.5), we have that $\alpha\|P_B^\sigma(\mathbf{x};m)\|_0+\frac{1}{2}\|P_B^\sigma(\mathbf{x};m)-\mathbf{x}\|_2^2\le\alpha\|\mathbf{z}\|_0+\frac{1}{2}\|\mathbf{z}-\mathbf{x}\|_2^2$, which by the definition of $\mathbf{z}$ (see (3.3)) implies that $P_B^\sigma(\mathbf{x};m)\in\operatorname{prox}_{\alpha g_B}(\mathbf{x})$. □

Theorem 3.6 states that essentially in order to find a sparse prox vector, we need to minimize the discrete function

$$V_{B,\sigma,\mathbf{x},\alpha}: i\mapsto\alpha\|P_B(\mathbf{x};i)\|_0+\frac{1}{2}\|\mathbf{x}-P_B^\sigma(\mathbf{x};i)\|_2^2.$$

When the identities of $B,\sigma,\mathbf{x}$ and α will be clear from the context, we will use the notation V instead of $V_{B,\sigma,\mathbf{x},\alpha}$. In this notation, $D_\alpha(\mathbf{x})=\underset{i=\ell_B,\ell_B+1,\ldots,n}{\operatorname{argmin}} V(i)$.

Theorem 3.6 naturally leads to the following algorithm for computing a sparse prox vector over simple symmetric sets.

3.3. The Second Order Monotonicity Property.

3.3.1. Definition and Basic Properties. Algorithm 1 requires the computation of n sparse projection vectors in order to compute a sparse prox vector. We will now show how this search can be done more efficiently when the set B, in addition to being simple symmetric, satisfies a submodularity-like monotonicity property that we will refer to as the *second order monotonicity property.*

DEFINITION 3.7 (second order monotonicity). *A simple symmetric set $B\subseteq\mathbb{R}^n$ is said to satisfy the* ***second order monotonicity*** *(SOM) property if for any $\mathbf{x}\in\mathbb{R}^n$, $i\in\{\ell_B,\ell_B+1,\ldots,n-2\}$ and $\sigma\in$*

Algorithm 1 Computing a Sparse Prox Vector

```
Input: $\mathbf{x} \in \mathbb{R}^n$, $\alpha > 0$, $B \subseteq \mathbb{R}^n$ simple symmetric.
Output: $\mathbf{y} \in \operatorname{prox}_{\alpha g_B}(\mathbf{x})$.
    1. Find $\sigma \in \tilde{\Sigma}(p_B(\mathbf{x}))$.
    2. Compute $P_B^\sigma(\mathbf{x}; i)$ for any $i \in \{\ell_B, \ell_B + 1, \dots, n\}$.
    3. Set $\mathbf{y} = P_B(\mathbf{x}; m)$, where $m \in \underset{i \in \{\ell_B, \ell_B+1, \dots, n\}}{\operatorname{argmin}} \left\{ \alpha \|P_B^\sigma(\mathbf{x}; i)\|_0 + \frac{1}{2}\|P_B^\sigma(\mathbf{x}; i) - \mathbf{x}\|_2^2 \right\}$.
```

$\tilde{\Sigma}(p_B(\mathbf{x}))$, *the following inequality holds:*

$$\|P_B^\sigma(\mathbf{x}; i) - \mathbf{x}\|_2^2 - \|P_B^\sigma(\mathbf{x}; i+1) - \mathbf{x}\|_2^2 \geq \|P_B^\sigma(\mathbf{x}; i+1) - \mathbf{x}\|_2^2 - \|P_B^\sigma(\mathbf{x}; i+2) - \mathbf{x}\|_2^2. \tag{3.6}$$

In Appendix A we prove that many important sets share this property, such as $\mathbb{R}^n$, $\mathbb{R}^n_+$, ℓ_1, ℓ_2, ℓ_∞-balls and the α-simplex set.
For any $i \in \{\ell_B, \ell_B + 1, \dots, n-1\}$ and $\mathbf{x} \in \mathbb{R}^n$ we use the notation

$$d_B^\sigma(\mathbf{x}; i) \equiv \frac{1}{2}\|P_B^\sigma(\mathbf{x}; i) - \mathbf{x}\|_2^2 - \frac{1}{2}\|P_B^\sigma(\mathbf{x}; i+1) - \mathbf{x}\|_2^2.$$

Obviously $d_B^\sigma(\mathbf{x}; j) \geq 0$ for any $\mathbf{x} \in \mathbb{R}^n$ and $j \in \{\ell_B, \ell_B + 1, \dots, n-1\}$. The SOM property can be written in terms of this notation as

$$d_B^\sigma(\mathbf{x}; \ell_B) \geq d_B^\sigma(\mathbf{x}; \ell_B + 1) \geq \cdots \geq d_B^\sigma(\mathbf{x}; n-1).$$

In the rest of this section we will show that the above monotonicity property of the sequence $\{d_B^\sigma(\mathbf{x}; i)\}_{i=\ell_B}^{n-1}$ implies that an index in $D_\alpha(\mathbf{x})$ can be attained by performing a binary search, thus reducing the amount of sparse projections that need to be computed from $O(n)$ to $O(\log_2(n))$.

The largest index for which $d_B^\sigma(\mathbf{x}; i-1) > 0$ will be denoted by $u_{B,\sigma,\mathbf{x}}$:

$$u_{B,\sigma,\mathbf{x}} = \begin{cases} \max\{i \in \{\ell_B + 1, \ell_B + 2, \dots, n\} : d_B^\sigma(\mathbf{x}; i-1) > 0\}, & d_B^\sigma(\mathbf{x}; \ell_B) > 0, \\ \ell_B, & d_B^\sigma(\mathbf{x}; \ell_B) = 0. \end{cases}$$

The next lemma establishes some basic properties of the sparse projection sequence under the SOM property that will be useful in our analysis.

Lemma 3.8. *Let $\mathbf{x} \in \mathbb{R}^n$, and suppose that B is a simple symmetric set satisfying the SOM property. Then*

(a) $\|P_B^\sigma(\mathbf{x}; i)\|_0 = i$ for any $i \in \{\ell_B, \ell_B + 1, \dots, u_{B,\sigma,\mathbf{x}}\}$;

(b) $\|P_B^\sigma(\mathbf{x}; i)\|_0 = u_{B,\sigma,\mathbf{x}}$ for any $i \in \{u_{B,\sigma,\mathbf{x}}, u_{B,\sigma,\mathbf{x}} + 1, \dots, n\}$. Moreover, for any such i, $P_B^\sigma(\mathbf{x}; i) = P_B^\sigma(\mathbf{x}; u_{B,\sigma,\mathbf{x}})$.

Proof. Claim (a) is obvious for the case where $u_{B,\sigma,\mathbf{x}} = \ell_B$. Suppose that $u_{B,\sigma,\mathbf{x}} > \ell_B$. By the definition of $u_{B,\sigma,\mathbf{x}}$, it follows that $d_B^\sigma(\mathbf{x}; i-1) > 0$ for any $i \in \{\ell_B + 1, \ell_B + 2, \dots, u_{B,\sigma,\mathbf{x}}\}$. In particular for any such i, $P_B^\sigma(\mathbf{x}; i-1) \neq P_B^\sigma(\mathbf{x}; i)$, and thus part (a) follows by invoking Theorem 3.5. To prove claim (b), note that by the definition of $u_{B,\sigma,\mathbf{x}}$ and the SOM property $d_B^\sigma(\mathbf{x}; i) = 0$ for any $i \in \{u_{B,\sigma,\mathbf{x}}, u_{B,\sigma,\mathbf{x}} + 1, \dots, n-1\}$, meaning that for such i, $\|\mathbf{x} - P_B^\sigma(\mathbf{x}; i)\|_2^2 = \|\mathbf{x} - P_B^\sigma(\mathbf{x}; i+1)\|_2^2$. Thus, invoking Lemma 3.4 , it follows that $P_B^\sigma(\mathbf{x}; i) = P_B^\sigma(\mathbf{x}; u_{B,\sigma,\mathbf{x}})$ for any $i \in \{u_{B,\sigma,\mathbf{x}}, u_{B,\sigma,\mathbf{x}} + 1, \dots, n\}$, and in particular for such i, $\|P_B^\sigma(\mathbf{x}; i)\|_0 = \|P_B^\sigma(\mathbf{x}; u_{B,\sigma,\mathbf{x}})\|_0 = u_{B,\sigma,\mathbf{x}}$. □

The main result connecting d_B^σ and the value function V will now be stated and proved.

THEOREM 3.9. *Suppose that $B \subseteq \mathbb{R}^n$ is a simple symmetric set satisfying the SOM property. Let $\mathbf{x} \in \mathbb{R}^n$, $\sigma \in \tilde{\Sigma}(p_B(\mathbf{x}))$ and $\alpha > 0$. Denote $V \equiv V_{B,\sigma,\mathbf{x},\alpha}$.*

1. *For any $i \in \{\ell_B, \ell_B + 1, \dots, u_{B,\sigma,\mathbf{x}} - 1\}$, it holds that*
   (a) $d_B^\sigma(\mathbf{x}; i) > \alpha$ if and only if $V(i) > V(i+1)$.
   (b) $d_B^\sigma(\mathbf{x}; i) = \alpha$ if and only if $V(i) = V(i+1)$.
   (c) $d_B^\sigma(\mathbf{x}; i) < \alpha$ if and only if $V(i) < V(i+1)$.
2. *$V(u_{B,\sigma,\mathbf{x}}) = V(u_{B,\sigma,\mathbf{x}} + 1) = \cdots = V(n)$.*

Proof. 1. By Lemma 3.8, for any $i \in \{\ell_B, \ell_B + 1, \dots, u_{B,\sigma,\mathbf{x}} - 1\}$, $\|P_B^\sigma(\mathbf{x}; i)\|_0 = i$ and $\|P_B^\sigma(\mathbf{x}; i+1)\|_0 = i + 1$; thus,

$$V(i+1) - V(i) = \frac{1}{2}\|\mathbf{x} - P_B^\sigma(\mathbf{x}; i+1)\|_2^2 + \alpha(i+1) - \frac{1}{2}\|\mathbf{x} - P_B^\sigma(\mathbf{x}; i)\|_2^2 - \alpha i = \alpha - d_B^\sigma(\mathbf{x}; i),$$

from which (a),(b) and (c) follow.
2. By the definition of $u_{B,\sigma,\mathbf{x}}$ and the SOM property, $d_B^\sigma(\mathbf{x}; u_{B,\sigma,\mathbf{x}}) = d_B^\sigma(\mathbf{x}; u_{B,\sigma,\mathbf{x}}+1) = \cdots = d_B^\sigma(\mathbf{x}; n-1) = 0$, and hence $\|\mathbf{x} - P_B^\sigma(\mathbf{x}; i)\|_2^2 = \|\mathbf{x} - P_B^\sigma(\mathbf{x}; u_{B,\sigma,\mathbf{x}})\|_2^2$ for any $i \in \{u_{B,\sigma,\mathbf{x}}, U_{B,\sigma,\mathbf{x}} + 1, \dots, n\}$, which along with Lemma 3.8(b) implies that for any $i \in \{u_{B,\sigma,\mathbf{x}}, U_{B,\sigma,\mathbf{x}} + 1, \dots, n\}$

$$V(i) = \alpha\|P_B^\sigma(\mathbf{x}; i)\|_0 + \frac{1}{2}\|\mathbf{x} - P_B^\sigma(\mathbf{x}; i)\|_2^2 = \alpha\|P_B^\sigma(\mathbf{x}; u_{B,\sigma,\mathbf{x}})\|_0 + \frac{1}{2}\|\mathbf{x} - P_B^\sigma(\mathbf{x}; u_{B,\sigma,\mathbf{x}})\|_2^2 = V(u_{B,\sigma,\mathbf{x}}).$$

The connection between V and d_B^σ described in Theorem 3.9 implies that if $d_B^\sigma(\mathbf{x}; i) > \alpha$, then all the indices in $D_\alpha(\mathbf{x})$ are larger than i, and that if $d_B(\mathbf{x}; i) \le \alpha$, then $D_\alpha(\mathbf{x}) \cap \{\ell_B, \ell_B + 1, \dots, i\} \neq \emptyset$. This naturally implies that we can define a binary search procedure for finding an index in $D_\alpha(\mathbf{x})$. Specifically, the following procedure finds the smallest index in $D_\alpha(\mathbf{x})$, which can be explicitly written as

$$D_\alpha^{\min} = \begin{cases} \max\{i \in \{\ell_B, \ell_B + 1, \dots, n\} : d_B^\sigma(\mathbf{x}; i-1) > \alpha\}, & d_B^\sigma(\mathbf{x}; \ell_B) > \alpha, \\ \ell_B, & d_B^\sigma(\mathbf{x}; \ell_B) \le \alpha. \end{cases} \tag{3.7}$$

Algorithm 2 Binary Search for Computing a Sparse Prox Vector

Input: $\mathbf{x} \in \mathbb{R}^n$, $\alpha > 0$, $B \subseteq \mathbb{R}^n$ simple symmetric ($n \ge 3$).
Output: $\mathbf{y} \in \text{prox}_{\alpha g_B}(\mathbf{x})$.

```
1. Find $\sigma \in \tilde{\Sigma}(p_B(\mathbf{x}))$.
2. If $d_B^\sigma(\mathbf{x}; \ell_B) \le \alpha$, then $D_\alpha^{\min} = \ell_B$ and go to step 6.
3. If $d_B^\sigma(\mathbf{x}; n-1) > \alpha$, then $D_\alpha^{\min} = n$ and go to step 6.
4. Set $k_{\text{low}} = \ell_B, k_{\text{up}} = n - 1$.
5. Repeat:
   (a) if $k_{\text{up}} = k_{\text{low}} + 1$, then $D_\alpha^{\min} = k_{\text{up}}$ and go to step 6.
   (b) $k_{\text{mid}} = \lceil (k_{\text{up}} + k_{\text{low}})/2 \rceil$
       i. if $d_B(\mathbf{x}; k_{\text{mid}}) > \alpha$ set $k_{\text{low}} \leftarrow k_{\text{mid}}$;
       ii. otherwise set $k_{\text{up}} \leftarrow k_{\text{mid}}$.
6. Set $\mathbf{y} = P_B^\sigma\left(\mathbf{x}; D_\alpha^{\min}\right)$.
```

As was illustrated above, the validity of the SOM property enables a more efficient computation of a sparse prox vector. We will later show in Section 4 that the SOM property is fundamental in the study of optimality conditions of problem (1.1). It is thus important to detect simple symmetric sets that satisfy the property. Table 1 contains a list of subsets of $\mathbb{R}^n$ on which we will prove the validity of the SOM property. The constant α is assumed to be positive. The proofs are very long and technical, and since they are not essential for the developments in the sequel, they are postponed to Appendix A.

