Optimization Problems Involving Group Sparsity Terms

Amir Beck* Nadav Hallak†

May 18, 2017

Abstract

This paper studies a general form problem in which a lower bounded continuously differentiable function is minimized over a block separable set incorporating a group sparsity expression as a constraint or a penalty (or both) in the group sparsity setting. This class of problems is generally hard to solve, yet highly applicable in numerous practical settings. Particularly, we study the proximal mapping that includes group-sparsity terms, and derive an efficient method to compute it. Necessary optimality conditions for the problem are devised, and a hierarchy between stationary-based and coordinate-wised based conditions is established. Methods that obtain points satisfying the optimality conditions are presented, analyzed and tested in applications from the fields of investment and graph theory.

1 Introduction

The group-sparsity setting, in which the decision vector's components are grouped together into several distinguishable index sets, has been extensively researched in recent years due to its applicability in numerous fields and practical problems, such as signal and image processing, compressed sensing (CS), gene selection and analysis, and many more (see the provided citations below and therein).

In statistical and machine learning literature ([23, 25, 30], [20],[27, chapter 2] and references therein) the group structure is imperative in many cases in which the explanatory variables demonstrate high correlative nature that can be used to classify them into groups, or a priory belong to categories based on the properties of the model. In these fields, the problem studied usually belongs to the family of regression models (linear, logistic, etc.) and the sparsity is induced using ℓ_1-norm type convex relaxations, also known as group lasso-type models.

In the setting of compressed sensing, group-sparsity is referred to as 'block-sparsity', and is a particular example of 'structured sparsity' [15]. Block-sparsity generalizes the standard sparse-signal model (see the in-depth reviews [13, 15, 16, 29]), making it suitable for dealing with a

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

†Faculty of Industrial Engineering and Management, Technion - Israel Institute of Technology, Haifa, Israel. Email: nadav_hallak@outlook.com.

larger family of problems and can lead to improved performances; for example, it was shown that the block-sparse structure enables signal recovery from a reduced number of CS measurements, as presented in [18, 28]. Many results and notions known for the standard sparse model were generalized to the block-sparse settings. Among them are the well-known *restricted isometry property* (RIP) ([2, 10, 18]), block convex relaxation techniques ([18, 17]), and several ℓ_0-based methods (the ℓ_0-norm counts the number of nonzero blocks) such as the generalized CoSaMP and block IHT in [2, 14]. For an additional comprehensive discussion on this topic in the CS setting, the reader can refer to [15, section V, part B] and references therein.

Recently, the two papers [1] and [19] studied the setting of group-sparsity with possibly overlapping groups, and with no other constraints other than the group-sparsity bound. In [1], the authors studied the orthogonal projection operator for different structures of the overlapping groups. In [19], the linear least-squares problem and a greedy-IHT method that solves it were studied in the context of some restrictive RIP-like condition and distribution assumptions.

Our research deals with group-sparsity of non-overlapping groups, in problems consisting of minimizing a continuously differentiable objective function over a set composed of blocks corresponding to the groups' partition, such that each block is constrained to be in the the union of a closed and possibly convex set with the zeros vector. The group-sparsity term (which is a discrete function) appears both in the objective function and in the constraints, forming a versatile model that can be tuned to suit any non-overlapping group-sparsity problem. We do not assume that all the blocks are of the same size or that they are constrained to be at the same set.

In this work, we will mainly address two topics; the computation of the proximal mapping with respect to the group-sparsity term, and optimality conditions of the general form problem. These two topics were tied together and studied in the series of papers [4, 6, 5] dealing with the problem of minimizing a general continuously differentiable function with a vector-sparsity penalty or constraint term over a symmetric set (such as the ℓ_p-norm ball or the unit simplex). With no additional constraints other than the sparsity of the solution, [4] provided the foundations for the following papers in the series by establishing a hierarchy between stationary-based optimality conditions and coordinate-based optimality conditions, and developing methods to obtain points satisfying these conditions. In [6], the constraint of belonging to a symmetric set (together with sparsity) was added to the problem, thus generalizing the results of [4]. Efficient methods for computing sparse orthogonal projections under various symmetry assumptions were devised, and a more general hierarchy between stationary-based and coordinate-based optimality conditions was proved. Methods for generating points satisfying the various optimality conditions were also provided and analyzed. The paper [5] studied a class of problems consisting of minimizing a continuously differentiable function penalized with the sparsity term over a symmetric set.

Roughly speaking, the conclusions from the above mentioned papers are that although sparsity can render a problem hard to solve, the proximal mapping involving a sparsity term can be computed efficiently in many standard problems as long as the underlying set possesses a certain symmetry property, and that optimality conditions based on stationarity are less restrictive than optimality conditions based on a coordinate-wise comparison. These two conclusions will be established in our group-sparsity setting as well, as we will prove a closed form explicit formula of

the proximal mapping which will lead to an efficient procedure to compute it, and will establish a hierarchy of optimality conditions in which the coordinate-wise conditions are more restrictive than the stationarity condition based on the proximal gradient operator. We note that in this paper we do not assume any symmetry assumptions on the underlying set, and heavily utilize its block separable structure. Therefore, the techniques used in this paper are completely different from those used in [4, 6, 5].

Paper layout. We first properly formulate the problem and setting in Section 1.1. In Section 2 we recall necessary mathematical preliminaries – stationarity in smooth problems over convex sets, and important results for the class of functions with Lipshcitz continuous gradient. Section 3 studies the group-sparse proximal mapping, providing a characterization of the proximal mapping, and deriving an efficient procedure that obtain it. The results of Section 3 are then used in Section 4 to develop necessary optimality conditions for the underlying problem, and to prove their hierarchy. Methods that generate points satisfying the latter are devised in Section 5. Finally, in Section 6, our results are demonstrated on an investment problem as well as on a cardinality constrained maximum weight clique problem.

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 $\mathbf{1}$. For a vector $\mathbf{x} \in \mathbb{R}^n$, the vector $|\mathbf{x}|$ is the vector of absolute values of the components of $\mathbf{x}$.

Given a function $h : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$, the proximal mapping of $\mathbf{x}$ with respect to h is defined as

$$\mathrm{prox}_h(\mathbf{x}) = \operatorname*{argmin}_{\mathbf{u}} \left\{ h(\mathbf{u}) + \frac{1}{2}\|\mathbf{u} - \mathbf{x}\|_2^2 \right\}.$$

This concept was introduced and studied extensively by Moreau [24]. When h is not convex, the proximal mapping might return multiple vectors and should therefore be considered as a multivalued mapping.

The indicator function of a given set $C \subseteq \mathbb{R}^n$ is denoted by δ_C and is given by $\delta_C(\mathbf{x}) = 0$ for $\mathbf{x} \in C$ and ∞ otherwise. The proximal mapping of the indicator function δ_C amounts to the orthogonal projection mapping onto C: for a given set $C \subseteq \mathbb{R}^n$, the orthogonal projection of $\mathbf{x}$ onto S is defined as

$$P_C(\mathbf{x}) \equiv \mathrm{prox}_{\delta_C}(\mathbf{x}) = \mathrm{argmin}\{\|\mathbf{y} - \mathbf{x}\|_2^2 : \mathbf{y} \in C\}.$$

The norm notation $\|\cdot\|$ without any subscript stands for the l_2-norm. The so-called l_0-norm which counts the number of nonzero elements in the vector is defined by $\|\mathbf{x}\|_0 \equiv |\{i : x_i \neq 0\}|$. 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. Given a vector $\mathbf{v} \in \mathbb{R}^m$, $v_{[i]}$ denotes the ith largest value in $\mathbf{v}$. In particular, $v_{[1]} \geq v_{[2]} \geq \cdots \geq v_{[m]}$. The set $S_j(\mathbf{v})$ comprises all the index sets containing the j-largest elements in $\mathbf{v}$. The set $S_j(\mathbf{v})$ is not necessarily a singleton as $\mathbf{v}$ might contain identical values. For example, if $\mathbf{v} = (1, 1)^T$, then $S_1(\mathbf{v}) = \{\{1\}, \{2\}\}$.

1.1 Problem Formulation

In order to properly formulate the problem, we first require to define some group-related notation.