TABLE 1
List of sets satisfying the SOM property.

Name of Set	Set	Reference
ℓ_∞-ball	$B_\infty[0,\alpha]$	Theorem A.6
nonnegative α-box	$[0,\alpha]^n$	Theorem A.5
–	$\mathbb{R}^n$	Theorem A.6
nonnegative orthant	$\mathbb{R}^n_+$	Theorem A.5
ℓ_2-ball	$B_2[0,\alpha]$	Theorem A.7
α-simplex	$\Delta_n(\alpha) = \{\mathbf{x} : \mathbf{e}^T\mathbf{x} = \alpha, \mathbf{x} \geq \mathbf{0}\}$	Theorem A.13
full α-simplex	$\Delta^F_n(\alpha) = \{\mathbf{x} : \mathbf{e}^T\mathbf{x} \leq \alpha, \mathbf{x} \geq \mathbf{0}\}$	Theorem A.15
ℓ_1-ball	$B_1[0,\alpha]$	Theorem A.16

4. Optimality Conditions.

In this section we will discuss some key properties of minimizers of problem (1.1). For the sake of simplicitly, we will rewrite the problem as

$$\text{(P)} \qquad \min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) + \lambda g_B(\mathbf{x}),$$

where as usual $g_B(\mathbf{x}) = \delta_B(\mathbf{x}) + \|\mathbf{x}\|_0$. We will focus on three types of optimality conditions: *support optimality (SO)*, *L-stationarity*, and *coordinate-wise (CW) optimality*.

4.1. Preliminaries.

We make the following set of standing assumptions that will be assumed from now on.

ASSUMPTION 4.1. *(A) $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable over $\mathbb{R}^n$ and its gradient has a Lipschitz constant $L_f > 0$:*

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2 \leq L_f\|\mathbf{x} - \mathbf{y}\|_2 \text{ for all } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n.$$

(B) $B \subseteq \mathbb{R}^n$ is a simple sparse set.

(C) $\lambda > 0$.

The class of differentiable functions with Lipschitz gradient with constant L is denoted by $C^{1,1}_L$, so part (A) of Assumption 4.1 can also be written as $f \in C^{1,1}_{L_f}$. In our analysis we will frequently use the operator $T_L : \mathbb{R}^n \to \mathbb{R}^n$ denoting a gradient step at $\mathbf{y} \in \mathbb{R}^n$ with stepsize $\frac{1}{L}$:

$$T_L(\mathbf{y}) \equiv \mathbf{y} - \frac{1}{L}\nabla f(\mathbf{y}). \tag{4.1}$$

The analysis of the optimality conditions will be based on two well-known claims, the most fundamental is the so-called descent lemma (recall that Assumption 4.1 is a standing assumption).

LEMMA 4.2 (descent lemma [6]). *For any $\mathbf{x}, \mathbf{d} \in \mathbb{R}^n$ and $L \geq L_f$, it holds that $f(\mathbf{x} + \mathbf{d}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^T\mathbf{d} + \frac{L}{2}\|\mathbf{d}\|_2^2$.*

The *sufficient decrease lemma* for the proximal gradient mapping is given next.

Lemma 4.3 (sufficient decrease lemma [9, Lemma 3.2]). *Let $L > L_f$. Then for any $\mathbf{y} \in B$ and $\mathbf{z} \in \operatorname{prox}_{\frac{\lambda}{L} g_B}(T_L(\mathbf{y}))$, it holds that $f(\mathbf{y}) + \lambda g_B(\mathbf{y}) - f(\mathbf{z}) - \lambda g_B(\mathbf{z}) \geq \frac{L-L_f}{2} \|\mathbf{z} - \mathbf{y}\|^2$.*

We also recall some basic facts (see for example [6]) about stationarity in problems consisting of minimizing smooth functions over closed convex sets, meaning problems of the form

$$\text{(H)} \quad \min\{h(\mathbf{x}) : \mathbf{x} \in S\},$$

where $h : \mathbb{R}^n \to \mathbb{R}$ is differentiable and $S \subseteq \mathbb{R}^n$ is nonempty closed and convex. A point $\mathbf{x}^* \in S$ is called a *stationary point of (H)* if $\langle \nabla h(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle \geq 0$ for all $\mathbf{x} \in S$.

It is well-known that stationarity is a necessary local optimality condition for problem (H), and in case where f is convex, it is a necessary and sufficient global optimality condition. It is also known that for a given $L > 0$, a point $\mathbf{x}^* \in S$ is a stationary point of (H) if and only if

$$\mathbf{x}^* = P_S\left(\mathbf{x}^* - \frac{1}{L}\nabla h(\mathbf{x}^*)\right). \tag{4.2}$$

Combining this with the fact that the original definition of stationarity is independent of any parameter, we can conclude that (4.2) holds for a *specific* $L > 0$ if and only if it holds *for any* $L > 0$.

4.2. Support Optimality. We begin with the condition of optimality over the support.

Definition 4.4 (support optimality). *A vector $\mathbf{x} \in B$ is called a* ***support optimal (SO) point of (P)*** *if*

$$\mathbf{x} \in \underset{\mathbf{u} \in B}{\operatorname{argmin}} \left\{f(\mathbf{u}) : I_1(\mathbf{u}) \subseteq I_1(\mathbf{x})\right\}.$$

Remark 4.5. *By the definition of B_T for a given index set T, it follows that $\mathbf{x}$ is support optimal if and only if*

$$\mathbf{x}_{I_1(\mathbf{x})} \in \underset{\mathbf{d}}{\operatorname{argmin}}\{f(\mathbf{U}_{I_1(\mathbf{x})}\mathbf{d}) : \mathbf{d} \in B_{I_1}(\mathbf{x})\}.$$

The support optimality condition is a necessary optimality condition for problem (P).

Theorem 4.6 (optimality ⇒ SO). *Let $\mathbf{x}^*$ be an optimal solution of problem (P), then $\mathbf{x}^*$ is a support optimal point of (P).*

Proof. Let $\mathbf{z} \in B$ satisfy $I_1(\mathbf{z}) \subseteq I_1(\mathbf{x}^*)$. The latter condition implies that $\|\mathbf{z}\|_0 \leq \|\mathbf{x}^*\|_0$, which combined with the optimality of $\mathbf{x}^*$ yields

$$f(\mathbf{x}^*) + \lambda\|\mathbf{x}^*\|_0 \leq f(\mathbf{z}) + \lambda\|\mathbf{z}\|_0 \leq f(\mathbf{z}) + \lambda\|\mathbf{x}^*\|_0,$$

and hence $f(\mathbf{x}^*) \leq f(\mathbf{z})$, proving the support optimality of $\mathbf{x}^*$. □

The next lemma establishes two technical results. The first states that if a point is support optimal, then it is a stationary point over the support. The second result is only relevant when the underlying set is an absolutely symmetric set, and will be useful later on in the analysis.

Lemma 4.7. *Let $\mathbf{x} \in \mathbb{R}^n$ be a support optimal point of (P). Then* *(a) for any $L > 0$,*

$$\mathbf{x}_{I_1(\mathbf{x})} = P_{B_{I_1(\mathbf{x})}}\left(T_L(\mathbf{x})_{I_1(\mathbf{x})}\right); \tag{4.3}$$

(b) if B is absolutely symmetric, then $x_i \nabla_i f(\mathbf{x}) \leq 0$ for any $i \in I_1(\mathbf{x})$.

Proof. (a) By Remark 4.5, $\mathbf{x}_{I_1(\mathbf{x})}$ is an optimal solution of

$$\min_{\mathbf{d} \in B_{I_1(\mathbf{x})}} f(\mathbf{U}_{I_1(\mathbf{x})}\mathbf{d}), \tag{4.4}$$

and thus, $\mathbf{x}_{I_1(\mathbf{x})}$ is a stationary point of (4.4). That is, for any $L > 0$ we have that $\mathbf{x}_{I_1(\mathbf{x})}$ satisfies

$$\mathbf{x}_{I_1(\mathbf{x})} = P_{B_{I_1(\mathbf{x})}}\left(\mathbf{x}_{I_1(\mathbf{x})} - \frac{1}{L}\mathbf{U}_{I_1(\mathbf{x})}^T \nabla f(\mathbf{U}_{I_1(\mathbf{x})}\mathbf{x}_{I_1(\mathbf{x})})\right).$$

Since $\mathbf{U}_{I_1(\mathbf{x})}^T \nabla f(\mathbf{U}_{I_1(\mathbf{x})}\mathbf{x}_{I_1(\mathbf{x})}) = \nabla_{I_1(\mathbf{x})} f(\mathbf{x})$, $\mathbf{x}$ satisfies (4.3) for any $L > 0$.
(b) Let $i \in I_1(\mathbf{x})$. Since $\mathbf{x}_{I_1(\mathbf{x})}$ is a stationary point of (4.4), it follows that

$$\mathbf{U}_{I_1(\mathbf{x})}^T \nabla f(\mathbf{U}_{I_1(\mathbf{x})}\mathbf{x}_{I_1(\mathbf{x})})^T(\mathbf{y}_{I_1(\mathbf{x})} - \mathbf{x}_{I_1(\mathbf{x})}) \geq 0 \text{ for any } \mathbf{y} \in B \text{ s.t. } I_1(\mathbf{y}) \subseteq I_1(\mathbf{x}). \tag{4.5}$$ □

Since B is absolutely symmetric, the vector $\tilde{\mathbf{x}} = \mathbf{x} \odot (-\mathbf{e}_i)$ is also in B. Obviously, $I_1(\tilde{\mathbf{x}}) = I_1(\mathbf{x})$, and therefore we can plug $\mathbf{x} = \tilde{\mathbf{x}}$ into (4.5) and obtain the desired inequality $x_i \nabla_i f(\mathbf{x}) \leq 0$.

4.3. L-stationarity.

In Section 5.1, we will consider the proximal gradient method for solving problem (P), whose general update step is of the form $\mathbf{x}^{k+1} \in \text{prox}_{\frac{\lambda}{L} g_B}(T_L(\mathbf{x}))$. *$L$-stationary* points are defined as fixed points of this process.

DEFINITION 4.8 (L-stationarity). *Let $L > 0$. A vector $\mathbf{x} \in \mathbb{R}^n$ is called an **L-stationary point of (P)** if*

$$\mathbf{x} \in \text{prox}_{\frac{\lambda}{L} g_B}\left(T_L(\mathbf{x})\right). \tag{4.6}$$

Related optimality conditions expressed in terms of the orthogonal projection operator were considered in [2, 3] in the context of problems with sparsity constraints.

The next theorem shows that L-stationarity is a necessary optimality condition whenever $L > L_f$. Note that the condition is stronger than L-stationarity since it states that $\mathbf{x}^*$ is the only vector in $\text{prox}_{\frac{\lambda}{L} g_B}\left(T_L(\mathbf{x}^*)\right)$.

THEOREM 4.9 (optimality ⇒ L-stationarity). *Let $\mathbf{x}^*$ be an optimal solution of problem (P). Then for any $L > L_f$ it holds that $\{\mathbf{x}^*\} = \text{prox}_{\frac{\lambda}{L} g_B}\left(T_L(\mathbf{x}^*)\right)$.*

Proof. Let $L > L_f$, and let $\mathbf{z} \in \text{prox}_{\frac{\lambda}{L} g_B}(T_L(\mathbf{x}^*))$. Then by the sufficient decrease lemma (Lemma 4.3) and by the optimality of $\mathbf{x}^*$,

$$f(\mathbf{x}^*) + \lambda g_B(\mathbf{x}^*) \geq \frac{L - L_f}{2}\left\|\mathbf{z} - \mathbf{x}^*\right\|^2 + f(\mathbf{z}) + \lambda g_B(\mathbf{z}) \geq \frac{L - L_f}{2}\left\|\mathbf{z} - \mathbf{x}^*\right\|^2 + f(\mathbf{x}^*) + \lambda g_B(\mathbf{x}^*).$$

Since $L > L_f$, we conclude that $\mathbf{z} = \mathbf{x}^*$. □

The next result shows a relation between L-stationarity and support optimality – in case where f is convex, L-stationarity implies support optimality.

THEOREM 4.10 (L-stationarity $\Rightarrow$ SO optimality (f convex)). *Suppose that f is convex. Let $\mathbf{x} \in \mathbb{R}^n$ be an L-stationary point of (P) for some $L > 0$. Then $\mathbf{x}$ is a support optimal point of (P).*

Proof. Since (4.6) holds, it follows that

$$\begin{aligned}\frac{\lambda}{L}\|\mathbf{x}\|_0 + \frac{1}{2}\|\mathbf{x} - T_L(\mathbf{x})\|_2^2 &\le \min_{\mathbf{u}} \left\{ \frac{\lambda}{L}\|\mathbf{u}\|_0 + \frac{1}{2}\|\mathbf{u} - T_L(\mathbf{x})\|_2^2 : \mathbf{u} \in B \right\} \\ &\le \min_{\mathbf{u}} \left\{ \frac{\lambda}{L}\|\mathbf{u}\|_0 + \frac{1}{2}\|\mathbf{u} - T_L(\mathbf{x})\|_2^2 : \mathbf{u} \in B, I_1(\mathbf{u}) \subseteq I_1(\mathbf{x}) \right\} \\ &\le \min_{\mathbf{u}} \left\{ \frac{\lambda}{L}\|\mathbf{x}\|_0 + \frac{1}{2}\|\mathbf{u} - T_L(\mathbf{x})\|_2^2 : \mathbf{u} \in B, I_1(\mathbf{u}) \subseteq I_1(\mathbf{x}) \right\},\end{aligned}$$

and hence,

$$\|\mathbf{x} - T_L(\mathbf{x})\|_2^2 \le \min_{\mathbf{u}} \left\{ \|\mathbf{u} - T_L(\mathbf{x})\|_2^2 : \mathbf{u} \in B, I_1(\mathbf{u}) \subseteq I_1(\mathbf{x}) \right\} = \min_{\mathbf{d}} \left\{ \|\mathbf{U}_{I_1(\mathbf{x})}\mathbf{d} - T_L(\mathbf{x})\|_2^2 : \mathbf{d} \in B_{I_1(\mathbf{x})} \right\}.$$

Decomposing the expressions in both sides of the above inequality w.r.t. the two sets of indices $I_1(\mathbf{x})$ and $I_0(\mathbf{x})$, we obtain

$$\|\mathbf{x}_{I_1(\mathbf{x})} - T_L(\mathbf{x})_{I_1(\mathbf{x})}\|_2^2 + \|T_L(\mathbf{x})_{I_0(\mathbf{x})}\|_2^2 \le \min_{\mathbf{d}} \left\{ \|\mathbf{d} - T_L(\mathbf{x})_{I_1(\mathbf{x})}\|_2^2 : \mathbf{d} \in B_{I_1(\mathbf{x})} \right\} + \|T_L(\mathbf{x})_{I_0(\mathbf{x})}\|_2^2,$$

that is,

$$\|\mathbf{x}_{I_1(\mathbf{x})} - T_L(\mathbf{x})_{I_1(\mathbf{x})}\|_2^2 \le \min_{\mathbf{d}} \left\{ \|\mathbf{d} - T_L(\mathbf{x})_{I_1(\mathbf{x})}\|_2^2 : \mathbf{d} \in B_{I_1(\mathbf{x})} \right\},$$

meaning that $\mathbf{x}_{I_1(\mathbf{x})} = P_{B_{I_1(\mathbf{x})}}\left(T_L(\mathbf{x})_{I_1(\mathbf{x})}\right)$, which is precisely the condition that $\mathbf{x}_{I_1(\mathbf{x})}$ is a stationary point of the problem

$$\min\{f(\mathbf{U}_{I_1(\mathbf{x})}\mathbf{d}) : \mathbf{d} \in B_{I_1(\mathbf{x})}\}. \tag{4.7}$$

Since problem (4.7) is convex (by the convexity of f), it follows that $\mathbf{x}_{I_1(\mathbf{x})}$ is an optimal solution of (4.7), establishing the fact that it is a support optimal point. □

We will now state and prove a sufficient condition for L-stationarity that will play an important role in establishing the hierarchy between L-stationary and the condition that will be discussed in the next section.

LEMMA 4.11 (L-stationarity sufficient conditions). *Let $L \ge L_f$. Suppose that $\mathbf{x} = P_B^{\sigma}(T_L(\mathbf{x}); \|\mathbf{x}\|_0)$ for some $\sigma \in \tilde{\Sigma}(p_B(T_L(\mathbf{x})))$. Denote $\mathbf{v}^+ \equiv P_B^{\sigma}(T_L(\mathbf{x}); \|\mathbf{x}\|_0 + 1)$ and $\mathbf{v}^- \equiv P_B^{\sigma}(T_L(\mathbf{x}); \|\mathbf{x}\|_0 - 1)$. Then $\mathbf{x}$ is an L-stationary point of (P) if the following two conditions hold:*
(a) if $\|\mathbf{x}\|_0 \le n - 1$ then $f(\mathbf{x}) + \lambda\|\mathbf{x}\|_0 \le f(\mathbf{v}^+) + \lambda\|\mathbf{v}^+\|_0$;
(b) if $\|\mathbf{x}\|_0 \ge \ell_B + 1$ then $f(\mathbf{x}) + \lambda\|\mathbf{x}\|_0 \le f(\mathbf{v}^-) + \lambda\|\mathbf{v}^-\|_0$.

Proof. To simplify the exposition of the proof, we will make the convention that whenever $\mathbf{v}^+$ appears in an expression, then we assume that $\|\mathbf{x}\|_0 \le n-1$ and that each time $\mathbf{v}^-$ appears, it means that we assume that $\|\mathbf{x}\|_0 \ge \ell_B + 1$. For any $\mathbf{v} \in B$ satisfying $f(\mathbf{x}) + \lambda\|\mathbf{x}\|_0 \le f(\mathbf{v}) + \lambda\|\mathbf{v}\|_0$, the descent lemma implies that

$$f(\mathbf{x}) + \lambda\|\mathbf{x}\|_0 \le f(\mathbf{v}) + \lambda\|\mathbf{v}\|_0 \le f(\mathbf{x}) + \langle \mathbf{v} - \mathbf{x}, \nabla f(\mathbf{x})\rangle + \frac{L}{2}\|\mathbf{v} - \mathbf{x}\|^2 + \lambda\|\mathbf{v}\|_0.$$

Consequently, by (a) and (b) it follows that

$$\lambda\left(\|\mathbf{v}^+\|_0 - \|\mathbf{x}\|_0\right) \geq \left\langle \mathbf{x}-\mathbf{v}^+, \nabla f(\mathbf{x})\right\rangle - \frac{L}{2}\left\|\mathbf{v}^+ - \mathbf{x}\right\|^2, \tag{4.8}$$

$$\lambda\left(\|\mathbf{x}\|_0 - \|\mathbf{v}^-\|_0\right) \leq \left\langle \mathbf{v}^- - \mathbf{x}, \nabla f(\mathbf{x})\right\rangle + \frac{L}{2}\left\|\mathbf{v}^- - \mathbf{x}\right\|^2. \tag{4.9}$$

Since $\mathbf{x} = P_B^\sigma\left(T_L(\mathbf{x}); \|\mathbf{x}\|_0\right)$, $\mathbf{v}^+ \equiv P_B^\sigma\left(T_L(\mathbf{x}); \|\mathbf{x}\|_0 + 1\right)$, and $\mathbf{v}^- \equiv P_B^\sigma\left(T_L(\mathbf{x}); \|\mathbf{x}\|_0 - 1\right)$, we have that

$$Ld_B^\sigma\left(T_L(\mathbf{x}); \|\mathbf{x}\|_0\right) = \frac{L}{2}\left\|\mathbf{x} - T_L(\mathbf{x})\right\|_2^2 - \frac{L}{2}\left\|\mathbf{v}^+ - T_L(\mathbf{x})\right\|_2^2 = \left\langle \mathbf{x}-\mathbf{v}^+, \nabla f(\mathbf{x})\right\rangle - \frac{L}{2}\left\|\mathbf{v}^+ - \mathbf{x}\right\|^2,$$

and

$$Ld_B^\sigma\left(T_L(\mathbf{x}); \|\mathbf{x}\|_0 - 1\right) = \frac{L}{2}\left\|\mathbf{v}^- - T_L(\mathbf{x})\right\|_2^2 - \frac{L}{2}\left\|\mathbf{x} - T_L(\mathbf{x})\right\|_2^2 = \left\langle \mathbf{v}^- - \mathbf{x}, \nabla f(\mathbf{x})\right\rangle + \frac{L}{2}\left\|\mathbf{v}^- - \mathbf{x}\right\|^2.$$

Plugging the former and latter into (4.8) and (4.9) respectively implies that

$$d_B^\sigma\left(T_L(\mathbf{x}); \|\mathbf{x}\|_0\right) \leq \frac{\lambda}{L}\left(\|\mathbf{v}^+\|_0 - \|\mathbf{x}\|_0\right) \text{ if } \|\mathbf{x}\|_0 \leq n-1, \tag{4.10}$$

$$d_B^\sigma\left(T_L(\mathbf{x}); \|\mathbf{x}\|_0 - 1\right) \geq \frac{\lambda}{L}\left(\|\mathbf{x}\|_0 - \|\mathbf{v}^-\|_0\right) \text{ if } \|\mathbf{x}\|_0 \geq \ell_B + 1. \tag{4.11}$$

By the definition of $\mathbf{v}^+$, it follows that $\|\mathbf{v}^+\|_0 \leq \|\mathbf{x}\|_0 + 1$, and hence (4.10) implies that

$$d_B^\sigma\left(T_L(\mathbf{x}); \|\mathbf{x}\|_0\right) \leq \frac{\lambda}{L} \text{ if } \|\mathbf{x}\|_0 \leq n-1. \tag{4.12}$$

Since $\mathbf{x} = P_B^\sigma\left(T_L(\mathbf{x}); \|\mathbf{x}\|_0\right)$, we have that $\|\mathbf{x}\|_0 = \|P_B^\sigma\left(T_L(\mathbf{x}); \|\mathbf{x}\|_0\right)\|_0$, and consequently by Lemma 3.8, $\|\mathbf{x}\|_0 \leq u_{B,\sigma,T_L(\mathbf{x})}$. Subsequently $\|\mathbf{x}\|_0 - 1 < u_{B,\sigma,T_L(\mathbf{x})}$, and by Lemma 3.8 again, it holds that $\|\mathbf{v}^-\|_0 = \|\mathbf{x}\|_0 - 1$. Therefore, (4.11) implies that

$$d_B^\sigma\left(T_L(\mathbf{x}); \|\mathbf{x}\|_0 - 1\right) \geq \frac{\lambda}{L} \text{ if } \|\mathbf{x}\|_0 \geq \ell_B + 1. \tag{4.13}$$

By Theorem 3.9, combining (4.12), (4.13), we obtain that $\|\mathbf{x}\|_0 \in D_{\lambda/L}(T_L(\mathbf{x}))$, which means that $\mathbf{x} = P_B^\sigma(T_L(\mathbf{x}); \|\mathbf{x}\|_0) \in \operatorname{prox}_{\frac{\lambda}{L}g_B}(T_L(\mathbf{x}))$, showing that $\mathbf{x}$ is an L-stationary point. ◻

4.4. Partial Coordinate-Wise Optimality. Loosely speaking, in the context of sparsity-related problems, coordinate-wise optimality conditions are conditions that state that the function value does not improve if a small change in the support is performed. For a given support optimal point $\mathbf{x}$, the condition that we will consider will compare the function value of $\mathbf{x}$ with those of the following three support optimal points:

$$\mathbf{v}_\mathbf{x}^- \in \operatorname{argmin}\left\{f(\mathbf{u}) : I_1(\mathbf{u}) \subseteq I_1(\mathbf{x})\backslash\{i_\mathbf{x}\}, \mathbf{u} \in B\right\}, \tag{4.14}$$

$$\mathbf{v}_\mathbf{x}^{\text{swap}} \in \operatorname{argmin}\left\{f(\mathbf{u}) : I_1(\mathbf{u}) \subseteq (I_1(\mathbf{x})\backslash\{i_\mathbf{x}\}) \cup \{j_\mathbf{x}\}, \mathbf{u} \in B\right\}, \tag{4.15}$$

$$\mathbf{v}_\mathbf{x}^+ \in \operatorname{argmin}\left\{f(\mathbf{u}) : I_1(\mathbf{u}) \subseteq I_1(\mathbf{x}) \cup \{j_\mathbf{x}\}, \mathbf{u} \in B\right\}, \tag{4.16}$$

where

$$i_{\mathbf{x}} \in \operatorname*{argmin}_{\ell \in C(\mathbf{x})} \{p_B(-\nabla_\ell f(\mathbf{x}))\} \text{ with } C(\mathbf{x}) = \operatorname*{argmin,}_{k \in I_1(\mathbf{x})} p_B(x_k) \tag{4.17}$$

$$j_{\mathbf{x}} \in \operatorname*{argmin}_{\ell \in I_0(\mathbf{x})} \left\{-p_B(-\nabla_\ell f(\mathbf{x}))\right\}. \tag{4.18}$$

Note that there might be several options of how to choose $i_{\mathbf{x}}$ and $j_{\mathbf{x}}$ given $\mathbf{x} \in \mathbb{R}^n$ and we assume that there exists some rule for choosing $i_{\mathbf{x}}$ and $j_{\mathbf{x}}$. We can now define the optimality condition, which we refer to as *partial coordinate-wise (CW) optimality*

DEFINITION 4.12 (partial CW optimality). *A support optimal vector $\mathbf{x} \in B$ is called a* ***partial coordinate-wise (CW) optimal point of problem (P)*** *if it satisfies*

$$f(\mathbf{x}) + \lambda\|\mathbf{x}\|_0 \leq \min\left\{f(\mathbf{v}) + \lambda\|\mathbf{v}\|_0 : \mathbf{v} \in \{\mathbf{v}_{\mathbf{x}}^-, \mathbf{v}_{\mathbf{x}}^{swap}, \mathbf{v}_{\mathbf{x}}^+\}\right\}.$$

The partial CW optimality condition is obviously a necessary optimality condition for problem (P).

THEOREM 4.13 (optimality ⇒ partial CW optimality). *Let $\mathbf{x}^*$ be an optimal solution of (P), then $\mathbf{x}^*$ is a partial CW optimal point of (P).*

The partial CW optimality condition implies L-stationarity for any $L \geq L_f$. To prove this claim, we will use the following result proved in [3], and stated in the terminology of the current paper.

LEMMA 4.14 ([3, Theorem 6.1]). *Let $\mathbf{x} \in \mathbb{R}^n$ be a support optimal point. If the inequality $f(\mathbf{x}) \leq f(\mathbf{v}^{swap})$ holds, then for any $L \geq L_f$ it holds that $\mathbf{x} \in P_{C_{\|\mathbf{x}\|_0} \cap B}\left(T_L(\mathbf{x})\right)$.*

By combining Lemma 4.14 together with Lemma 3.3 we obtain the following result.

LEMMA 4.15. *Let $\mathbf{x} \in \mathbb{R}^n$ be a support optimal point of (P) and let $L \geq L_f$. If*

$$f(\mathbf{x}) \leq f(\mathbf{v}^{swap}),$$

then there exists $\sigma \in \tilde{\Sigma}(p_B(\mathbf{x})) \cap \tilde{\Sigma}(p_B(T_L(\mathbf{x})))$ for which $\mathbf{x} = P_B^\sigma\left(T_L(\mathbf{x}); \|\mathbf{x}\|_0\right)$ and

$$\sigma(\|\mathbf{x}\|_0) \in \operatorname*{argmin}_{\ell \in C(\mathbf{x})} \{p_B(-\nabla_\ell f(\mathbf{x}))\} \text{ with } C(\mathbf{x}) = \operatorname*{argmin}_{t \in I_1(\mathbf{x})} p_B(x_t), \tag{4.19}$$

$$\sigma(\|\mathbf{x}\|_0 + 1) \in \operatorname*{argmin}_{\ell \in I_0(\mathbf{x})} \left\{-p_B(-\nabla_\ell f(\mathbf{x}))\right\}. \tag{4.20}$$

Proof. Denote $k = \|\mathbf{x}\|_0$. Lemma 4.14 implies that $\mathbf{x} \in P_{C_k \cap B}\left(T_L(\mathbf{x})\right)$, and thus by Lemma 3.3 there exists a permutation

$$\sigma \in \tilde{\Sigma}(p_B(\mathbf{x})) \cap \tilde{\Sigma}(p_B(T_L(\mathbf{x}))) \tag{4.21}$$

for which $\mathbf{x} = P_B^\sigma\left(T_L(\mathbf{x}); k\right)$. Then in particular

$$\sigma(k) \in C(\mathbf{x}) \cap \operatorname*{argmin}_{\ell \in I_1(\mathbf{x})} \{p_B(T_L(\mathbf{x}))_\ell\} \text{ and } x_{\sigma(m)} = 0 \text{ for all } m \geq k+1. \tag{4.22}$$

We will now show that (4.19) and (4.20) are satisfied. Since $x_{\sigma(m)} = 0$ for all $m \geq k+1$, relation (4.20) follows trivially from (4.21). To prove relation (4.19), we will consider two cases.