1.1.1 Groups Notation

Throughout the paper we will regard $\{G_i\}_{i=1}^m$ as a predetermined partition of $\{1, 2, \ldots, n\}$ comprising m groups of sizes $n_1, n_2, \ldots, n_m$ respectively. Without loss of generality, we will hereafter assume that the groups are given by

$$G_1 = \{1, 2, \ldots, n_1\}, G_2 = \{n_1 + 1, n_1 + 2, \ldots, n_1 + n_2\}, \ldots, G_m = \{n_{m-1} + 1, n_{m-1} + 2, \ldots, n\}.$$

The mapping $g : \mathbb{R}^n \to \{0, 1\}^m$ that indicates which groups contain indices corresponding to nonzero components in a given vector is defined for any $i = 1, 2, \ldots, m$ by

$$g(\mathbf{x})_i = \begin{cases} 1, & \mathbf{x}_{G_i} \neq \mathbf{0}, \\ 0, & \text{otherwise.} \end{cases} \tag{1.1}$$

For example, if $G_1 = \{1, 2\}$ and $G_2 = \{3\}$, then

$$g\left((0, 0, 5)^T\right) = (0, 1)^T,\ g\left((0, 5, 0)^T\right) = (1, 0)^T,\ g\left((5, 0, 5)^T\right) = (1, 1)^T.$$

A group G_i will be called *active* at a point $\mathbf{x} \in \mathbb{R}^n$ if $g(\mathbf{x})_i = 1$, and *inactive* otherwise. The set of all real vectors with at most $s \in \{1, 2, \ldots, m\}$ active groups will be denoted by C_s:

$$C_s = \{\mathbf{x} \in \mathbb{R}^n : \|g(\mathbf{x})\|_0 \leq s\}. \tag{1.2}$$

When the groups are singletons, C_s amounts to the set of all s-sparse vectors. The set of the active groups of a vector $\mathbf{x} \in \mathbb{R}^n$ will be called the *group-support of* $\mathbf{x}$, and is defined by $I_1(\mathbf{x}) = \{i \in \{1, 2, \ldots, m\} : g(\mathbf{x})_i = 1\}$.

The operator $\mathcal{A} : 2^{\{1,2,\ldots,m\}} \to 2^{\{1,2,\ldots,n\}}$ that returns the set comprising all indices in the groups in a given index set is defined by

$$\mathcal{A}(T) \equiv \bigcup_{i \in T} G_i.$$

For example, if $G_1 = \{1, 2\}$, $G_2 = \{3, 4\}$, and $G_3 = \{5, 6\}$, then

$$\mathcal{A}(\{1\}) = \{1, 2\},\ \mathcal{A}(\{1, 3\}) = \{1, 2, 5, 6\},\ \mathcal{A}(\{1, 2, 3\}) = \{1, 2, 3, 4, 5, 6\}.$$

We can now properly formulate the problem.

1.1.2 Problem Formulation

In this paper we study the following optimization problem:

$$\begin{aligned} \min \quad & f(\mathbf{x}) + \lambda \|g(\mathbf{x})\|_0 \\ \text{s.t.} \quad & \mathbf{x} \in C_s \cap B, \end{aligned} \tag{P}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a lower bounded continuously differentiable function, $\lambda \geq 0$ is a penalty parameter on the number of active groups, $s \in \{1, 2, \ldots, m\}$ is an upper bound on the number of active groups, and the set $B \subseteq \mathbb{R}^n$ is defined by

$$B \equiv \prod_{i=1}^{m} (D_i \cup \{\mathbf{0}\}) = (D_1 \cup \{\mathbf{0}\}) \times (D_2 \cup \{\mathbf{0}\}) \times \cdots \times (D_m \cup \{\mathbf{0}\}), \tag{1.3}$$

where $D_i \subseteq \mathbb{R}^{n_i}$ is a nonempty closed set for any i. In Section 4 we will add the assumption that D_i is also convex for any $i = 1, 2, \ldots, m$. Denoting $h(\mathbf{x}) \equiv \lambda \|g(\mathbf{x})\|_0 + \delta_{B \cap C_s}(\mathbf{x})$, problem (P) can be rewritten as

$$\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) + h(\mathbf{x}).$$

We will sometimes refer to a restriction of the set $\prod_{i=1}^{m} D_i$ to an index set $T \subseteq \{1, 2, \ldots, m\}$ by using the operator $\mathcal{B}$ defined by

$$\mathcal{B}(T) \equiv \prod_{i \in T} D_i.$$

Following are four simple examples for minimization problems which are special cases of the general model (P).

Example 1.1 (group-sparsity constrained minimization)**.** In this problem $D_i \equiv \mathbb{R}^{n_i}$, $s < m$, and $\lambda = 0$.

$$\min_{\mathbf{x}} \{ f(\mathbf{x}) \ : \ \mathbf{x} \in C_s \}.$$

Example 1.2 (group-sparsity penalized minimization)**.** In this problem $D_i \equiv \mathbb{R}^{n_i}$, $s = m$, and $\lambda > 0$.

$$\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) + \lambda \|g(\mathbf{x})\|_0.$$

Example 1.3 (double-sparsity constrained minimization)**.** This model incorporates group-sparsity and sparsity within each group. In this problem $D_i \equiv \{\mathbf{y} \in \mathbb{R}^{n_i} : \|\mathbf{y}\|_0 \leq s_i\}$ where $s_i \leq n_i$ is the sparsity level within each group, $s < m$, and $\lambda = 0$.

$$\min_{\mathbf{x}} \{ f(\mathbf{x}) \ : \ \mathbf{x} \in C_s, \ \|\mathbf{x}_{G_i}\|_0 \leq s_i \quad \forall i = 1, 2, \ldots, m \}.$$

Example 1.4 (binary constrained minimization)**.** In this model $D_i \equiv \{1\}$ for any $i = 1, 2, \ldots, m = n$, $s < m$, and $\lambda = 0$, which results with a minimization problem over n-length binary vectors with a cardinality constraint.

$$\min_{\mathbf{x} \in \{0,1\}^n} \left\{ f(\mathbf{x}) \ : \ \sum_{i=1}^{n} x_i \leq s \right\}.$$

The necessary notation and assumptions regarding the optimization problem discussed in this paper are summarized in the following.

Standing notation and assumptions (made throughout the paper)

- $g : \mathbb{R}^n \to \mathbb{R}$ is given in (1.1).
- $s \in \{1, 2, \ldots, m\}$ is given and $C_s = \{\mathbf{x} \in \mathbb{R}^n : \|g(\mathbf{x})\|_0 \leq s\}$.
- $D_1, D_2, \ldots, D_m$ are nonempty closed sets and B is given in (1.3).
- $\lambda \geq 0$ is given.
- $h(\mathbf{x}) \equiv \lambda\|g(\mathbf{x})\|_0 + \delta_{B\cap C_s}(\mathbf{x})$.
- $f : \mathbb{R}^n \to \mathbb{R}$ is a lower bounded continuously differentiable function.

2 Mathematical Preliminaries

2.1 Stationarity in Smooth Problems over Convex Sets

We begin by recalling the notion of stationarity in problems comprising the minimization of smooth functions over closed and convex sets. Consider the problem

$$\min\{f_0(\mathbf{x}) : \mathbf{x} \in C\}, \tag{2.1}$$

where $f_0 : \mathbb{R}^n \to \mathbb{R}$ is a continuously differentiable function and $C \subseteq \mathbb{R}^n$ is a nonempty closed and convex set. A vector $\mathbf{x}^*$ is called *a stationary point* of (2.1) if

$$\nabla f_0(\mathbf{x}^*)^T(\mathbf{x} - \mathbf{x}^*) \geq 0 \text{ for any } \mathbf{x} \in C. \tag{2.2}$$

This necessary optimality condition means that there are no feasible descent directions at $\mathbf{x}^*$. It is well known (see for example [3, 8]) that the condition can be rewritten as

$$\mathbf{x}^* = P_C\left(\mathbf{x}^* - \frac{1}{L}\nabla f_0(\mathbf{x}^*)\right) \tag{2.3}$$

for some $L > 0$. Even though condition (2.3) is expressed in terms of the parameter L, it is independent of L by its equivalence to condition (2.2). When the objective function f_0 is convex, then stationarity is a necessary *and sufficient* condition for optimality.

2.2 The Class of $C_L^{1,1}$ Functions

A function $f_0 : \mathbb{R}^n \to \mathbb{R}$ is said to belong to $C_L^{1,1}$ if it is continuously differentiable and its gradient is Lipschitz continuous with parameter $L > 0$, meaning that

$$\|\nabla f_0(\mathbf{x}) - \nabla f_0(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\| \text{ for all } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n.$$

An important property of $C_L^{1,1}$ functions is described in the well-known descent lemma.