- ***B* is nonnegative symmetric**. For any $t \in C(\mathbf{x})$ it holds that $x_t = x_{\sigma(k)}$, and by (4.22),

$T_L(\mathbf{x})_{\sigma(k)} \le T_L(\mathbf{x})_t$. Combining these two fact we conclude that (4.19) holds.

- ***B* is absolutely symmetric**. Since $\mathbf{x}$ is support optimal and B is absolutely symmetric, it follows by Lemma 4.7(b) that $x_t \nabla_t f(\mathbf{x}) \le 0$ for any $t \in I_1(\mathbf{x})$. Therefore, for any $t \in C(\mathbf{x})$ (noting that $C(\mathbf{x}) \subseteq I_1(\mathbf{x})$),

$$\left| x_t - \frac{1}{L}\nabla_t f(\mathbf{x}) \right| = |x_t| + \frac{1}{L}\left|\nabla_t f(\mathbf{x})\right|. \tag{4.23}$$

Now for any $t \in C(\mathbf{x})$ it holds that $|x_t| = |x_{\sigma(k)}|$, and by (4.22) it holds that $|T_L(\mathbf{x})_{\sigma(k)}| \le |T_L(\mathbf{x})_t|$, which by (4.23) implies that $|\nabla_{\sigma(k)} f(\mathbf{x})| \le |\nabla_t f(\mathbf{x})|$, implying the validity of (4.20) in this case. □

In the next claim we show that any partial CW optimal point is an L-stationary point for any $L \ge L_f$.

THEOREM 4.16 (partial CW optimality ⇒ L-stationarity for $L \ge L_f$). *If* $\mathbf{x}$ *is a partial CW optimal point of (P), then it is an* L*-stationary point of (P) for any* $L \ge L_f$.

Proof. Suppose that $\mathbf{x}$ is a partial CW optimal point of (P). Then by definition, $\mathbf{x}$ is a support optimal point of (P) that satisfies

$$f(\mathbf{x}) + \lambda\|\mathbf{x}\|_0 \le \min\left\{ f(\mathbf{v}) + \lambda\|\mathbf{v}\|_0 : \mathbf{v} \in \{\mathbf{v}_\mathbf{x}^-, \mathbf{v}_\mathbf{x}^{\text{swap}}, \mathbf{v}_\mathbf{x}^+\}\right\}. \tag{4.24}$$

where $\mathbf{v}_\mathbf{x}^-, \mathbf{v}_\mathbf{x}^{\text{swap}}, \mathbf{v}_\mathbf{x}^+$ are defined in (4.14), (4.15), and (4.16) respectively, with $i_\mathbf{x}$ and $j_\mathbf{x}$ defined in (4.17) and (4.18) respectively.

Denote $k = \|\mathbf{x}\|_0$. By its definition, $\mathbf{v}_\mathbf{x}^{\text{swap}}$ satisfies $\|\mathbf{v}_\mathbf{x}^{\text{swap}}\|_0 \le k$, and consequently by (4.24) we have that

$$f(\mathbf{v}_\mathbf{x}^{\text{swap}}) - f(\mathbf{x}) \ge \lambda(k - \|\mathbf{v}_\mathbf{x}^{\text{swap}}\|_0) \ge 0.$$

Thus, since $\mathbf{x}$ is a support optimal point of (P) that satisfies $f(\mathbf{x}) \le f(\mathbf{v}^{\text{swap}})$, Corollary 4.15 implies that there exists a permutation

$$\sigma \in \tilde{\Sigma}(p_B(\mathbf{x})) \cap \tilde{\Sigma}(p_B(T_L(\mathbf{x}))) \tag{4.25}$$

for which

$$\mathbf{x} = P_B^\sigma\left(T_L(\mathbf{x}); k\right) \tag{4.26}$$

and

$$\sigma(k) \in \operatorname*{argmin}_{\ell \in C(\mathbf{x})}\{p_B(-\nabla_\ell f(\mathbf{x}))\} \text{ with } C(\mathbf{x}) = \operatorname*{argmin}_{t \in I_1(\mathbf{x})} p_B(x_t) \tag{4.27}$$

$$\sigma(k+1) \in \operatorname*{argmin}_{\ell \in I_0(\mathbf{x})}\left\{-p_B(-\nabla_\ell f(\mathbf{x}))\right\}. \tag{4.28}$$

In particular,

$$p_B(x_{i_\mathbf{x}}) = p_B(x_{\sigma(k)}), p_B(-\nabla_{i_\mathbf{x}} f(\mathbf{x})) = p_B(-\nabla_{\sigma(k)} f(\mathbf{x}))$$

and

$$p_B(-\nabla_{j_\mathbf{x}} f(\mathbf{x})) = p_B(-\nabla_{\sigma(k+1)} f(\mathbf{x})).$$

Hence, if $i_\mathbf{x} \ne \sigma(k)$ or $j_\mathbf{x} \ne \sigma(k+1)$, we can swap in σ (employing a transposition permutation) between $i_\mathbf{x}$

and $\sigma(k)$ and between $j_{\mathbf{x}}$ and $\sigma(k+1)$ in σ, implying that we can assume that in addition for the properties (4.25),(4.26), (4.27),(4.28), σ also satisfies $\sigma(k) = i_{\mathbf{x}}$ and $\sigma(k+1) = j_{\mathbf{x}}$. Consequently, $S^{\sigma}_{k-1} = I_1(\mathbf{x})\backslash\{i_{\mathbf{x}}\}$ and $S^{\sigma}_{k+1} = I_1(\mathbf{x})\cup\{j_{\mathbf{x}}\}$, and thus, $I_1\left(P^{\sigma}_{B}\left(T_L(\mathbf{x}); k-1\right)\right) \subseteq I_1(\mathbf{x})\backslash\{i_{\mathbf{x}}\}$, and $I_1\left(P^{\sigma}_{B}\left(T_L(\mathbf{x}); k+1\right)\right) \subseteq I_1(\mathbf{x})\cup\{j_{\mathbf{x}}\}$, which in turn, by (4.14) and (4.16), implies that

$$f(\mathbf{v}^{-}_{\mathbf{x}}) \le f\left(P^{\sigma}_{B}\left(T_L(\mathbf{x}); k-1\right)\right), \text{ and } f(\mathbf{v}^{+}_{\mathbf{x}}) \le f\left(P^{\sigma}_{B}\left(T_L(\mathbf{x}); k+1\right)\right). \tag{4.29}$$

We will now show that the sufficient conditions for L-stationarity involving the $(k \pm 1)$-sparse projections, given in Lemma 4.11, are satisfied.

Assume that $k \ge \ell_B + 1$. Since $k = \|P^{\sigma}_{B}\left(T_L(\mathbf{x}); k\right)\|_0$, it holds that

$$k \le u_{B,\sigma,T_L(\mathbf{x})}, \tag{4.30}$$

and thus, by Lemma 3.8, $\|P^{\sigma}_{B}\left(T_L(\mathbf{x}); k-1\right)\|_0 = k-1$. Consequently, by combining (4.24), (4.29) and the fact that $\|\mathbf{v}^{-}_{\mathbf{x}}\|_0 \le k-1$, we attain that if $k \ge \ell_B + 1$, it holds that

$$\lambda k + f(\mathbf{x}) \le \lambda\|\mathbf{v}^{-}_{\mathbf{x}}\|_0 + f(\mathbf{v}^{-}_{\mathbf{x}}) \le \lambda\|P^{\sigma}_{B}\left(T_L(\mathbf{x}); k-1\right)\|_0 + f\left(P^{\sigma}_{B}\left(T_L(\mathbf{x}); k-1\right)\right). \tag{4.31}$$

Now assume that $k \le n-1$. If $\|P^{\sigma}_{B}\left(T_L(\mathbf{x}); k+1\right)\|_0 < k+1$, then due to (4.26) and (4.30), Lemma 3.8 implies that $\|P^{\sigma}_{B}\left(T_L(\mathbf{x}); k+1\right)\|_0 = k$ and $P^{\sigma}_{B}\left(T_L(\mathbf{x}); k+1\right) = \mathbf{x}$. As a result, $\lambda k + f(\mathbf{x}) = \lambda\|P^{\sigma}_{B}\left(T_L(\mathbf{x}); k+1\right)\|_0 + f\left(P^{\sigma}_{B}\left(T_L; k+1\right)\right)$. Otherwise, $\|P^{\sigma}_{B}\left(T_L(\mathbf{x}); k+1\right)\|_0 = k+1$, and since $\|\mathbf{v}^{+}_{\mathbf{x}}\|_0 \le k+1$, we have that

$$\lambda k + f(\mathbf{x}) \le \lambda\|\mathbf{v}^{+}_{\mathbf{x}}\|_0 + f(\mathbf{v}^{+}_{\mathbf{x}}) \le \lambda\|P^{\sigma}_{B}\left(T_L(\mathbf{x}); k+1\right)\|_0 + f\left(P^{\sigma}_{B}\left(T_L(\mathbf{x}); k+1\right)\right),$$

concluding that in any case

$$\lambda k + f(\mathbf{x}) \le \lambda\|P^{\sigma}_{B}\left(T_L(\mathbf{x}); k+1\right)\|_0 + f\left(P^{\sigma}_{B}\left(T_L(\mathbf{x}); k+1\right)\right). \tag{4.32}$$

Since the sufficient conditions for L-stationarity, (4.25), (4.26), (4.31), and (4.32) hold, by Lemma 4.11 the vector $\mathbf{x}$ is an L-stationary point of (P) for any $L \ge L_f$. □

The hierarchy between the optimality conditions is summarized in the following diagram.

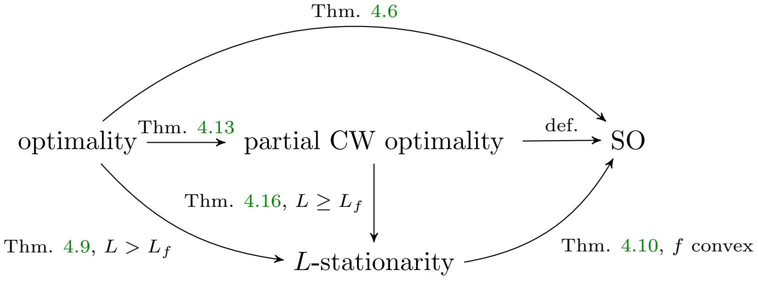

Fig. 1. *optimality points hierarchy*

The strictness of the hierarchy will be demonstrated in the following numerical example.

Example 4.17. *Consider the following optimization problem over the ℓ_1-norm unit ball, $C \equiv B_1[\mathbf{0}, 1]$:*

$$\min_{\mathbf{x}\in\mathbb{R}^{10}} \|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2 + 0.2\|\mathbf{x}\|_0 + \delta_C(\mathbf{x}),$$

where $\mathbf{A} \in \mathbb{R}^{10\times 10}$ and $\mathbf{b} \in \mathbb{R}^{10}$. We generated the components of $\mathbf{A}$ and $\mathbf{b}$ independently via a standard normal distribution. The following table contains, for a specific realizaion, the number of support optimal, L_f-stationary ($L_f = 2\lambda_{\max}(\mathbf{A}^T\mathbf{A})$), partial CW and optimal points. The superiority of the partial CW optimality condition is well illustrated in this example by the fact that there are significantly less partial CW optimal points (2) than L_f-stationary points (25) or support optimal points (887). We note that very similar results are obtained if other realizations are considered.

	supports	**support optimal**	**L-stationary**	**partial CW**	**optimal**
number of	1024	887	25	2	1

Table 2
points satisfying optimality conditions

In the next section we derive procedures to attain the defined optimality conditions.

5. Methods.

5.1. The Proximal Gradient Method. L-stationary points can be attained by the so-called proximal gradient method. In the setting of our problem, the prox operator can be computed using Algorithm 1 or Algorithm 2.

Algorithm 3 proximal gradient method

```
Input: $\mathbf{x}^0 \in \mathbb{R}^n$, $\epsilon, \lambda, L > 0$, $B$-simple symmetric set.
repeat
    1. $\mathbf{x}^{k+1} \in \operatorname{prox}_{\frac{\lambda}{L}g_B}\left(T_L(\mathbf{x}^k)\right)$;
    2. $k \leftarrow k+1$;
```

We will show that if in addition to our standing assumption (Assumption 4.1) we assume that f is lower bounded, then utilizing the sufficient decrease lemma (Lemma 4.3) allows us to to prove that limit points of the sequence generated by the proximal gradient method with $L > L_f$ are L-stationary points.

Theorem 5.1. *Assume that f is lower bounded. Let $\{\mathbf{x}^k\}_{k\geq 0}$ be the sequence generated by the proximal gradient method with with $L > L_f$. Then*
(a) $f(\mathbf{x}^k) + \lambda g_B(\mathbf{x}^k) - f(\mathbf{x}^{k+1}) - \lambda g_B(\mathbf{x}^{k+1}) \geq \frac{L-L(f)}{2}\left\|\mathbf{x}^{k+1}-\mathbf{x}^k\right\|^2$;
(b) any limit point of the sequence $\{\mathbf{x}^k\}_{k\geq 0}$ is an L-stationary point.

Proof. Part (a) readily follows from the sufficient decrease lemma (Lemma 4.3). To prove part (b), note that by part (a) the sequence of function values $\{f(\mathbf{x}^k) + \lambda g_B(\mathbf{x}^k)\}_{k\geq 0}$ is nonincreasing and in addition, by the assumption that f is lower bounded, it follows that the sequence is also lower bounded and hence convergent. We can thus conclude by part (a) that

$$\|\mathbf{x}^{k+1} - \mathbf{x}^k\| \to 0 \text{ as } k \to \infty. \tag{5.1}$$

Let $\mathbf{x}^*$ be a limit point of the sequence. Then there exists a subsequence $\{\mathbf{x}^{k_i}\}_{i\geq 1}$ that converges to $\mathbf{x}^*$, and hence, by (5.1), $\mathbf{x}^{k_i+1} \to \mathbf{x}^*$ as $i \to \infty$. Since $\mathbf{x}^{k_i+1} \in \operatorname{prox}_{\frac{\lambda}{L}g_B}(T_L(\mathbf{x}^{k_i}))$, by the definition of the prox operator we have

$$\frac{\lambda}{L}\|\mathbf{x}^{k_i+1}\|_0 + \frac{1}{2}\|\mathbf{x}^{k_i+1} - T_L(\mathbf{x}^{k_i})\|_2^2 \le \frac{\lambda}{L}\|\mathbf{x}\|_0 + \frac{1}{2}\|\mathbf{x} - T_L(\mathbf{x}^{k_i})\|_2^2 \quad \text{for all } \mathbf{x} \in B.$$

Taking the limit $i \to \infty$ yields

$$\frac{\lambda}{L}\|\mathbf{x}^*\|_0 + \frac{1}{2}\|\mathbf{x}^* - T_L(\mathbf{x}^*)\|_2^2 \le \frac{\lambda}{L}\|\mathbf{x}\|_0 + \frac{1}{2}\|\mathbf{x} - T_L(\mathbf{x}^*)\|_2^2 \quad \text{for all } \mathbf{x} \in B,$$

which along with the fact that $\mathbf{x}^* \in B$ (by the closedness of B) implies that $\mathbf{x}^* \in \text{prox}_{\frac{\lambda}{L}g_B}(T_L(\mathbf{x}^*))$, meaning that $\mathbf{x}^*$ is an L-stationary point. □

5.2. The Coordinate-Wise Support Optimality method. The hierarchy of the optimality conditions illustrated in (1) suggests that partial coordinate-wise optimality is a more restrictive optimality condition than L-stationarity for $L \ge L_f$. The following method creates a sequence of support optimal points and returns a partial CW optimal point in finite number of steps. We therefore refer to it as the **co**ordinate-**w**ise **s**upport optimality method.