Lemma 2.1 (descent lemma [8, Proposition A.24]). *Suppose that $f_0 \in C_{L_{f_0}}^{1,1}$. Then for any $\mathbf{x}, \mathbf{d} \in \mathbb{R}^n$ and $L \geq L_{f_0}$, the following inequality is satisfied:*

$$f_0(\mathbf{x}+\mathbf{d}) \leq f_0(\mathbf{x}) + \nabla f_0(\mathbf{x})^T\mathbf{d} + \frac{L}{2}\|\mathbf{d}\|^2.$$

Denote

$$h(\cdot) \equiv \lambda\|g(\cdot)\|_0 + \delta_{B\cap C_s}(\cdot). \tag{2.4}$$

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

Lemma 2.2 (sufficient decrease lemma [11, Lemma 3.2]). *Let $f_0 \in C_{L_{f_0}}^{1,1}$ and $L > L_{f_0}$. Let the functions g and h be defined in (1.1) and (2.4) respectively with B and C_s given in (1.3) and (1.2). Then for any $\lambda \geq 0, \mathbf{y} \in \mathbb{R}^n$ and $\mathbf{z} \in \operatorname{prox}_{\frac{1}{L}h}\left(\mathbf{y} - \frac{1}{L}\nabla f_0(\mathbf{y})\right)$, it holds that*

$$f_0(\mathbf{y}) + \lambda\|g(\mathbf{y})\|_0 - f_0(\mathbf{z}) - \lambda\|g(\mathbf{z})\|_0 \geq \frac{L - L_{f_0}}{2}\|\mathbf{z}-\mathbf{y}\|^2.$$

We will also be interested in a more refined version of the descent lemma, which we call *the group descent lemma*, in which the perturbation vector $\mathbf{d}$ has at most two active groups. For that, we will define the *group Lipschiz constant*. Let $f_0 \in C_{L_{f_0}}^{1,1}$. Then for any $i \neq j$ there exists a constant $L_{i,j}$ for which

$$\|\nabla_{G_i\cup G_j} f_0(\mathbf{x}) - \nabla_{G_i\cup G_j} f_0(\mathbf{x}+\mathbf{d})\| \leq L_{i,j}\|\mathbf{d}\|, \tag{2.5}$$

for any $\mathbf{x} \in \mathbb{R}^n$ and any $\mathbf{d} \in \mathbb{R}^n$ for which $g(\mathbf{d})_k = 0$ for any $k \notin \{i,j\}$. The group Lipschitz constant is defined as

$$L_{f_0}^G \equiv \max_{i\neq j} L_{i,j}. \tag{2.6}$$

Clearly, we can always pick $L_{f_0}^G = L_{f_0}$, but in general the group Lipschitz constant $L_{f_0}^G$ can be much smaller than the global Lipschitz constant L_{f_0}. The group Lipschitz constant is used in a more refined version of the descent lemma.

Lemma 2.3 (group descent lemma). *Suppose that $f_0 \in C_{L_{f_0}}^{1,1}$, and that $L \geq L_{f_0}^G$. Then*

$$f_0(\mathbf{x}+\mathbf{d}) \leq f_0(\mathbf{x}) + \nabla f_0(\mathbf{x})^T\mathbf{d} + \frac{L}{2}\|\mathbf{d}\|^2$$

for any vector $\mathbf{d} \in \mathbb{R}^n$ with at most two active groups.

3 The Group-Sparse Proximal Mapping

This section is devoted to the study of the proximal mapping operator with respect to the function $h(\cdot) \equiv \lambda\|g(\cdot)\|_0 + \delta_{B\cap C_s}(\cdot)$, where $\lambda \geq 0$ and $s \in \{1, 2\ldots, m\}$. The proximal mapping

with respect to h is given by

$$\mathrm{prox}_h(\mathbf{x}) = \underset{\mathbf{u}\in B\cap C_s}{\mathrm{argmin}} \left\{ \lambda\|g(\mathbf{u})\|_0 + \frac{1}{2}\|\mathbf{u}-\mathbf{x}\|_2^2 \right\}. \tag{3.1}$$

In general, it is hard to compute prox operators of nonconvex functions, and in particular those who contain sparsity terms that induce combinatorial elements into the problem. Yet in some cases, the properties of the set can be exploited in order to obtain a solution to (3.1) efficiently, such as in [9] and [14] where the lack of constraints (other than sparsity) was exploited to compute the orthogonal projection onto the set of s-sparse/s-group-sparse real vectors; in [6], some symmetry properties of the underlying sets were exploited to compute the orthogonal projection onto the intersection of a closed convex and symmetric set and the set of s-sparse vectors; the case in which the sparsity term appears as a penalty rather than as a constraint was studied in [5], where it was shown how to compute a member of the prox mapping under similar symmetry conditions and/or submodularity-like properties related to the underlying set. We will show in this section how the proximal mapping can be evaluated in our setting as well, with no additional symmetry assumptions on the underlying set B.

Given $\mathbf{x}\in\mathbb{R}^n$, for any $j=1,2,\ldots,m$ denote

$$d_{D_j}(\mathbf{x}_{G_j}) = \min_{\mathbf{z}\in D_j} \left\|\mathbf{x}_{G_j}-\mathbf{z}\right\|_2.$$

Note that for any j, $d_{D_j}(\mathbf{x}_{G_j})$ is well-defined due to the closedness of D_j. A key component of the analysis ahead is the mapping $\omega:\mathbb{R}^n\to\mathbb{R}^m$ defined below, which we will show to have a major role in determining the identities of the active groups of the vectors in the proximal mapping. It is defined as

$$\omega(\mathbf{x})_j = \left\|\mathbf{x}_{G_j}\right\|_2^2 - d_{D_j}^2(\mathbf{x}_{G_j}), \quad j=1,2,\ldots,m.$$

The next lemma formulates the main benefit from using the mapping ω.

Lemma 3.1. *Let $\mathbf{x}\in\mathbb{R}^n$, and $T,S\subseteq\{1,2,\ldots,m\}$. For any $\mathbf{z},\mathbf{y}\in\mathbb{R}^n$ satisfying*

$$\mathbf{z}_{G_i}\in\begin{cases} P_{D_i}(\mathbf{x}_{G_i}), & i\in T,\\ \{\mathbf{0}\}, & i\notin T,\end{cases} \qquad \mathbf{y}_{G_i}\in\begin{cases} P_{D_i}(\mathbf{x}_{G_i}), & i\in S,\\ \{\mathbf{0}\}, & i\notin S.\end{cases}$$

It holds that

$$\|\mathbf{z}-\mathbf{x}\|_2^2 - \|\mathbf{y}-\mathbf{x}\|_2^2 = \sum_{i\in S}\omega(\mathbf{x})_i - \sum_{i\in T}\omega(\mathbf{x})_i. \tag{3.2}$$

Proof. By rearrangement of terms,

$$\begin{aligned}\|\mathbf{z}-\mathbf{x}\|_2^2 - \|\mathbf{y}-\mathbf{x}\|_2^2 &= \sum_{i\notin T}\|\mathbf{x}_{G_i}\|_2^2 + \sum_{i\in T}\|\mathbf{z}_{G_i}-\mathbf{x}_{G_i}\|_2^2 - \sum_{i\notin S}\|\mathbf{x}_{G_i}\|_2^2 - \sum_{i\in S}\|\mathbf{y}_{G_i}-\mathbf{x}_{G_i}\|_2^2\\ &= \sum_{i\in S\setminus T}\|\mathbf{x}_{G_i}\|_2^2 + \sum_{i\in T\setminus S}\|\mathbf{z}_{G_i}-\mathbf{x}_{G_i}\|_2^2 - \sum_{i\in T\setminus S}\|\mathbf{x}_{G_i}\|_2^2 - \sum_{i\in S\setminus T}\|\mathbf{y}_{G_i}-\mathbf{x}_{G_i}\|_2^2\\ &= \sum_{i\in S\setminus T}\omega(\mathbf{x})_i - \sum_{i\in T\setminus S}\omega(\mathbf{x})_i.\end{aligned}$$

By adding the elements in $\{\omega(\mathbf{x})_i : i \in S \cap T\}$ to each of the sums, we obtain (3.2). □

The characterization of the proximal mapping with respect to h is given by the next theorem.

Theorem 3.2 (prox characterization)**.** *Let $\mathbf{x} \in \mathbb{R}^n$. Then $\mathbf{u} \in \text{prox}_h(\mathbf{x})$ if and only if the following conditions hold:*

(a) $\mathbf{u}_{G_i} \in P_{D_i}(\mathbf{x}_{G_i})$ for any $i \in I_1(\mathbf{u})$.

(b) There exists $T \in S_s(\omega(\mathbf{x}))$ for which

$$T \cap \{j : \omega(\mathbf{x})_j > 2\lambda\} \subseteq I_1(\mathbf{u}) \subseteq T \cap \{j : \omega(\mathbf{x})_j \geq 2\lambda\}. \tag{3.3}$$

Proof. Suppose that $\mathbf{u}$ satisfies (a) and (b) for some $T \in S_s(\omega(\mathbf{x}))$, and let $\mathbf{y} \in \text{prox}_h(\mathbf{x})$. It will be shown that $\mathbf{u} \in \text{prox}_h(\mathbf{x})$ and that $\mathbf{y}$ satisfies (a) and (b) for some $\tilde{T} \in S_s(\omega(\mathbf{x}))$. Since $\mathbf{y} \in \text{prox}_h(\mathbf{x})$,

$$2\lambda(\|g(\mathbf{u})\|_0 - \|g(\mathbf{y})\|_0) + \|\mathbf{u} - \mathbf{x}\|_2^2 - \|\mathbf{y} - \mathbf{x}\|_2^2 \geq 0. \tag{3.4}$$

Obviously, $\mathbf{y}_{G_i} \in P_{D_i}(\mathbf{x}_{G_i})$ for any $i \in I_1(\mathbf{y})$ (as otherwise a better solution for the problem defining $\text{prox}_h(\mathbf{x})$ could be obtained) and we assumed that $\mathbf{u}_i \in P_{D_i}(\mathbf{x}_{G_i})$ for any $i \in I_1(\mathbf{u})$. Thus by Lemma 3.1, (3.4) is the same as

$$\sum_{i \in I_1(\mathbf{y})} (\omega(\mathbf{x})_i - 2\lambda) - \sum_{i \in I_1(\mathbf{u})} (\omega(\mathbf{x})_i - 2\lambda) \geq 0,$$

which is the same as

$$\sum_{i \in I_1(\mathbf{y}) \cap \{j : \omega(\mathbf{x})_j \geq 2\lambda\}} (\omega(\mathbf{x})_i - 2\lambda) + \sum_{i \in I_1(\mathbf{y}) \cap \{j : \omega(\mathbf{x})_j < 2\lambda\}} (\omega(\mathbf{x})_i - 2\lambda) - \sum_{i \in I_1(\mathbf{u})} (\omega(\mathbf{x})_i - 2\lambda) \geq 0. \tag{3.5}$$

By (3.3), it follows that

$$\sum_{i \in I_1(\mathbf{u})} (\omega(\mathbf{x})_i - 2\lambda) = \sum_{i \in T \cap \{j : \omega(\mathbf{x})_j \geq 2\lambda\}} (\omega(\mathbf{x})_i - 2\lambda).$$

Since $T \cap \{j : \omega(\mathbf{x})_j \geq 2\lambda\}$ contains indices of a set of s largest elements of $\omega(\mathbf{x})$ that have the value of at least 2λ and $\mathbf{y} \in C_s$, we have

$$\sum_{i \in T \cap \{j : \omega(\mathbf{x})_j \geq 2\lambda\}} (\omega(\mathbf{x})_i - 2\lambda) \geq \sum_{i \in I_1(\mathbf{y}) \cap \{j : \omega(\mathbf{x})_j \geq 2\lambda\}} (\omega(\mathbf{x})_i - 2\lambda).$$

Combining the last two relations, we get

$$\sum_{i \in I_1(\mathbf{u})} (\omega(\mathbf{x})_i - 2\lambda) \geq \sum_{i \in I_1(\mathbf{y}) \cap \{j : \omega(\mathbf{x})_j \geq 2\lambda\}} (\omega(\mathbf{x})_i - 2\lambda). \tag{3.6}$$

Utilizing the valid inequality[1]

$$\sum_{i\in I_1(\mathbf{y})\cap\{j:\omega(\mathbf{x})_j<2\lambda\}} (\omega(\mathbf{x})_i - 2\lambda) \le 0$$

along with (3.6) yields

$$\begin{aligned} 0 &\ge \sum_{i\in I_1(\mathbf{y})\cap\{j:\omega(\mathbf{x})_j\ge 2\lambda\}} (\omega(\mathbf{x})_i - 2\lambda) - \sum_{i\in I_1(\mathbf{u})} (\omega(\mathbf{x})_i - 2\lambda) \\ &\ge \sum_{i\in I_1(\mathbf{y})\cap\{j:\omega(\mathbf{x})_j\ge 2\lambda\}} (\omega(\mathbf{x})_i - 2\lambda) + \sum_{i\in I_1(\mathbf{y})\cap\{j:\omega(\mathbf{x})_j<2\lambda\}} (\omega(\mathbf{x})_i - 2\lambda) - \sum_{i\in I_1(\mathbf{u})} (\omega(\mathbf{x})_i - 2\lambda) \quad (3.7) \\ &\ge 0, \end{aligned}$$

where the third inequality follows from (3.5). Thus, the chain of inequalities in (3.7) is satisfied as a chain of equalities, and subsequently (3.4) is satisfied as an equality, which implies that $\mathbf{u} \in \mathrm{prox}_h(\mathbf{x})$. The fact that (3.7) is a chain of equalities also implies that

$$\sum_{i\in I_1(\mathbf{y})\cap\{j:\omega(\mathbf{x})_j<2\lambda\}} (\omega(\mathbf{x})_i - 2\lambda) = 0,$$

which in turn implies that

$$\sum_{i\in I_1(\mathbf{y})} (\omega(\mathbf{x})_i - 2\lambda) = \sum_{i\in I_1(\mathbf{y})\cap\{j:\omega(\mathbf{x})_j\ge 2\lambda\}} (\omega(\mathbf{x})_i - 2\lambda) = \sum_{i\in I_1(\mathbf{u})} (\omega(\mathbf{x})_i - 2\lambda).$$

Hence, by the validity of (a) and (b) and the fact that $\mathbf{y} \in C_s$, $\mathbf{y}$ must also satisfy (b) (with $\mathbf{u}$ replaced by $\mathbf{y}$). Finally, as was already noted, $\mathbf{y}_{G_i} \in P_{D_i}(\mathbf{x}_{G_i})$ for any $i \in I_1(\mathbf{y})$, showing that (a) also holds for $\mathbf{y}$. □

Theorem 3.2 provides a characterization of the proximal mapping set (3.1) that can be written as a closed-form formula[2].

$$\mathrm{prox}_h(\mathbf{x}) = \begin{cases} \left\{\mathbf{U}_{\mathcal{A}(T)}\mathbf{y} : \mathbf{y} \in P_{\mathcal{B}(T)}(\mathbf{x}_{\mathcal{A}(T)}),\ I_+(\mathbf{x}) \subseteq T \subseteq I_+(\mathbf{x}) \cup I_?(\mathbf{x}),\ |T| = s\right\}, & \omega(\mathbf{x})_{[s]} > 2\lambda, \\ \left\{\mathbf{U}_{\mathcal{A}(T)}\mathbf{y} : \mathbf{y} \in P_{\mathcal{B}(T)}(\mathbf{x}_{\mathcal{A}(T)}),\ I_+(\mathbf{x}) \subseteq T \subseteq I_+(\mathbf{x}) \cup I_?(\mathbf{x}),\ |T| \le s\right\}, & \omega(\mathbf{x})_{[s]} = 2\lambda, \\ \left\{\mathbf{U}_{\mathcal{A}(T)}\mathbf{y} : \mathbf{y} \in P_{\mathcal{B}(T)}(\mathbf{x}_{\mathcal{A}(T)}),\ I_+(\mathbf{x}) = T\right\}, & \omega(\mathbf{x})_{[s]} < 2\lambda, \end{cases}$$

where

$$\begin{aligned} I_+(\mathbf{x}) &= \left\{j : \omega(\mathbf{x})_j > \max\{\omega(\mathbf{x})_{[s]}, 2\lambda\}\right\}, \\ I_?(\mathbf{x}) &= \left\{j : \omega(\mathbf{x})_j = \max\{\omega(\mathbf{x})_{[s]}, 2\lambda\},\ \mathbf{x}_{G_j} \ne \mathbf{0}\right\}. \end{aligned}$$

Theorem 3.2 suggests the following procedure to obtain a specific vector in $\mathrm{prox}_h(\mathbf{x})$.

[1] If $I_1(\mathbf{y}) \cap \{j : \omega(\mathbf{x})_j < 2\lambda\} = \emptyset$, then the sum equals 0, and otherwise it is negative.

[2] We use a convention that $T = \emptyset$, then $\mathbf{U}_{\mathcal{A}(T)}\mathbf{y} = \mathbf{0}$.

Algorithm 1: group-sparse proximal mapping

Input: $\mathbf{x} \in \mathbb{R}^n$;

Output: $\mathbf{u} \in \text{prox}_h(\mathbf{x})$;

1. compute $T \in S_s(\omega(\mathbf{x}))$
2. set $R = \{j \in T : \omega(\mathbf{x})_j > 2\lambda\}$
3. **return** $\mathbf{u} = \mathbf{U}_{\mathcal{A}(R)} P_{\mathcal{B}(R)}(\mathbf{x}_{\mathcal{A}(R)})$.

4 Necessary Optimality Conditions

We will now exploit the result of the previous section in order to analyze optimality conditions of the following problem:

$$\begin{aligned} \min \quad & f(\mathbf{x}) + \lambda \|g(\mathbf{x})\|_0 \\ \text{s.t.} \quad & \mathbf{x} \in C_s \cap B, \end{aligned} \tag{P}$$

where $B \subseteq \mathbb{R}^n$ is given by (1.3), and as before, $s \in \{1, 2, \ldots, m\}$, $\lambda \geq 0$.