Algorithm 4 CowS

```
Input: $\mathbf{y} \in \mathbb{R}^n$, $\lambda > 0$, $B$-simple symmetric set.
  1. set $\mathbf{x}^0 = \text{argmin}\{f(\mathbf{u}) : I_1(\mathbf{u}) \subseteq I_1(\mathbf{y})\}$ and $k \leftarrow 0$;
  2. set $\mathbf{x} = \mathbf{x}^k$ and compute
        $i_{\mathbf{x}} \in \underset{\ell \in C(\mathbf{x})}{\text{argmin}}\{p_B(-\nabla_\ell f(\mathbf{x}))\}$ with $C(\mathbf{x}) = \underset{k \in I_1(\mathbf{x})}{\text{argmin}}\, p_B(x_k)$,
        $j_{\mathbf{x}} \in \underset{\ell \in I_0(\mathbf{x})}{\text{argmin}}\{-p_B(-\nabla_\ell f(\mathbf{x}))\}$;
  3. compute
        $\mathbf{v}_{\mathbf{x}}^- \in \text{argmin}\{f(\mathbf{u}) : I_1(\mathbf{u}) \subseteq I_1(\mathbf{x})\backslash\{i_{\mathbf{x}}\}, \mathbf{u} \in B\}$,
        $\mathbf{v}_{\mathbf{x}}^{\text{swap}} \in \text{argmin}\{f(\mathbf{u}) : I_1(\mathbf{u}) \subseteq (I_1(\mathbf{x})\backslash\{i_{\mathbf{x}}\}) \cup \{j_{\mathbf{x}}\}, \mathbf{u} \in B\}$,
        $\mathbf{v}_{\mathbf{x}}^+ \in \text{argmin}\{f(\mathbf{u}) : I_1(\mathbf{u}) \subseteq I_1(\mathbf{x}) \cup \{j_{\mathbf{x}}\}, \mathbf{u} \in B\}$;
  4. if $f(\mathbf{x}) + \lambda\|\mathbf{x}\|_0 \le \min\{f(\mathbf{v}) + \lambda\|\mathbf{v}\|_0 : \mathbf{v} \in \{\mathbf{v}_{\mathbf{x}}^-, \mathbf{v}_{\mathbf{x}}^{\text{swap}}, \mathbf{v}_{\mathbf{x}}^+\}\}$, stop and return $\mathbf{x}$.
  5. set $\mathbf{x}^{k+1} \in \text{argmin}\{f(\mathbf{u}) + \lambda\|\mathbf{u}\|_0 : \mathbf{u} \in \{\mathbf{v}_{\mathbf{x}}^-, \mathbf{v}_{\mathbf{x}}^{\text{swap}}, \mathbf{v}_{\mathbf{x}}^+\}\}$, $k \leftarrow k+1$, and go to step 2.
```

Note that the vectors $\mathbf{v}_{\mathbf{x}}^-, \mathbf{v}_{\mathbf{x}}^{\text{swap}}, \mathbf{v}_{\mathbf{x}}^+$ are not necessarily well defined since the optimization problems defining them might have multiple optimal solutions. To make them well defined, and in order to assure the finiteness of the CowS algorithm, we make the assumption that there is some deterministic rule for choosing an optimal solution among multiple optimal solutions (if exist) for problems of the form

$$\min\{f(\mathbf{u}) : I_1(\mathbf{u}) \subseteq T, \mathbf{u} \in B\} \tag{5.2}$$

where $T \subseteq \{1, 2, \ldots, n\}$ is an index set. Under this assumption, the algorithm is obviously finite since it generates a sequence with strictly decreasing function values and there are only a finite amount of possible vectors through which it passes–the optimal solutions of problems of the form (5.2). Since there are only a finite amount of index sets (2^n), the finiteness of the algorithm follows.

It is easy to show that the outcome of the CowS method is a partial CW optimal point.

THEOREM 5.2. *Let* $\mathbf{x}$ *be the output of the CowS method. Then* $\mathbf{x}$ *is a partial CW optimal point of problem (P).*

Proof. By the definition of the method, $\mathbf{x}$ is a support optimal point satisfying

$$f(\mathbf{x}) + \lambda\|\mathbf{x}\|_0 \leq \min\{f(\mathbf{u}) + \lambda\|\mathbf{u}\|_0 : \mathbf{u} \in \{\mathbf{v}_{\mathbf{x}}^{-}, \mathbf{v}_{\mathbf{x}}^{\mathrm{swap}}, \mathbf{v}_{\mathbf{x}}^{+}\}\},$$

which means that it is a partial CW optimal point. ∎

REFERENCES

[1] A. Beck. *Introduction to Nonlinear Optimization: Theory, Algorithms, and Applications with MATLAB.* MOS-SIAM Series on Optimization. Society for Industrial and Applied Mathematics, 2014.

[2] A. Beck and Y.C. Eldar. Sparsity constrained nonlinear optimization: Optimality conditions and algorithms. *SIAM Journal on Optimization*, 23(3):1480–1509, 2013.

[3] A. Beck and N. Hallak. On the minimization over sparse symmetric sets: Projections, optimality conditions, and algorithms. *Mathematics of Operations Research*, 41(1):196–223, 2016.

[4] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009.

[5] A. Beck and Y. Vaisbourd. The sparse principal component analysis problem: Optimality conditions and algorithms. *Journal of Optimization Theory and Applications*, pages 1–25, 2016.

[6] D. P. Bertsekas. *Nonlinear programming.* Athena Scientific, Belmont, MA, 2nd edition, 1999.

[7] T. Blumensath. Compressed sensing with nonlinear observations and related nonlinear optimization problems. *IEEE Trans. Inform. Theory*, 59(6):3466–3474, 2013.

[8] T. Blumensath and M. E. Davies. Iterative thresholding for sparse approximations. *The Journal of Fourier Analysis and Applications*, 14(5):629–654, 2008.

[9] J. Bolte, S. Sabach, and M. Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1):459–494, 2014.

[10] J. M. Borwein and A. S. Lewis. *Convex analysis and nonlinear optimization: theory and examples.* Springer Science & Business Media, 2010.

[11] A.M. Bruckstein, D.L. Donoho, and M. Elad. From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM review*, 51(1):34–81, 2009.

[12] E. Candes and T. Tao. The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, December 2007.

[13] E.J. Candes and T. Tao. Decoding by linear programming. *Information Theory, IEEE Transactions on*, 40698(December):1–22, 2005.

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

[15] M.A. Davenport, M.F. Duarte, Y.C. Eldar, and G. Kutyniok. Introduction to compressed sensing. *Preprint*, pages 1–68, 2011.

[16] D. L. Donoho and M. Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via l1 minimization. *Proceedings of the National Academy of Sciences*, 100(5):2197–2202, 2003.

[17] D.L. Donoho. Compressed sensing. *Information Theory, IEEE Transactions on*, 52(4):1289–1306, 2006.

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

[19] M. Elad. *Sparse and redundant representations: from theory to applications in signal and image processing.* Springer, 2010.

[20] A. S. Lewis. The convex analysis of unitarily invariant matrix functions. *Journal of Convex Analysis*, 2(1):173–183, 1995.

[21] A. S. Lewis. Convex analysis on the Hermitian matrices. *SIAM Journal on Optimization*, 6(1):164–177, 1996.

[22] Z. Lu and Y. Zhang. Sparse approximation via penalty decomposition methods. *SIAM Journal on Optimization*, 23(4):2448–2478, 2013.

[23] B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM journal on computing*, 24(2):227–234, 1995.

[24] H. Ohlsson, A. Yang, R. Dong, and S. Sastry. Compressive phase retrieval from squared output measurements via

semidefinite programming. In *16th IFAC Symposium on System Identification, Brussels, Belgium, 11-13 July, 2012*, pages 89–94, 2012.

[25] Y. Shechtman, A. Beck, and Y. C. Eldar. Gespar: Efficient phase retrieval of sparse signals. *Signal Processing, IEEE Transactions on*, 62(4):928–938, 2014.

[26] J. A. Tropp. Just relax: Convex programming methods for identifying sparse signals in noise. *Information Theory, IEEE Transactions on*, 52(3):1030–1051, 2006.

[27] J.A Tropp and S.J. Wright. Computational Methods for Sparse Solution of Linear Inverse Problems. *Proceedings of the IEEE*, 98(6):948–958, June 2010.

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

[29] Y. B. Zhao and D. Li. Reweighted ℓ_1-minimization for sparse solutions to underdetermined linear systems. *SIAM Journal on Optimization*, 22(3):1065–1088, 2012.

Appendix A. Sets Satisfying the SOM property.

In this section we will prove that several notable sets satisfy the SOM property – separable sets, the ℓ_1 and ℓ_2 α-balls, the α-simplex and the full α-simplex ($\alpha > 0$). In proving the SOM property of this selection of sets, we will aim to prove that for any $\mathbf{x} \in \mathbb{R}^n$, $\sigma \in \tilde{\Sigma}(p_B(\mathbf{x}))$ and $i \in \{\ell_B, \ell_B+1, \dots, n-2\}$, the following relation holds:

$$\|\mathbf{x} - P_B^\sigma(\mathbf{x}; i)\|_2^2 - \|\mathbf{x} - P_B^\sigma(\mathbf{x}; i+1)\|_2^2 \geq \|\mathbf{x} - P_B^\sigma(\mathbf{x}; i+1)\|_2^2 - \|\mathbf{x} - P_B^\sigma(\mathbf{x}; i+2)\|_2^2. \tag{A.1}$$

For the sake of simplicity, we will use the following notation:
Given a permutation $\sigma \in \Sigma_n$ and $l \in \{0, 1, \dots, n-1\}$, the vector $\mathbf{y}_{\langle l \rangle_\sigma} \in \mathbb{R}^n$ has the original values of $\mathbf{y}$ in indices S_l^σ, and zeros elsewhere, that is, $(\mathbf{y}_{\langle l \rangle_\sigma})_{S_l^\sigma} = \mathbf{y}_{S_l^\sigma}$, $(\mathbf{y}_{\langle l \rangle_\sigma})_{(S_l^\sigma)^c} = \mathbf{0}$.

A.1. General results for absolutely symmetric sets. We begin by establishing several properties of the sparse orthogonal projection operator onto absolutely symmetric sets that tie it to the sparse orthogonal projection onto nonnegative symmetric sets. The first result extends the correspondence given in Lemma 2.10(b).

Corollary A.1. *Let $\mathbf{x} \in \mathbb{R}^n$ and $\sigma \in \tilde{\Sigma}(|\mathbf{x}|)$. Suppose that $D \subseteq \mathbb{R}^n$ is nonempty, absolutely symmetric, closed and convex, and $B = D \cap \mathbb{R}^n_+$. Then for any $i \in \{0, 1, \dots, n\}$*

$$P_D^\sigma(\mathbf{x}; i) = \operatorname{sign}(\mathbf{x}) \odot P_B^\sigma(|\mathbf{x}|; i). \tag{A.2}$$

Proof. Let $T = S_i^\sigma$. Then $P_D^\sigma(\mathbf{x}; i) = \mathbf{U}_T P_{D_T}(\mathbf{x}_T)$. By Lemma 2.10(c) (taking into account the closedness and convexity of D_T),

$$\mathbf{U}_T P_{D_T}(\mathbf{x}_T) = \mathbf{U}_T \left(\operatorname{sign}(\mathbf{x}_T) \odot P_{D_T \cap \mathbb{R}^{|T|}_+}(|\mathbf{x}_T|) \right) = \operatorname{sign}(\mathbf{x}) \odot \mathbf{U}_T P_{D_T \cap \mathbb{R}^{|T|}_+}(|\mathbf{x}_T|).$$

Since $B_T = D_T \cap \mathbb{R}^{|T|}_+$, $\mathbf{U}_T P_{D_T \cap \mathbb{R}^{|T|}_+}(|\mathbf{x}_T|) = \mathbf{U}_T P_{B_T}(|\mathbf{x}_T|)$, and by the choice of $\sigma \in \tilde{\Sigma}(|\mathbf{x}|)$, it holds that $\mathbf{U}_T P_{B_T}(|\mathbf{x}_T|) = P_B^\sigma(|\mathbf{x}|; i)$. Thus, by the derived chain of equalities, (A.2) holds. □

We can now establish that if the nonnegative part of an absolutely symmetric set satisfies the SOM property, then so does the entire set.

Lemma A.2. *Let $D \subseteq \mathbb{R}^n$ be a nonempty, absolutely symmetric, closed and convex set, and $B = D \cap \mathbb{R}^n_+$. If B satisfies the SOM property, then D satisfies the SOM property.*

Proof. Let $\mathbf{x} \in \mathbb{R}^n$ and $\sigma \in \tilde{\Sigma}(|\mathbf{x}|)$. By Corollary A.1, $P_D^\sigma(\mathbf{x}; i) = \operatorname{sign}(\mathbf{x}) \odot P_B^\sigma(|\mathbf{x}|; i)$ for any $i \in \{0, 1, \dots, n\}$. Since $\|P_D^\sigma(\mathbf{x}; i) - \mathbf{x}\|_2 = \|\operatorname{sign}(\mathbf{x}) \odot (P_B^\sigma(|\mathbf{x}|; i) - |\mathbf{x}|)\|_2 = \|P_B^\sigma(|\mathbf{x}|; i) - |\mathbf{x}|\|_2$, the set D satisfies the SOM property if and only if B satisfies it for $|\mathbf{x}|$. Therefore, by the underlying assumption that B

satisfies the SOM property, the required holds. □

The former corollary suggests that members of the sparse projection sequences onto an absolutely symmetric set D can easily be obtained from the corresponding members of the sparse projection sequence onto the intersection $D \cap \mathbb{R}^n_+$. It will now be shown that for the members of the sparse projection sequence onto such intersections, the components corresponding to the non-positive elements in the input vector are zeros.