This study begins with a stationarity-based condition that is defined as a fixed point of a proximal gradient procedure. Then, under the additional assumption that the sets $D_i \subseteq \mathbb{R}^{n_i}$ are convex, we devise and study coordinate-based conditions that are defined with respect to a small change of the support. Some of the presented results require the standard assumption that $f \in C^{1,1}_{L_f}$, which will be stated upon use.

Our approach is similar to that taken in the context of sparse optimization in [4, 6, 5], and somewhat in [7]. In all these studies the underlying set enjoyed some symmetry properties and the coordinate-wise based conditions were proved to be more restrictive than the stationary based conditions under the assumption that $f \in C^{1,1}_{L_f}$. We will show that similar results can be established in our setting which does not involve any symmetry properties, and consequently requires a different line of 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 $L > 0$:

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

4.1 L-stationarity

The following optimality condition is defined as a fixed point condition of the process $\mathbf{x}^{k+1} \in \text{prox}_{\frac{1}{L}h}\left(T_L(\mathbf{x}^k)\right)$, where we recall that $h(\cdot) \equiv \lambda\|g(\cdot)\|_0 + \delta_{B \cap C_s}(\cdot)$. In Section 5 we will study and analyze this method in more depth.

Definition 4.1 (L-stationarity)**.** Let $L > 0$. A vector $\mathbf{x} \in \mathbb{R}^n$ is called an **L-stationary point** of (P) if

$$\mathbf{x} \in \text{prox}_{\frac{1}{L}h}\left(T_L(\mathbf{x})\right). \tag{4.2}$$

If $f \in C^{1,1}_{L_f}$, then the L-stationarity condition is a necessary optimality condition whenever $L \geq L_f$.

Theorem 4.2 (optimality $\Rightarrow$ L-stationarity)**.** *Let $\mathbf{x}^* \in B \cap C_s$ be an optimal solution of (P), and suppose that $f \in C^{1,1}_{L_f}$. Then for any $L \geq L_f$*

$$\mathbf{x}^* \in \operatorname{prox}_{\frac{1}{L}h}\left(T_L(\mathbf{x}^*)\right). \tag{4.3}$$

Proof. Let $L > L_f$, and let $\mathbf{z} \in \operatorname{prox}_{\frac{1}{L}h}\left(T_L(\mathbf{x}^*)\right)$. Then by the sufficient decrease lemma (Lemma 2.2) and by the optimality of $\mathbf{x}^*$,

$$f(\mathbf{x}^*)+\lambda\|g(\mathbf{x}^*)\|_0 \geq \frac{L-L_f}{2}\left\|\mathbf{z}-\mathbf{x}^*\right\|^2+f(\mathbf{z})+\lambda\|g(\mathbf{z})\|_0 \geq \frac{L-L_f}{2}\left\|\mathbf{z}-\mathbf{x}^*\right\|^2+f(\mathbf{x}^*)+\lambda\|g(\mathbf{x}^*)\|_0.$$

Since $L > L_f$, we conclude that $\mathbf{z} = \mathbf{x}^*$, implying the validity of (4.3).

Now, for any $L > L_f$, $\mathbf{x}^*$ satisfies (4.3) and thus for any $\mathbf{u} \in B \cap C_s$

$$\frac{\lambda}{L}\|g(\mathbf{x}^*)\|_0 + \frac{1}{2}\|\mathbf{x}^* - T_L(\mathbf{x}^*)\|_2^2 \leq \frac{\lambda}{L}\|g(\mathbf{u})\|_0 + \frac{1}{2}\|\mathbf{u} - T_L(\mathbf{x}^*)\|_2^2.$$

By the continuity of the expressions in the above inequality as a function of L, taking $L \to L_f$ results with

$$\frac{\lambda}{L_f}\|g(\mathbf{x}^*)\|_0 + \frac{1}{2}\|\mathbf{x}^* - T_{L_f}(\mathbf{x}^*)\|_2^2 \leq \frac{\lambda}{L_f}\|g(\mathbf{u})\|_0 + \frac{1}{2}\|\mathbf{u} - T_{L_f}(\mathbf{x}^*)\|_2^2,$$

which implies that $\mathbf{x}^*$ satisfies (4.3) for $L = L_f$ as well. □

In the next part we will define coordinate-based conditions, under the additional assumption of convexity of the sets D_i $(i = 1, 2, \ldots, m)$.

4.2 Coordinate based conditions

Throughout this subsection we will assume that $D_i \subseteq \mathbb{R}^{n_i}$ is, in addition to the underlying assumptions, convex for any $i = 1, 2, \ldots, m$. Note that since the orthogonal projection onto a nonempty closed and convex set is unique, this assumption together with Theorem 3.2 imply the following characterization of the L-stationarity condition.

Corollary 4.3 (L-stationarity characterization)**.** *Let $\mathbf{x} \in \mathbb{R}^n$. Then $\mathbf{x}$ is an L-stationary point of (P) if and only if*

(a) $\mathbf{x}_{G_i} = P_{D_i}(T_L(\mathbf{x})_{G_i})$ *for any* $i \in I_1(\mathbf{x})$.

(b) There exists $Q \in S_s(\omega(T_L(\mathbf{x})))$ *for which*

$$Q \cap \{j : \omega(T_L(\mathbf{x}))_j > 2\lambda/L\} \subseteq I_1(\mathbf{x}) \subseteq Q \cap \{j : \omega(T_L(\mathbf{x}))_j \geq 2\lambda/L\}.$$

For a given set of indices $S \subseteq \{1, 2 \ldots, m\}$, we define $\mathcal{O}$ to be an oracle that produces the set of optimal solutions of f restricted to the index set S by

$$\mathcal{O}(S) \equiv \operatorname*{argmin}_{\mathbf{u}} \left\{f(\mathbf{u}) : I_1(\mathbf{u}) \subseteq S,\ \mathbf{u}_{G_i} \in D_i\ \forall i \in S\right\}. \tag{4.4}$$

In practice, we will only require one solution from $\mathcal{O}(S)$.

4.2.1 Support optimality

We begin by presenting an optimality condition called *group support optimality (GSO)* that, as its name suggests, states that the vector is an optimal solution of the restriction of f to its own support.

Definition 4.4 (support optimality)**.** A vector $\mathbf{x} \in B \cap C_s$ is called a **group support optimal (GSO) point** of (P) if $\mathbf{x} \in \mathcal{O}(I_1(\mathbf{x}))$.

It is easy to show that group support optimality is a necessary optimality condition.

Theorem 4.5 (optimality $\Rightarrow$ GSO)**.** *Let $\mathbf{x}^* \in B \cap C_s$ be an optimal solution of (P). Then $\mathbf{x}^*$ is a GSO point of (P).*

Proof. For any $\mathbf{z} \in \{\mathbf{u} : I_1(\mathbf{u}) \subseteq I_1(\mathbf{x}^*),\ \mathbf{u}_{G_i} \in D_i\ \ \forall i \in I_1(\mathbf{x}^*)\}$ it holds that $\|g(\mathbf{z})\|_0 \leq \|g(\mathbf{x}^*)\|_0$, and subsequently by the optimality of $\mathbf{x}^*$ (recalling that $\mathbf{z}$ is feasible),

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

Thus, $f(\mathbf{x}^*) \leq f(\mathbf{z})$, and consequently $\mathbf{x}^* \in \mathcal{O}(I_1(\mathbf{x}^*))$. □

The next lemma shows that group support optimality implies a condition that can be seen as a "support stationarity" condition.

Lemma 4.6. *Let $\mathbf{x} \in B \cap C_s$ be a GSO point of (P). Then for any $L > 0$, $\mathbf{x}$ satisfies*

$$\mathbf{x}_{G_i} = P_{D_i}\left(T_L(\mathbf{x})_{G_i}\right) \ \textit{for any } i \in I_1(\mathbf{x}). \tag{4.5}$$

Proof. Let $\tilde{C} = \prod_{i=1}^m \tilde{D}_i$ where for any $i = 1, 2, \ldots, m$,

$$\tilde{D}_i = \begin{cases} D_i, & i \in I_1(\mathbf{x}), \\ \{\mathbf{0}\}, & \text{otherwise.} \end{cases}$$

Then by the definition of $\tilde{C}$, $\mathbf{x} \in \mathcal{O}(I_1(\mathbf{x}))$ holds if and only if $\mathbf{x}$ is an optimal solution of

$$\min_{\mathbf{u}} \{f(\mathbf{u}) : \mathbf{u} \in \tilde{C}\}. \tag{4.6}$$

Since $\tilde{C}$ is a nonempty closed convex set, it follows that $\mathbf{x}$ must be a stationary point of (4.6), meaning that

$$\mathbf{x} = P_{\tilde{C}}\left(\mathbf{x} - \frac{1}{L}\nabla f(\mathbf{x})\right),$$

which in turn implies the validity of (4.5). □

In the case where $B = \mathbb{R}^n$, the condition of Lemma (4.6) translates to the property that the components of $\nabla f(\mathbf{x})$ at the active groups are zeros.

Corollary 4.7. *Suppose that $B = \mathbb{R}^n$. Let $\mathbf{x} \in \mathbb{R}^n$ be a GSO point of problem (P). Then*[3]

$$\nabla_{G_i} f(\mathbf{x}) = 0 \text{ for any } i \in I(\mathbf{x}).$$

Proof. Since $D_i = \mathbb{R}^{n_i}$, it follows by Lemma 4.6 that for any $i \in I_1(\mathbf{x})$

$$\mathbf{x}_{G_i} = P_{\mathbb{R}^{n_i}}\left(\mathbf{x}_{G_i} - \frac{1}{L}\nabla_{G_i} f(\mathbf{x})\right) = \mathbf{x}_{G_i} - \frac{1}{L}\nabla_{G_i} f(\mathbf{x}),$$

and hence, $\nabla_{G_i} f(\mathbf{x}) = \mathbf{0}$. □

When f is convex, the L-stationarity condition implies the GSO condition.

Lemma 4.8 (L-stationarity ⇒ GSO (f convex))**.** *Let $L > 0$. Suppose that f is convex and that $\mathbf{x} \in \mathbb{R}^n$ is an L-stationary point of (P). Then $\mathbf{x}$ is a GSO point of (P).*

Proof. Denote $S = \mathcal{A}(I_1(\mathbf{x}))$ and $\tilde{C} = \mathcal{B}(I_1(\mathbf{x}))$. Since

$$\mathbf{x} \in \operatorname{prox}_{\frac{1}{L}h}\left(T_L(\mathbf{x})\right),$$

it follows that

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

and hence,

$$\begin{aligned}
\|\mathbf{x} - T_L(\mathbf{x})\|_2^2 &\leq \min_{\mathbf{u}}\left\{\|\mathbf{u} - T_L(\mathbf{x})\|_2^2 : \mathbf{u} \in B, I_1(\mathbf{u}) \subseteq I_1(\mathbf{x})\right\} \\
&\leq \min_{\mathbf{d}}\left\{\|\mathbf{U}_S\mathbf{d} - T_L(\mathbf{x})\|_2^2 : \mathbf{d} \in \tilde{C}\right\}.
\end{aligned}$$

Decomposing the expressions in both sides of the above inequality with respect to the two sets of indices S and S^c, we obtain

$$\|\mathbf{x}_S - T_L(\mathbf{x})_S\|_2^2 + \|T_L(\mathbf{x})_{S^c}\|_2^2 \leq \min_{\mathbf{d}}\left\{\|\mathbf{d} - T_L(\mathbf{x})_S\|_2^2 : \mathbf{d} \in \tilde{C}\right\} + \|T_L(\mathbf{x})_{S^c}\|_2^2,$$

that is,

$$\|\mathbf{x}_S - T_L(\mathbf{x})_S\|_2^2 \leq \min_{\mathbf{d}}\left\{\|\mathbf{d} - T_L(\mathbf{x})_S\|_2^2 : \mathbf{d} \in \tilde{C}\right\},$$

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

$$\min\{f(\mathbf{U}_S\mathbf{d}) : \mathbf{d} \in \tilde{C}\}. \tag{4.7}$$

[3] $\nabla_{G_i} f(\mathbf{x})$ stands for the components of $\nabla f(\mathbf{x})$ corresponding to the indices in G_i, that is, $\nabla_{G_i} f(\mathbf{x}) = [\nabla f(\mathbf{x})]_{G_i}$.

Since problem (4.7) is convex (by the convexity of f and $\tilde{C}$), it follows that $\mathbf{x}_S$ is an optimal solution of (4.7), establishing the fact that it is a GSO point. □

4.2.2 Coordinate-wise optimality

In the rest of this section we will consider two coordinate-wise based conditions–the *partial coordinate-wise optimality (PCWO)* condition and the *coordinate-wise optimality (CWO)* condition. Loosely speaking, these conditions state that the function value does not improve if a small change in the support is performed. For a given GSO point $\mathbf{x}$, the conditions that we will consider will compare the function value of $\mathbf{x}$ with those of other GSO points defined by:

$$\mathbf{x}^{i,-} \in \mathcal{O}(J^i) \ : J^i = I_1(\mathbf{x}) \setminus \{i\}, \tag{4.8}$$

$$\mathbf{x}^{j,+} \in \mathcal{O}(J_j) \ : J_j = I_1(\mathbf{x}) \cup \{j\}, \tag{4.9}$$

$$\mathbf{x}^{i,j} \in \mathcal{O}(J^i_j) \ : J^i_j = (I_1(\mathbf{x}) \cup \{j\}) \setminus \{i\}, \tag{4.10}$$

for indices $i \in I_1(\mathbf{x})$ and $j \notin I_1(\mathbf{x})$. The PCWO property is defined for a specific parameter $L > 0$ and indices $i_{\mathbf{x},L}$, $j_{\mathbf{x},L}$, chosen according to the rule:

$$i_{\mathbf{x},L} \in \underset{\ell \in I_1(\mathbf{x})}{\operatorname{argmin}} \{\omega(T_L(\mathbf{x}))_\ell\}, \tag{4.11}$$

$$j_{\mathbf{x},L} \in \underset{\ell \notin I_1(\mathbf{x})}{\operatorname{argmax}} \{\omega(T_L(\mathbf{x}))_\ell\}. \tag{4.12}$$

Note that the choice of $i_{\mathbf{x},L}$ and $j_{\mathbf{x},L}$ in (4.11) and (4.12) respectively is affected by the parameter $L > 0$ as it changes the order of the elements in $\omega(T_L(\mathbf{x}))$. A special case in which the order is not affected by L is when $B = \mathbb{R}^n$.

Remark 4.9. Suppose that $B = \mathbb{R}^n$ and let $\mathbf{x}$ be a GSO point. Then since $D_i = \mathbb{R}^{n_i}$,

$$\omega(T_L(\mathbf{x}))_\ell = \|T_L(\mathbf{x})_{G_\ell}\|_2^2 - d^2_{\mathbb{R}^{n_\ell}}(T_L(\mathbf{x})_{G_\ell}) = \|T_L(\mathbf{x})_{G_\ell}\|_2^2 - 0 = \|T_L(\mathbf{x})_{G_\ell}\|_2^2.$$

If $\ell \in I(\mathbf{x})$, then by Corollary 4.7 it follows that $\nabla_{G_\ell} f(\mathbf{x}) = 0$ and hence

$$\omega(T_L(\mathbf{x}))_\ell = \|T_L(\mathbf{x})_{G_\ell}\|_2^2 = \left\| \mathbf{x}_{G_\ell} - \frac{1}{L} \nabla_{G_\ell} f(\mathbf{x}) \right\|_2^2 = \|\mathbf{x}_{G_\ell}\|_2^2.$$

If $\ell \notin I(\mathbf{x})$, then $\mathbf{x}_{G_\ell} = \mathbf{0}$, and thus,

$$\omega(T_L(\mathbf{x}))_\ell = \|T_L(\mathbf{x})_{G_\ell}\|_2^2 = \left\| \mathbf{x}_{G_\ell} - \frac{1}{L} \nabla_{G_\ell} f(\mathbf{x}) \right\|_2^2 = \frac{1}{L^2} \|\nabla_{G_\ell} f(\mathbf{x})\|_2^2.$$

Therefore, in the setting of $B = \mathbb{R}^n$, conditions (4.11) and (4.12) translate into the following relations which are independent of L:

$$i_{\mathbf{x},L} \in \underset{\ell \in I_1(\mathbf{x})}{\operatorname{argmin}} \|\mathbf{x}_{G_\ell}\|_2,$$

$$j_{\mathbf{x},L} \in \underset{\ell \notin I_1(\mathbf{x})}{\operatorname{argmax}} \|\nabla_{G_\ell} f(\mathbf{x})\|_2.$$

We make two important assumptions regarding the possible ambiguity in the choices of $\mathbf{x}^{i,-}, \mathbf{x}^{j,+}, \mathbf{x}^{i,j}, i_{\mathbf{x},L}, j_{\mathbf{x},L}$.