Lemma A.3. *Let $\mathbf{x} \in \mathbb{R}^n$, $\sigma \in \tilde{\Sigma}(\mathbf{x})$ and $J = \{l : x_l > 0\}$. Suppose that $D \subseteq \mathbb{R}^n$ is nonempty, absolutely symmetric, closed and convex, and let $B = D \cap \mathbb{R}^n_+$. Then $P^{\sigma}_B(\mathbf{x}; i) = P^{\sigma}_B(\mathbf{x}; r_i)$ for any $i \in \{0, 1, \ldots, n\}$, where $r_i = \min\{i, |J|\}$.*

Proof. We will first show that the assertion is correct for $i = n$, in which case $P^{\sigma}_B(\mathbf{x}; i) = P_B(\mathbf{x})$. To prove the required in this case it is enough to prove that $P_B(\mathbf{x})_{J^c} = \mathbf{0}$. Since B is nonnegative, $P_B(\mathbf{x})_{J^c} \geq \mathbf{0} \geq \mathbf{x}_{J^c}$, and we thus have that $\|P_B(\mathbf{x})_{J^c} - \mathbf{x}_{J^c}\|_2^2 \geq \|\mathbf{x}_{J^c}\|_2^2$. Consequently,

$$\begin{aligned} \|P_B(\mathbf{x}) - \mathbf{x}\|_2^2 = \|P_B(\mathbf{x})_{J^c} - \mathbf{x}_{J^c}\|_2^2 + \|P_B(\mathbf{x})_J - \mathbf{x}_J\|_2^2 &\geq \|\mathbf{x}_{J^c}\|_2^2 + \|P_B(\mathbf{x})_J - \mathbf{x}_J\|_2^2 \\ &= \|\mathbf{U}_J P_B(\mathbf{x})_J - \mathbf{x}\|_2^2 . \end{aligned} \tag{A.3}$$

Since D is absolutely symmetric, $P_B(\mathbf{x}) \in B \subseteq D$ implies that $[\mathbf{U}_J P_B(\mathbf{x})_J - \mathbf{U}_{J^c} P_B(\mathbf{x})_{J^c}] \in D$, and therefore, by the convexity of D,

$$\mathbf{U}_J P_B(\mathbf{x})_J = 0.5 \left(P_B(\mathbf{x}) + [\mathbf{U}_J P_B(\mathbf{x})_J - \mathbf{U}_{J^c} P_B(\mathbf{x})_{J^c}]\right) \in D.$$

Hence, $\mathbf{U}_J P_B(\mathbf{x})_J \in D \cap \mathbb{R}^n_+ = B$, and by (A.3) and the uniqueness of the orthogonal projection onto close and convex sets, it follows that $P_B(\mathbf{x}) = \mathbf{U}_J P_B(\mathbf{x})_J$. Subsequently, $P_B(\mathbf{x})_{J^c} = \mathbf{0}$, and the required holds for $i = n$.

Let $i \in \{0, 1, \ldots, n\}$ and $T = S^{\sigma}_i$. Then $P^{\sigma}_B(\mathbf{x}; i) = \mathbf{U}_T P_{B_T}(\mathbf{x}_T)$. We will use the 'tilde' notation to refer to terms in dimension $\mathbb{R}^{|T|}$ (T is fixed). For an index set $W \subseteq \{1, \ldots, |T|\}$, the matrix $\tilde{\mathbf{U}}_W$ denotes the submatrix of the $|T|$-dimensional identity matrix $\mathbf{I}_{|T|}$ constructed from the columns corresponding to the index set W, $\tilde{\mathbf{x}} \equiv \mathbf{x}_T$, $\tilde{B} \equiv B_T$, and $\tilde{J} = \{l : (\mathbf{x}_T)_l > 0\}$. By applying the result of the first part (with adjusted dimensions) we have that $\mathbf{U}_T P_{B_T}(\tilde{\mathbf{x}}) = \mathbf{U}_T \tilde{\mathbf{U}}_{\tilde{J}} P_{\tilde{B}_{\tilde{J}}}(\tilde{\mathbf{x}}_{\tilde{J}})$. Since $\sigma \in \tilde{\Sigma}(\mathbf{x})$ and $J = \{l : x_l > 0\}$, it holds that $J = S^{\sigma}_{|J|}$. Therefore, $T \cap J = S^{\sigma}_i \cap S^{\sigma}_{|J|} = S^{\sigma}_{r_i}$, and subsequently $|\tilde{J}| = |T \cap J| = r_i$. Hence, by the equalities $\mathbf{U}_T \tilde{\mathbf{U}}_{\tilde{J}} = \mathbf{U}_{T \cap J}$ and $\tilde{\mathbf{x}}_{\tilde{J}} = \mathbf{x}_{T \cap J}$, we have that $\mathbf{U}_T \tilde{\mathbf{U}}_{\tilde{J}} P_{\tilde{B}_{\tilde{J}}}(\tilde{\mathbf{x}}_{\tilde{J}}) = \mathbf{U}_{S^{\sigma}_{r_i}} P_{B_{S^{\sigma}_{r_i}}}(\mathbf{x}_{S^{\sigma}_{r_i}}) = P^{\sigma}_B(\mathbf{x}; r_i)$, that proves the required. □

The following lemma shows that the relation (A.1) defining the SOM property is satisfied in some general cases when B is absolutely symmetric.

Lemma A.4. *Let $B \subseteq \mathbb{R}^n$ be a nonempty, absolutely symmetric, closed and convex set. Then for any $\mathbf{x} \in \mathbb{R}^n, \sigma \in \tilde{\Sigma}(|\mathbf{x}|)$ and $i \in \{0, 1, \ldots, n-2\}$ satisfying $\mathbf{x}_{\langle i+1 \rangle_\sigma} \in B$, the inequality (A.1) holds.*

Proof. Let $i \in \{0, 1, \ldots, n-2\}$. Since $\mathbf{x}_{\langle i+1 \rangle_\sigma} \in B$, and since B is absolutely symmetric, it follows that $\mathbf{x}_{\langle i \rangle_\sigma} - x_{\sigma(i+1)} \mathbf{e}_{\sigma(i+1)} \in B$, and hence, by the convexity of B,

$$\mathbf{x}_{\langle i \rangle_\sigma} = 0.5 \left(\mathbf{x}_{\langle i+1 \rangle_\sigma} + \mathbf{x}_{\langle i \rangle_\sigma} - x_{\sigma(i+1)} \mathbf{e}_{\sigma(i+1)}\right) \in B.$$

Since $\mathbf{x}_{\langle i \rangle_\sigma} \in B$ and $\mathbf{x}_{\langle i+1 \rangle_\sigma} \in B$, we have that $P^{\sigma}_B(\mathbf{x}; i) = \mathbf{x}_{\langle i \rangle_\sigma}$ and $P_B(\mathbf{x}; i+1) = \mathbf{x}_{\langle i+1 \rangle_\sigma}$. Thus, (A.1) is equivalent to $x^2_{\sigma(i+1)} \geq x^2_{\sigma(i+2)} - \|\mathbf{x}_{\langle i+2 \rangle_\sigma} - P^{\sigma}_B(\mathbf{x}; i+2)\|_2^2$, which is a valid inequality since $|x_{\sigma(i+1)}| \geq$

case	**pair**	**result in (A.5)**
$0 \geq x_{\sigma(i+1)} \geq x_{\sigma(i+2)}$	$(0,0)$	$0 \geq 0$
$\alpha > x_{\sigma(i+1)} > 0 \geq x_{\sigma(i+2)}$	$(x_{\sigma(i+1)}, 0)$	$x^2_{\sigma(i+1)} \geq 0$
$\alpha > x_{\sigma(i+1)} \geq x_{\sigma(i+2)} > 0$	$(x_{\sigma(i+1)}, x_{\sigma(i+2)})$	$x^2_{\sigma(i+1)} \geq x^2_{\sigma(i+2)}$
$x_{\sigma(i+1)} \geq \alpha > 0 \geq x_{\sigma(i+2)}$	$(\alpha, 0)$	$x^2_{\sigma(i+1)} - (x_{\sigma(i+1)} - \alpha)^2 \geq 0$
$x_{\sigma(i+1)} \geq \alpha > x_{\sigma(i+2)} > 0$	$(\alpha, x_{\sigma(i+2)})$	$2\alpha x_{\sigma(i+1)} \geq \alpha^2 + x^2_{\sigma(i+2)}$
$x_{\sigma(i+1)} \geq x_{\sigma(i+2)} \geq \alpha > 0$	(α, α)	$2\alpha x_{\sigma(i+1)} \geq 2\alpha x_{\sigma(i+2)}$

TABLE 3
possible cases for (A.5)

$|x_{\sigma(i+2)}|$. ∎

We will now prove the SOM property individually per set.

A.2. SOM property of the box. This subsection will show that the symmetric α-box defined by

$$D = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_\infty \leq \alpha\}, \tag{A.4}$$

and its intersection with the nonnegative orthant, satisfy the SOM property for any $\alpha \in (0, \infty]$, where $\alpha = \infty$ corresponds to the case $D \equiv \mathbb{R}^n$. To show that (A.4) satisfies the SOM property, we will first prove that $B = D \cap \mathbb{R}^n_+$ satisfies the SOM property.

THEOREM A.5 (SOM of nonnegative α-box). *Let $B = D \cap \mathbb{R}^n_+$ where D is defined in (A.4) with $\alpha \in (0, \infty]$. Then B satisfies the SOM property.*

Proof. Let $\mathbf{x} \in \mathbb{R}^n$ and $\sigma \in \tilde{\Sigma}(\mathbf{x})$. The projection of $\mathbf{x} \in \mathbb{R}^n$ onto B is given for $l = 1, \ldots, n$, by

$$P_B(\mathbf{x})_l = \begin{cases} 0, & x_l \leq 0, \\ x_l, & x_l \in (0, \alpha), \\ \alpha, & \alpha \leq x_l. \end{cases}$$

In addition, as long as $j \leq i$, it holds that $P^\sigma_B(\mathbf{x}; i)_{\sigma(j)} = P^\sigma_B(\mathbf{x}; i+1)_{\sigma(j)} = P^\sigma_B(\mathbf{x}; i+2)_{\sigma(j)} = P_B(\mathbf{x})_{\sigma(j)}$. Consequently, for any $i \in \{0, 1, \ldots, n-1\}$

$$\|P^\sigma_B(\mathbf{x}; i) - \mathbf{x}\|^2_2 - \|P^\sigma_B(\mathbf{x}; i+1) - \mathbf{x}\|^2_2 = x^2_{\sigma(i+1)} - (x_{\sigma(i+1)} - P^\sigma_B(\mathbf{x}; i+1)_{\sigma(i+1)})^2,$$

and the required in (A.1) can be transformed into

$$x^2_{\sigma(i+1)} - (x_{\sigma(i+1)} - P^\sigma_B(\mathbf{x}; i+1)_{\sigma(i+1)})^2 \geq x^2_{\sigma(i+2)} - (x_{\sigma(i+2)} - P^\sigma_B(\mathbf{x}; i+2)_{\sigma(i+2)})^2. \tag{A.5}$$

To show that (A.5) holds, we explore in Table 3 all the possibilities for $x_{\sigma(i+1)}$ and $x_{\sigma(i+2)}$ (first column), and their corresponding pairs $(P^\sigma_B(\mathbf{x}; i+1)_{\sigma(i+1)}, P^\sigma_B(\mathbf{x}; i+2)_{\sigma(i+2)})$ (second column), and show the resulting inequality (A.5) (third column). It can be easily seen that the derived inequality (A.5) holds in all cases. ∎

By combining Lemma A.2 and Theorem A.5, we conclude that D defined in (A.4) satisfies the SOM property.

THEOREM A.6 (SOM property of the α-box). *Let D be defined in (A.4) with $\alpha \in (0, \infty]$. Then D satisfies the SOM property.*

A.3. SOM property of the ℓ_2 ball.

THEOREM A.7 (SOM property of the ℓ_2 ball). *Let $B = B_2[\mathbf{0}, \alpha] = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_2 \leq \alpha\}$, where $\alpha > 0$. Then B satisfies the SOM property.*

Proof. Let $\mathbf{x} \in \mathbb{R}^n$ and $\sigma \in \tilde{\Sigma}(|\mathbf{x}|)$. Let $i \in \{0, 1, \ldots, n-2\}$. If $\|\mathbf{x}_{\langle i+1\rangle_\sigma}\|_2 \leq \alpha$, then $\mathbf{x}_{\langle i+1\rangle_\sigma} \in B$, and hence by Lemma A.4 the required (A.1) is satisfied. We will hereafter assume that $\|\mathbf{x}_{\langle i+1\rangle_\sigma}\|_2 > \alpha$.

For any $j \in \{0, 1, \ldots, n\}$, $\|\mathbf{x}_{\langle j\rangle_\sigma}\|_2 \leq \alpha$ implies that $P_B^\sigma(\mathbf{x}; j) = \mathbf{x}_{\langle j\rangle_\sigma}$, and $\|\mathbf{x}_{\langle j\rangle_\sigma}\|_2 > \alpha$ implies that $P_B^\sigma(\mathbf{x}; j) = \frac{\alpha}{\|\mathbf{x}_{\langle j\rangle_\sigma}\|_2}\mathbf{x}_{\langle j\rangle_\sigma}$. Therefore,

$$\|\mathbf{x} - P_B^\sigma(\mathbf{x}; j)\|_2^2 = \begin{cases} \|\mathbf{x}\|_2^2 - \|\mathbf{x}_{\langle j\rangle_\sigma}\|_2^2, & \|\mathbf{x}_{\langle j\rangle_\sigma}\|_2 \leq \alpha; \\ \|\mathbf{x}\|_2^2 - 2\alpha\|\mathbf{x}_{\langle j\rangle_\sigma}\|_2 + \alpha^2, & \|\mathbf{x}_{\langle j\rangle_\sigma}\|_2 > \alpha. \end{cases} \tag{A.6}$$

If $\|\mathbf{x}_{\langle i\rangle_\sigma}\|_2 > \alpha$, then $\|\mathbf{x}_{\langle i+1\rangle_\sigma}\|_2, \|\mathbf{x}_{\langle i+2\rangle_\sigma}\|_2 > \alpha$. Substituting (A.6) for $j = i, i+1, i+2$ into (A.1) results with: $-2\alpha\|\mathbf{x}_{\langle i\rangle_\sigma}\|_2 + 2\alpha\|\mathbf{x}_{\langle i+1\rangle_\sigma}\|_2 \geq -2\alpha\|\mathbf{x}_{\langle i+1\rangle_\sigma}\|_2 + 2\alpha\|\mathbf{x}_{\langle i+2\rangle_\sigma}\|_2$, which is the same as

$$\frac{\|\mathbf{x}_{\langle i+1\rangle_\sigma}\|_2^2 - \|\mathbf{x}_{\langle i\rangle_\sigma}\|_2^2}{\|\mathbf{x}_{\langle i+1\rangle_\sigma}\|_2 + \|\mathbf{x}_{\langle i\rangle_\sigma}\|_2} \geq \frac{\|\mathbf{x}_{\langle i+2\rangle_\sigma}\|_2^2 - \|\mathbf{x}_{\langle i+1\rangle_\sigma}\|_2^2}{\|\mathbf{x}_{\langle i+1\rangle_\sigma}\|_2 + \|\mathbf{x}_{\langle i+2\rangle_\sigma}\|_2}. \tag{A.7}$$

Since

$$\|\mathbf{x}_{\langle i+1\rangle_\sigma}\|_2^2 - \|\mathbf{x}_{\langle i\rangle_\sigma}\|_2^2 = x_{\sigma(i+1)}^2 \geq x_{\sigma(i+2)}^2 = \|\mathbf{x}_{\langle i+2\rangle_\sigma}\|_2^2 - \|\mathbf{x}_{\langle i+1\rangle_\sigma}\|_2^2,$$

and $\|\mathbf{x}_{\langle i+1\rangle_\sigma}\|_2 + \|\mathbf{x}_{\langle i\rangle_\sigma}\|_2 \leq \|\mathbf{x}_{\langle i+1\rangle_\sigma}\|_2 + \|\mathbf{x}_{\langle i+2\rangle_\sigma}\|_2$, the relation (A.7) holds, and consequently the required is satisfied.

Now suppose that $\|\mathbf{x}_{\langle i\rangle_\sigma}\|_2 \leq \alpha$ and $\|\mathbf{x}_{\langle i+1\rangle_\sigma}\|_2 > \alpha$ (implying that $\|\mathbf{x}_{\langle i+2\rangle_\sigma}\|_2 > \alpha$). Plugging the corresponding equations (A.6) for $j = i, i+1, i+2$, (A.1) becomes

$$4\alpha\|\mathbf{x}_{\langle i+1\rangle_\sigma}\|_2 - \|\mathbf{x}_{\langle i\rangle_\sigma}\|_2^2 - 2\alpha\|\mathbf{x}_{\langle i+2\rangle_\sigma}\|_2 - \alpha^2 \geq 0. \tag{A.8}$$

Denote $t = \|\mathbf{x}_{\langle i\rangle_\sigma}\|_2, a = \|\mathbf{x}_{\langle i+1\rangle_\sigma}\|_2, b = \sqrt{\|\mathbf{x}_{\langle i\rangle_\sigma}\|_2^2 + x_{\sigma(i+2)}^2}$. Then $\|\mathbf{x}_{\langle i+2\rangle_\sigma}\|_2 = \sqrt{a^2 + b^2 - t^2}$. Therefore, inequality (A.8) can be expressed in terms of a, b, t, α as: $4\alpha a - t^2 - \alpha^2 \geq 2\alpha\sqrt{a^2 + b^2 - t^2}$. As both sides are positive ($a > \alpha \geq t$), it is equivalent (after squaring and rearranging terms) to

$$q_t(a, b) \equiv 4\alpha^2(a^2 - b^2) + 8\alpha^2 a^2 - 8\alpha a(t^2 + \alpha^2) + (t^2 + \alpha^2)^2 + 4\alpha^2 t^2 \geq 0.$$

Since $a^2 = \|\mathbf{x}_{\langle i+1\rangle_\sigma}\|_2^2 = t^2 + x_{\sigma(i+1)}^2 \geq t^2 + x_{\sigma(i+2)}^2 = b^2$, it follows that

$$q_t(a, b) \geq q_t(a, a) = 8\alpha^2 a^2 - 8\alpha a(t^2 + \alpha^2) + (t^2 + \alpha^2)^2 + 4\alpha^2 t^2 \equiv g_t(a).$$

Recalling that $\alpha \geq t$, we obtain that $g_t'(a) = 16\alpha^2 a - 8\alpha(t^2 + \alpha^2) \geq 0$ for any $a \geq \alpha$, and hence g_t is nondecreasing over $[\alpha, \infty)$. Consequently,

$$g_t(a) \geq g_t(\alpha) = 8\alpha^4 - 8\alpha^2(t^2 + \alpha^2) + (t^2 + \alpha^2)^2 + 4\alpha^2 t^2 = \alpha^4 - 2\alpha^2 t^2 + t^4 = (\alpha^2 - t^2)^2 \geq 0. \qquad \square$$

A.4. SOM property of the α-simplex. In this subsection we will show that the SOM property holds for the α-simplex set ($\alpha > 0$) given by $\Delta_n(\alpha) = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{e}^T\mathbf{x} = \alpha, \mathbf{x} \geq \mathbf{0}\}$. We begin by recalling the form of the orthogonal projection onto $\Delta_n(\alpha)$.

Lemma A.8 (projection onto the α-simplex set [1, Section 12.3.6]). *Let $\mathbf{x} \in \mathbb{R}^n$. Then the orthogonal projection onto $\Delta_n(\alpha)$ is given by $P_B(\mathbf{x}) = [\mathbf{x} + \gamma \mathbf{e}]_+$, where γ satisfies $\sum_{l=1}^n [x_l + \gamma]_+ = \alpha$.*

Obviously, for any $\gamma \in \mathbb{R}$ it holds that $x_{\sigma(1)} + \gamma \geq x_{\sigma(2)} + \gamma \geq \cdots \geq x_{\sigma(n)} + \gamma$. This fact suggests the following corollary.

Corollary A.9. *Let $\mathbf{x} \in \mathbb{R}^n$ and $\sigma \in \tilde{\Sigma}(\mathbf{x})$. Suppose that γ satisfies $\sum_{l=1}^n [x_l + \gamma]_+ = \alpha$. Then there exists $k \in \{1, \ldots, n\}$ such that $\gamma = \frac{1}{k}\left(\alpha - \sum_{l=1}^k x_{\sigma(l)}\right)$, and $x_{\sigma(1)} + \gamma, \ldots, x_{\sigma(k)} + \gamma > 0; x_{\sigma(k+1)} + \gamma, \ldots, x_{\sigma(n)} + \gamma \leq 0$.*

Lemma A.10 (α-simplex projection properties). *Let $B = \Delta_n(\alpha)$ for some $\alpha > 0$, and let $\mathbf{x} \in \mathbb{R}^n, \sigma \in \tilde{\Sigma}(\mathbf{x})$. Define*

$$\gamma^j = \frac{1}{j}\left(\alpha - \sum_{l=1}^{j} x_{\sigma(l)}\right), \ j \in \{1, \ldots, n\}, \ \text{and } q = \max\left\{j \in \{1, 2, \ldots, n\} : x_{\sigma(j)} + \gamma^j > 0\right\}. \tag{A.9}$$

Then

(a) for any $j \in \{1, \ldots, n-1\}$,

$$x_{\sigma(j+1)} + \gamma^{j+1} = \frac{j}{j+1}\left(x_{\sigma(j+1)} + \gamma^j\right), \tag{A.10}$$

and consequently $x_{\sigma(j+1)} + \gamma^{j+1} > 0$ if and only if $x_{\sigma(j+1)} + \gamma^j > 0$;

(b) the sequence $\{[x_{\sigma(j)} + \gamma^j]_+\}_{j=1}^n$ is non-increasing;

(c) $P_B(\mathbf{x}) = [\mathbf{x} + \gamma^q \mathbf{e}]_+$.

Proof. (a). For any $j \in \{1, \ldots, n-1\}$,

$$x_{\sigma(j+1)} + \gamma^{j+1} = \frac{1}{j+1}\left((j+1)x_{\sigma(j+1)} + \alpha - \sum_{l=1}^{j} x_{\sigma(l)} - x_{\sigma(j+1)}\right) = \frac{j}{j+1}\left(x_{\sigma(j+1)} + \gamma^j\right).$$

(b) Let $j \in \{1, \ldots, n-1\}$. If $[x_{\sigma(j+1)} + \gamma^{j+1}]_+ = 0$, then trivially $[x_{\sigma(j)} + \gamma^j]_+ \geq [x_{\sigma(j+1)} + \gamma^{j+1}]_+$. Otherwise, $[x_{\sigma(j+1)} + \gamma^{j+1}]_+ > 0$, and consequently, by part (a), $x_{\sigma(j+1)} + \gamma^j > 0$. Combining this with the fact that $x_{\sigma(j)} \geq x_{\sigma(j+1)}$, we conclude that

$$x_{\sigma(j)} + \gamma^j \geq x_{\sigma(j+1)} + \gamma^j \geq \frac{j}{j+1}\left(x_{\sigma(j+1)} + \gamma^j\right) = x_{\sigma(j+1)} + \gamma^{j+1}.$$

(c) By Lemma A.8 and Corollary A.9, there exists $k \in \{1 \ldots, n\}$ such that $P_B(\mathbf{x}) = [\mathbf{x} + \gamma^k \mathbf{e}]_+$ and

$$x_{\sigma(1)} + \gamma^k, \ldots, x_{\sigma(k)} + \gamma^k > 0; x_{\sigma(k+1)} + \gamma^k, \ldots, x_{\sigma(n)} + \gamma^k \leq 0. \tag{A.11}$$

We will show that $k = q$. Assume by contradiction that $k \neq q$. By the definition of q, the relations in (A.11) imply that $k < q$. Thus, by part (b), $x_{\sigma(k+1)} + \gamma^{k+1} \geq x_{\sigma(q)} + \gamma^q > 0$, and hence, by (A.10), it follows that $x_{\sigma(k+1)} + \gamma^k > 0$, which is a contradiction to (A.11). ∎

Lemma A.10 suggests that the projection onto the α-simplex might be sparse even without a sparsity constraint. In the context of the sparse projection sequence, this implies that there might be a minimal sparsity level from which all the sparse projections are equal. That sparsity level is equal to the sparsity level of the full-dimension projection (the n-sparse projection vector). By utilizing this insight together with the definition of the i-sparse projection vector as $P_B^\sigma(\mathbf{x}; i) = \mathbf{U}_T P_{B_T}(\mathbf{x}_T)$ where $T = S_i^\sigma$, it is easy to show that the formula for the projection onto the α-simplex set can be extended to compute the sparse projection sequence of the α-simplex in the following way.

Lemma A.11. *Let $B = \Delta_n(\alpha)$ for some $\alpha > 0$. Let $\mathbf{x} \in \mathbb{R}^n$ and $\sigma \in \tilde{\Sigma}(\mathbf{x})$. Let q and γ^i be defined as in (A.9). Then for any $i \in \{1, 2, \ldots, n\}$,*

$$P_B^\sigma(\mathbf{x}; i) = \begin{cases} \mathbf{x}_{\langle i\rangle_\sigma} + \gamma^i \mathbf{e}_{\langle i\rangle_\sigma}, & i \leq q, \\ [\mathbf{x} + \gamma^q \mathbf{e}]_+, & i > q. \end{cases}$$

We next present a direct consequence of the last lemma characterizing the point from which the projection sequence onto the α-simplex becomes fixed.

Corollary A.12. *Let $B = \Delta_n(\alpha)$ for some $\alpha > 0$. Let $\mathbf{x} \in \mathbb{R}^n$ and $\sigma \in \tilde{\Sigma}(\mathbf{x})$. Let q be defined as in (A.9). Then $i \geq q$ if and only if $P_B^\sigma(\mathbf{x}; i+1) = P_B^\sigma(\mathbf{x}; i)$.*

Proof. If $i \geq q$, then by Lemma A.11, $P_B^\sigma(\mathbf{x}; i) = P_B^\sigma(\mathbf{x}; i+1) = [\mathbf{x} + \gamma^q \mathbf{e}]_+$. If $i < q$, then $P_B^\sigma(\mathbf{x}; i)_{\sigma(i+1)} = 0$ and by Lemma A.10(b), $P_B^\sigma(\mathbf{x}; i+1)_{\sigma(i+1)} = [x_{\sigma(i+1)} + \gamma^{i+1}]_+ \geq [x_{\sigma(q)} + \gamma^q]_+ > 0$, and hence $P_B^\sigma(\mathbf{x}; i) \neq P_B^\sigma(\mathbf{x}; i+1)$. ∎

The claim that the α-simplex satisfies the SOM property will now be stated and proved.

Theorem A.13 (SOM property of the α-simplex). *Let $B = \Delta_n(\alpha)$. Then B satisfies the SOM property.*

Proof. Let $i \in \{1, 2, \ldots, n-2\}$ and $\mathbf{x} \in \mathbb{R}^n, \sigma \in \tilde{\Sigma}(\mathbf{x})$. We will prove that (A.1) holds. Let q be defined as in (A.9). If $q < i+2$, then by Corollary A.12, $P_B^\sigma(\mathbf{x}; i+2) = P_B^\sigma(\mathbf{x}; i+1)$, and consequently (A.1) trivially holds. We will hereafter assume that $q \geq i+2$, which in particular implies by Lemma A.11 that $P_B^\sigma(\mathbf{x}; l) = \mathbf{x}_{\langle l\rangle_\sigma} + \gamma^l \mathbf{e}_{\langle l\rangle_\sigma}$ for $l = i, i+1, i+2$. Consequently, for $j \in \{i, i+1\}$,

$$\begin{aligned} &\|\mathbf{x} - P_B^\sigma(\mathbf{x}; j)\|_2^2 - \|\mathbf{x} - P_B^\sigma(\mathbf{x}; j+1)\|_2^2 \\ &= \|\mathbf{x} - \mathbf{x}_{\langle j\rangle_\sigma} - \gamma^j \mathbf{e}_{\langle j\rangle_\sigma}\|_2^2 - \|\mathbf{x} - \mathbf{x}_{\langle j+1\rangle_\sigma} - \gamma^{j+1}\mathbf{e}_{\langle j+1\rangle_\sigma}\|_2^2 \\ &= x_{\sigma(j+1)}^2 + j(\gamma^j)^2 - (j+1)(\gamma^{j+1})^2. \end{aligned}$$

By (A.10) we have that $\gamma^j = \frac{j+1}{j}\gamma^{j+1} + \frac{1}{j}x_{\sigma(j+1)}$, and subsequently,

$$\begin{aligned} &\|\mathbf{x} - P_B^\sigma(\mathbf{x}; j)\|_2^2 - \|\mathbf{x} - P_B^\sigma(\mathbf{x}; j+1)\|_2^2 \\ &= x_{\sigma(j+1)}^2 + j\left(\frac{j+1}{j}\gamma^{j+1} + \frac{1}{j}x_{\sigma(j+1)}\right)^2 - (j+1)(\gamma^{j+1})^2 \\ &= \frac{j+1}{j}(x_{\sigma(j+1)} + \gamma^{j+1})^2. \end{aligned}$$

Thus, what we need to prove is that $\frac{i+1}{i}\left(x_{\sigma(i+1)} + \gamma^{i+1}\right)^2 \geq \frac{i+2}{i+1}\left(x_{\sigma(i+2)} + \gamma^{i+2}\right)^2$, which is a valid inequality since $\frac{i+1}{i} > \frac{i+2}{i+1}$ and $x_{\sigma(i+1)} + \gamma^{i+1} \geq x_{i+2} + \gamma^{i+2} \geq x_q + \gamma^q > 0$. ∎

A.5. SOM property of the full α-simplex and the ℓ_1 α-ball. In this subsection we will prove that the SOM property holds for the full α-simplex set and the ℓ_1 α-ball ($\alpha > 0$). The full α-simplex set is given by $\Delta_n^F(\alpha) = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{e}^T\mathbf{x} \leq \alpha, \mathbf{x} \geq \mathbf{0}\}$. By noting that $\Delta_n^F(\alpha) = D \cap \mathbb{R}_+^n$ where $D = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_1 \leq \alpha\}$, the projection onto $\Delta_n^F(\alpha)$ can be derived using Lemma A.3 and simple convex optimization arguments.