Remark 4.10. To simplify the exposition of the coordinate-wise based conditions analysis, we will assume that whenever $\mathbf{x}^{i,-}$,$\mathbf{x}^{i,j}$ or $\mathbf{x}^{j,+}$ (or other vectors similarly defined) appear, the corresponding required conditions on $\|g(\mathbf{x})\|_0$ given below are satisfied:

- when $\mathbf{x}^{i,-}$ appears it holds that $\|g(\mathbf{x})\|_0 > 0$,
- when $\mathbf{x}^{j,+}$ appears it holds that $\|g(\mathbf{x})\|_0 < s$,
- when $\mathbf{x}^{i,j}$ appears it holds that $0 < \|g(\mathbf{x})\|_0 \leq s < m$.

Remark 4.11. Note that the choice of the GSO points in (4.8),(4.9),(4.10), or the choice of $i_{\mathbf{x},L}$ and $j_{\mathbf{x},L}$ in (4.11) and (4.12), is not necessarily unique. We assume that there exists some well-defined deterministic policy by which the selection is made.

The partial coordinate-wise optimality (PCWO) property will now be defined.

Definition 4.12 (L-partial coordinate wise optimality)**.** Let $L > 0$ and $\mathbf{x} \in B \cap C_s$ be a GSO point of (P). Then $\mathbf{x}$ is called an **L-partial coordinate wise optimal (L-PCWO) point** of (P) if for $i = i_{\mathbf{x},L}$ and $j = j_{\mathbf{x},L}$

$$f(\mathbf{x}) + \lambda\|g(\mathbf{x})\|_0 \leq \min\left\{f(\mathbf{y}) + \lambda\|g(\mathbf{y})\|_0 : \mathbf{y} \in \{\mathbf{x}^{i,-}, \mathbf{x}^{j,+}, \mathbf{x}^{i,j}\}\right\}. \tag{4.13}$$

The CWO property is similar to the L-PCWO property with a substantial modification – it imposes the condition that the function value does not decrease by *any* change of at most two indices in the support.

Definition 4.13 (coordinate wise optimality)**.** Let $\mathbf{x} \in B \cap C_s$ be a GSO point of (P). Then $\mathbf{x}$ is a **coordinate wise optimal (CWO) point** of (P) if for any $i \in I_1(\mathbf{x})$ and $j \notin I_1(\mathbf{x})$ relation (4.13) is satisfied.

Obviously, the CWO property implies the L-PCWO property for any $L > 0$.

Theorem 4.14 (CWO $\Rightarrow$ L-PCWO)**.** *Let $\mathbf{x}^*$ be a CWO point of (P). Then $\mathbf{x}^*$ is an L-PCWO point of (P) for any $L > 0$.*

Another straightforward observation is that both CWO and L-PCWO conditions are necessary optimality conditions.

Theorem 4.15 (optimality $\Rightarrow$ CWO)**.** *Let $\mathbf{x}^*$ be an optimal solution of (P), then $\mathbf{x}^*$ is a CWO as well as an L-PCWO point of (P) for any $L > 0$.*

The next theorem states that when $f \in C^{1,1}_{L_f}$, the L-PCWO condition with parameter $L \geq L^G_f$ implies L-stationarity. Recall that $f \in C^{1,1}_{L_f}$ is a required assumption for the necessity of the L-stationarity condition (together with $L \geq L_f$, Theorem 4.2), and that L^G_f might be smaller than L_f (and in any case can be chosen as $L^G_f = L_f$).

Theorem 4.16 (L-PCWO $\Rightarrow$ L-stationarity)**.** *Suppose that $f \in C^{1,1}_{L_f}$. Let $L \geq L^G_f$, and $\mathbf{x} \in \mathbb{R}^n$ be an L-PCWO point of (P). Then $\mathbf{x}$ is an L-stationary point of (P).*

Proof. Denote $i \equiv i_{\mathbf{x},L}$ and $j \equiv j_{\mathbf{x},L}$. We will show that the condition for L-stationarity given in Corollary 4.3 holds. Utilizing the group descent lemma (Lemma 2.3) for any $\mathbf{z} \in \{\mathbf{z}_{i,-}, \mathbf{z}_{j,+}, \mathbf{z}_{i,j}\}$, we have

$$f(\mathbf{z}) - f(\mathbf{x}) \leq \langle \nabla f(\mathbf{x}), \mathbf{z} - \mathbf{x} \rangle + \frac{L}{2}\|\mathbf{z} - \mathbf{x}\|_2^2, \tag{4.14}$$

where

$$\begin{aligned} \mathbf{z}_{i,-} &= \mathbf{x} - \mathbf{U}_{G_i}\mathbf{x}_{G_i}, & (4.15)\\ \mathbf{z}_{j,+} &= \mathbf{x} + \mathbf{U}_{G_j}P_{D_j}\left(T_L(\mathbf{x})_{G_j}\right), & (4.16)\\ \mathbf{z}_{i,j} &= \mathbf{x} - \mathbf{U}_{G_i}\mathbf{x}_{G_i} + \mathbf{U}_{G_j}P_{D_j}\left(T_L(\mathbf{x})_{G_j}\right). & (4.17) \end{aligned}$$

Note that since $\mathbf{x}$ is a GSO point, by Lemma 4.6 it satisfies

$$\mathbf{x}_{G_\ell} = P_{D_\ell}(T_L(\mathbf{x})_{G_\ell}) \text{ for any } \ell \in I_1(\mathbf{x}). \tag{4.18}$$

By the definitions of $\mathbf{x}^{i,-}, \mathbf{x}^{j,+}$ and $\mathbf{x}^{i,j}$, we have that for $\mathbf{y} \in \{\mathbf{x}^{i,-}, \mathbf{x}^{j,+}, \mathbf{x}^{i,j}\}$,

$$\|g(\mathbf{x})\|_0 - \|g(\mathbf{y})\|_0 \geq \begin{cases} 1, & \mathbf{y} = \mathbf{x}^{i,-},\\ -1, & \mathbf{y} = \mathbf{x}^{j,+},\\ 0, & \mathbf{y} = \mathbf{x}^{i,j}, \end{cases}$$

and thus, by the L-PCWO property of $\mathbf{x}$, it holds that

$$f(\mathbf{y}) - f(\mathbf{x}) \geq \lambda(\|g(\mathbf{x})\|_0 - \|g(\mathbf{y})\|_0) \geq \begin{cases} \lambda, & \mathbf{y} = \mathbf{x}^{i,-},\\ -\lambda, & \mathbf{y} = \mathbf{x}^{j,+},\\ 0, & \mathbf{y} = \mathbf{x}^{i,j}. \end{cases} \tag{4.19}$$

Since $\mathbf{x}^{i,-} \in \mathcal{O}(J^i)$ and $I_1(\mathbf{z}_{i,-}) \subseteq J^i$, $\mathbf{x}^{j,+} \in \mathcal{O}(J_j)$ and $I_1(\mathbf{z}_{j,+}) \subseteq J_j$, $\mathbf{x}^{i,j} \in \mathcal{O}(J^i_j)$ and $I_1(\mathbf{z}_{i,j}) \subseteq J^i_j$, we have (by invoking (4.19)) that

$$\begin{aligned} f(\mathbf{z}_{i,-}) - f(\mathbf{x}) \geq \quad & f(\mathbf{x}^{i,-}) - f(\mathbf{x}) \geq \quad \lambda, & (4.20)\\ f(\mathbf{z}_{j,+}) - f(\mathbf{x}) \geq \quad & f(\mathbf{x}^{j,+}) - f(\mathbf{x}) \geq \quad -\lambda, & (4.21)\\ f(\mathbf{z}_{i,j}) - f(\mathbf{x}) \geq \quad & f(\mathbf{x}^{i,j}) - f(\mathbf{x}) \geq \quad 0. & (4.22) \end{aligned}$$

For the ith component in $\omega(T_L(\mathbf{x}))$ we have that

$$\begin{aligned}
\omega(T_L(\mathbf{x}))_i &= \|T_L(\mathbf{x})_{G_i}\|_2^2 - \|T_L(\mathbf{x})_{G_i} - P_{D_i}(T_L(\mathbf{x})_{G_i})\|_2^2 && \text{[def. of } \omega\text{]} \\
&= -\frac{2}{L}\langle \nabla_{G_i} f(\mathbf{x}), \mathbf{x}_{G_i}\rangle + \|\mathbf{x}_{G_i}\|_2^2 && \text{[algebra, (4.18), def. of } T_L\text{]} \quad (4.23) \\
&= \frac{2}{L}\langle \nabla f(\mathbf{x}), \mathbf{z}_{i,-} - \mathbf{x}\rangle + \|\mathbf{z}_{i,-} - \mathbf{x}\|_2^2 && \text{[(4.15)]} \\
&\geq \frac{2}{L}(f(\mathbf{z}_{i,-}) - f(\mathbf{x})) \geq \frac{2\lambda}{L}. && \text{[(4.14) and (4.20)]}
\end{aligned}$$