Lemma A.14 (full α-simplex projection). *Let $B = \Delta_n^F(\alpha)$, where $\alpha > 0$. Denote $J = \{l : x_l > 0\}$.*

Then

$$P_B(\mathbf{x}) = \begin{cases} \mathbf{U}_J\mathbf{x}_J, & \textit{if } \mathbf{e}_J^T\mathbf{x}_J \le \alpha \\ \mathbf{U}_J P_{\Delta_{|J|}(\alpha)}(\mathbf{x}_J), & \textit{if } \mathbf{e}_J^T\mathbf{x}_J > \alpha. \end{cases} \tag{A.12}$$

Proof. By the fact that $\Delta_n^F(\alpha) = D \cap \mathbb{R}_+^n$ where $D = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_1 \le \alpha\}$ is absolutely symmetric, Lemma A.3 implies that

$$P_B(\mathbf{x}) = \mathbf{U}_J P_{\Delta_{|J|}^F(\alpha)}(\mathbf{x}_J), \tag{A.13}$$

where by definition

$$P_{\Delta_{|J|}^F(\alpha)}(\mathbf{x}_J) = \operatorname{argmin}\left\{\|\mathbf{u} - \mathbf{x}_J\|_2 : \mathbf{e}^T\mathbf{u} \le \alpha, \mathbf{u} \ge \mathbf{0}\right\}. \tag{A.14}$$

Since $\mathbf{x}_J > \mathbf{0}$, if $\mathbf{e}^T\mathbf{x}_J \le \alpha$, then $\mathbf{x}_J \in \Delta_{|J|}^F(\alpha)$, and subsequently

$$P_{\Delta_{|J|}^F(\alpha)}(\mathbf{x}_J) = \mathbf{x}_J. \tag{A.15}$$

Otherwise, $\mathbf{e}^T\mathbf{x}_J > \alpha$ and thus $\mathbf{x}_J \notin \Delta_{|J|}^F(\alpha)$. We will show that in this case, $\mathbf{e}^T P_{\Delta_{|J|}^F(\alpha)}(\mathbf{x}_J) = \alpha$ must hold. Suppose in contradiction that $\tilde{\mathbf{x}} = P_{\Delta_{|J|}^F(\alpha)}(\mathbf{x}_J)$ satisfies $\mathbf{e}^T\tilde{\mathbf{x}} < \alpha$. Then $\mathbf{z}_\beta = \beta\mathbf{x}_J + (1-\beta)\tilde{\mathbf{x}} \ge 0$ for any $\beta \in (0,1)$, and $\mathbf{e}^T\mathbf{z}_\beta = \mathbf{e}^T\tilde{\mathbf{x}} + \beta(\mathbf{e}^T\mathbf{x}_J - \mathbf{e}^T\tilde{\mathbf{x}}) \le \alpha$ for all $\beta \le \frac{\alpha - \mathbf{e}^T\tilde{\mathbf{x}}}{\mathbf{e}^T\mathbf{x}_J - \mathbf{e}^T\tilde{\mathbf{x}}} = \beta_1$. In particular, $\mathbf{z}_{\beta_1} \in \Delta_{|J|}^F(\alpha)$, and in addition $\|\mathbf{z}_{\beta_1} - \mathbf{x}_J\|_2 = (1-\beta_1)\|\tilde{\mathbf{x}} - \mathbf{x}_J\|_2 < \|\tilde{\mathbf{x}} - \mathbf{x}_J\|_2$, which is a contradiction to the fact that $\tilde{\mathbf{x}} = P_{\Delta_{|J|}^F(\alpha)}(\mathbf{x}_J)$. We have thus shown that if $\mathbf{e}^T\mathbf{x}_J > \alpha$ then $\mathbf{e}^T P_{\Delta_{|J|}^F(\alpha)}(\mathbf{x}_J) = \alpha$, and consequently, (A.14) is equivalent to

$$P_{\Delta_{|J|}^F(\alpha)}(\mathbf{x}_J) = \operatorname{argmin}\left\{\|\mathbf{u} - \mathbf{x}_J\|_2 : \mathbf{e}^T\mathbf{u} = \alpha, \mathbf{u} \ge \mathbf{0}\right\} = P_{\Delta_{|J|}(\alpha)}(\mathbf{x}_J). \tag{A.16}$$

Plugging (A.15) and (A.16) into (A.13), we obtain the desired formula (A.12). ∎

The SOM property of the full α-simplex will now be proved.

Theorem A.15 (SOM property of the full α-simplex). *Let $B = \Delta_n^F(\alpha)$. Then B satisfies the SOM property.*

Proof. Let $\mathbf{x} \in \mathbb{R}^n, \sigma \in \tilde{\Sigma}(\mathbf{x})$, and $i \in \{0, 1, \ldots, n-2\}$. Denote $J = \{l : x_l > 0\}$. We will prove that (A.1) holds. Combining Lemma A.14 and Lemma A.3, we have that for any $j \in \{0, 1, \ldots, n\}$ the j-sparse projection vector is given by

$$P_B^\sigma(\mathbf{x}; j) = \begin{cases} \mathbf{x}_{\langle r_j\rangle_\sigma}, & \text{if } \mathbf{e}_{\langle r_j\rangle_\sigma}^T \mathbf{x}_{\langle r_j\rangle_\sigma} \le \alpha, \\ P_{\Delta_n(\alpha)}^\sigma(\mathbf{x}; r_j), & \text{if } \mathbf{e}_{\langle r_j\rangle_\sigma}^T \mathbf{x}_{\langle r_j\rangle_\sigma} > \alpha, \end{cases}$$

where $r_j = \min\{j, |J|\}$. If $r_{i+2} \ne i+2$, then $r_{i+2} = |J|$ and $|J| \le i+1$. Thus, $r_{i+1} = |J|$ as well, and consequently, $P_B^\sigma(\mathbf{x}; i+2) = P_B^\sigma(\mathbf{x}; i+1)$, which implies that the required inequality (A.1) trivially holds. We will hereafter assume that $P_B^\sigma(\mathbf{x}; i+2) \ne P_B^\sigma(\mathbf{x}; i+1)$, and subsequently $r_{i+2} = i+2$, implying that $|J| \ge i+2$ and $x_{\sigma(i)} \ge x_{\sigma(i+1)} \ge x_{\sigma(i+2)} > 0$. Several cases will now be addressed.

Suppose that $\mathbf{e}_{\langle i+1\rangle_\sigma}^T \mathbf{x}_{\langle i+1\rangle_\sigma} \le \alpha$. Then for $j \in \{i, i+1\}$, $P_B^\sigma(\mathbf{x}; j) = \mathbf{x}_{\langle j\rangle_\sigma}$, and thus (A.1) is equivalent to $x_{\sigma(i+1)}^2 \ge x_{\sigma(i+2)}^2 - \|\mathbf{x}_{\langle i+2\rangle_\sigma} - P_B^\sigma(\mathbf{x}; i+2)\|_2^2$, which is a valid inequality since $x_{\sigma(i+1)} \ge x_{\sigma(i+2)} > 0$.

Suppose that $\mathbf{e}_{\langle i\rangle_\sigma}^T \mathbf{x}_{\langle i\rangle_\sigma} \leq \alpha$ and $\mathbf{e}_{\langle i+1\rangle_\sigma}^T \mathbf{x}_{\langle i+1\rangle_\sigma} > \alpha$. Then

$$P_B^\sigma(\mathbf{x}; i) = \mathbf{x}_{\langle i\rangle_\sigma},\ P_B^\sigma(\mathbf{x}; i+1) = P_{\Delta_n(\alpha)}^\sigma(\mathbf{x}; i+1),\ P_B^\sigma(\mathbf{x}; i+2) = P_{\Delta_n(\alpha)}^\sigma(\mathbf{x}; i+2). \tag{A.17}$$

Therefore, by the underlying assumption that $P_B^\sigma(\mathbf{x}; i+2) \neq P_B^\sigma(\mathbf{x}; i+1)$, we have that

$$P_{\Delta_n(\alpha)}^\sigma(\mathbf{x}; i+2) \neq P_{\Delta_n(\alpha)}^\sigma(\mathbf{x}; i+1). \tag{A.18}$$

Corollary A.12 in view of (A.18) implies that q defined in (A.9) satisfies $q \geq i+2$. Thus, for $j \in \{i+1, i+2\}$ it holds that

$$P_B^\sigma(\mathbf{x}; j) = \mathbf{x}_{\langle j\rangle_\sigma} + \gamma^j \mathbf{e}_{\langle j\rangle_\sigma} = \mathbf{x}_{\langle j\rangle_\sigma} + \frac{1}{j}\left(\alpha - \sum_{l=1}^{j} x_{\sigma(l)}\right)\mathbf{e}_{\langle j\rangle_\sigma}. \tag{A.19}$$

Denote $t = \alpha - \sum_{l=1}^{i+1} x_{\sigma(l)}$. By equations (A.17) and (A.19), we have the following equalities: $\|\mathbf{x} - P_B^\sigma(\mathbf{x}; i)\|_2^2 - \|\mathbf{x} - P_B^\sigma(\mathbf{x}; i+1)\|_2^2 = x_{\sigma(i+1)}^2 - \frac{1}{i+1}t^2$, and $\|\mathbf{x} - P_B^\sigma(\mathbf{x}; i+1)\|_2^2 - \|\mathbf{x} - P_B^\sigma(\mathbf{x}; i+2)\|_2^2 = x_{\sigma(i+2)}^2 + \frac{1}{i+1}t^2 - \frac{1}{i+2}\left(t - x_{\sigma(i+2)}\right)^2$. Hence, in order to prove the required we need to show that $x_{\sigma(i+1)}^2 - \frac{2}{i+1}t^2 - x_{\sigma(i+2)}^2 + \frac{1}{i+2}\left(t - x_{\sigma(i+2)}\right)^2 \geq 0$. By rearranging terms and multiplying by $(i+2)$, the above inequality can be rewritten as

$$(i+1)(x_{\sigma(i+1)}^2 - x_{\sigma(i+2)}^2) + (x_{\sigma(i+1)}^2 - t^2) - 2t\left(\frac{t}{i+1} + x_{\sigma(i+2)}\right) \geq 0. \tag{A.20}$$

The non-negativity of the first term trivially follows from the standing assumption that $x_{\sigma(i+1)} \geq x_{\sigma(i+2)} > 0$. For the second term, since $x_{\sigma(i+1)} > 0$ and $t < 0$, we have that

$$x_{\sigma(i+1)}^2 - t^2 = (x_{\sigma(i+1)} + |t|)\left(x_{\sigma(i+1)} + \alpha - \sum_{l=1}^{i+1} x_{\sigma(l)}\right) = (x_{\sigma(i+1)} + |t|)\left(\alpha - \sum_{l=1}^{i} x_{\sigma(l)}\right) \geq 0,$$

where the last inequality follows from the standing assumption that $\mathbf{e}_{\langle i\rangle_\sigma}^T \mathbf{x}_{\langle i\rangle_\sigma} \leq \alpha$. For the third term, we have that

$$\begin{aligned}
-2t\left(\frac{t}{i+1} + x_{\sigma(i+2)}\right) &= -\frac{2t(i+2)}{i+1}\left(x_{\sigma(i+2)} + \frac{1}{i+2}\left(\alpha - \sum_{l=1}^{i+2} x_{\sigma(l)}\right)\right) \\
&= -\frac{2t(i+2)}{i+1} P_{\Delta_n(\alpha)}(\mathbf{x}; i+2)_{\sigma(i+2)} \geq 0,
\end{aligned}$$

where the second equality follows from (A.19), and the third inequality from the fact that $t < 0$ and $P_{\Delta_n(\alpha)}(\mathbf{x}; i+2)_{\sigma(i+2)} \geq 0$ (as $\Delta_n(\alpha)$ is nonnegative). Thus, (A.20) holds, and (A.1) is satisfied in this case.

Finally, suppose that $\mathbf{e}_{\langle i\rangle_\sigma}^T \mathbf{x}_{\langle i\rangle_\sigma} > \alpha$. Then $P_B^\sigma(\mathbf{x}; j) = P_{\Delta_n(\alpha)}^\sigma(\mathbf{x}; j)$ for $j \in \{i, i+1, i+2\}$, and consequently (A.1) holds if and only if it holds for $B = \Delta_n(\alpha)$. By Theorem A.13 the set $\Delta_n(\alpha)$ satisfies the SOM property, and in particular required inequality (A.1). $\square$

Since $\Delta_n^F(\alpha) = D \cap \mathbb{R}_+^n$ where $D = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_1 \leq \alpha\}$, Lemma A.2 together with Theorem A.15 readily imply that the ℓ_1 α-ball satisfies the SOM property.

Theorem A.16 (SOM property of the ℓ_1 ball). *Let $B = B_1[\mathbf{0}, \alpha] = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_1 \leq \alpha\}$, where $\alpha > 0$. Then B satisfies the SOM property.*