Thus, if $0 < \|g(\mathbf{x})\|_0$, then since $i = i_{\mathbf{x},L}$,

$$\omega(T_L(\mathbf{x}))_{l_1} \geq \omega(T_L(\mathbf{x}))_i \geq \frac{2\lambda}{L} \text{ for any } l_1 \in I_1(\mathbf{x}). \quad (4.24)$$

In particular, if $\|g(\mathbf{x})\|_0 = m$ (in this case $s = m$), then $I_1(\mathbf{x}) = \{1, 2, \ldots, m\}$, and by (4.24) for any $l_1 = 1, 2, \ldots, m$ it holds that $\omega(T_L(\mathbf{x}))_i \geq \frac{2\lambda}{L}$. Hence,

$$I_1(\mathbf{x}) \cap \{j : \omega(\mathbf{x})_j \geq 2\lambda/L\} = I_1(\mathbf{x}). \quad (4.25)$$

Since $S_s(\omega(T_L(\mathbf{x}))) = S_m(\omega(T_L(\mathbf{x}))) = \{I_1(\mathbf{x})\}$, we have that $I_1(\mathbf{x}) \in S_s(\omega(T_L(\mathbf{x})))$. Thus, the latter together with (4.25) and (4.18) imply by Corollary 4.3 that $\mathbf{x}$ is an L-stationary point.

For the jth component in $\omega(T_L(\mathbf{x}))$ we have that

$$\begin{aligned}
\omega(T_L(\mathbf{x}))_j &= \left\|T_L(\mathbf{x})_{G_j}\right\|_2^2 - \left\|T_L(\mathbf{x})_{G_j} - P_{D_j}(T_L(\mathbf{x})_{G_j})\right\|_2^2 && \text{[def. of } \omega\text{]} \\
&= -\frac{2}{L}\langle \nabla_{G_j} f(\mathbf{x}), (\mathbf{z}_{j,+})_{G_j}\rangle - \|(\mathbf{z}_{j,+})_{G_j}\|_2^2 && \text{[algebra, (4.16), def. of } T_L\text{]} \quad (4.26) \\
&= -\frac{2}{L}\langle \nabla f(\mathbf{x}), \mathbf{z}_{j,+} - \mathbf{x}\rangle - \|\mathbf{z}_{j,+} - \mathbf{x}\|_2^2 && \text{[(4.16)]} \\
&\leq -\frac{2}{L}(f(\mathbf{z}_{j,+}) - f(\mathbf{x})) \leq \frac{2\lambda}{L}, && \text{[(4.14) and (4.21)]}
\end{aligned}$$

which implies that if[4] $\|g(\mathbf{x})\|_0 < s$, then since $j = j_{\mathbf{x},L}$,

$$\omega(T_L(\mathbf{x}))_{l_2} \leq \omega(T_L(\mathbf{x}))_j \leq \frac{2\lambda}{L} \text{ for any } l_2 \notin I_1(\mathbf{x}). \quad (4.27)$$

In particular, if $\|g(\mathbf{x})\|_0 = 0$ then $I_1(\mathbf{x}) = \emptyset$, and thus by (4.27) $\omega(T_L(\mathbf{x}))_{l_2} \leq \frac{2\lambda}{L}$ for any $l_2 = 1, 2, \ldots, m$. Consequently,

$$\{j : \omega(\mathbf{x})_j > 2\lambda/L\} = \emptyset = I_1(\mathbf{x}),$$

and thus for any $Q \in S_s(\omega(T_L(\mathbf{x})))$ it holds that $Q \cap \{j : \omega(\mathbf{x})_j > 2\lambda/L\} = I_1(\mathbf{x})$, which implies by Corollary 4.3 that $\mathbf{x}$ is an L-stationary point.

[4]This result assumes that $\mathbf{x}^{j,+}$ exists, which happens only if $\|g(\mathbf{x})\|_0 < s$, see Remark 4.10.

If $0 < \|g(\mathbf{x})\|_0 < s$, then by combining (4.24) and (4.27) we have that

$$\omega(T_L(\mathbf{x}))_{l_1} \geq \omega(T_L(\mathbf{x}))_i \geq \frac{2\lambda}{L} \geq \omega(T_L(\mathbf{x}))_j \geq \omega(T_L(\mathbf{x}))_{l_2} \text{ for all } l_1 \in I_1(\mathbf{x}), l_2 \notin I_1(\mathbf{x}). \quad (4.28)$$

Hence, there exists a $Q \in S_s(\omega(T_L(\mathbf{x})))$ for which $I_1(\mathbf{x}) \subseteq Q$, and $\{j : \omega(T_L(\mathbf{x}))_j > 2\lambda/L\} \subseteq I_1(\mathbf{x}) \subseteq \{j : \omega(T_L(\mathbf{x}))_j \geq 2\lambda/L\}$. Therefore,

$$Q \cap \{j : \omega(T_L(\mathbf{x}))_j > 2\lambda/L\} \subseteq I_1(\mathbf{x}) \subseteq Q \cap \{j : \omega(T_L(\mathbf{x}))_j \geq 2\lambda/L\},$$

which together with (4.18) implies by Corollary 4.3 that $\mathbf{x}$ is an L-stationary point.

For the ith and jth components of $\omega(T_L(\mathbf{x}))$ we have that

$$\begin{aligned}
&\omega(T_L(\mathbf{x}))_i - \omega(T_L(\mathbf{x}))_j \\
&\quad = \|\mathbf{x}_{G_i}\|_2^2 - \frac{2}{L}\langle \nabla_{G_i} f(\mathbf{x}), \mathbf{x}_{G_i} \rangle + \frac{2}{L}\langle \nabla_{G_j} f(\mathbf{x}), (\mathbf{z}_{j,+})_{G_j} \rangle + \|(\mathbf{z}_{j,+})_{G_j}\|_2^2 && [(4.23), (4.26)] \\
&\quad = \frac{2}{L}\langle \nabla f(\mathbf{x}), \mathbf{z}_{i,j} - \mathbf{x} \rangle + \|\mathbf{z}_{i,j} - \mathbf{x}\|_2^2 && [\text{algebra}, (4.17)] \\
&\quad \geq \frac{2}{L}(f(\mathbf{z}_{i,j}) - f(\mathbf{x})) \geq 0. && [(4.14), (4.22)]
\end{aligned}$$

Thus, if $0 < \|g(\mathbf{x})\|_0 = s < m$, then

$$\omega(T_L(\mathbf{x}))_{l_1} \geq \omega(T_L(\mathbf{x}))_i \geq \omega(T_L(\mathbf{x}))_j \geq \omega(T_L(\mathbf{x}))_{l_2} \text{ for all } l_1 \in I_1(\mathbf{x}), l_2 \notin I_1(\mathbf{x}). \quad (4.29)$$

Noting that $|I_1(\mathbf{x})| = s$, (4.29) implies that $I_1(\mathbf{x}) \in S_s(\omega(T_L(\mathbf{x})))$. In addition, by (4.24) we have that

$$I_1(\mathbf{x}) \cap \{l : \omega(T_L(\mathbf{x}))_l \geq 2\lambda/L\} = I_1(\mathbf{x}),$$

which together with (4.18) implies by Corollary 4.3 that $\mathbf{x}$ is an L-stationary point.

Hence, for any value of $\|g(\mathbf{x})\|_0$ the L-PCWO point $\mathbf{x}$ is an L-stationary point, as required. □

The hierarchy of the optimality conditions, under the assumption that D_i's are convex, is illustrated by Figure 1.

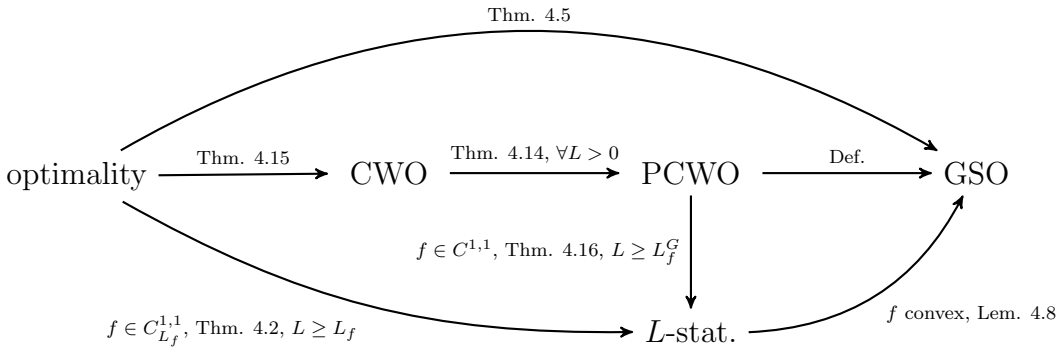

Figure 1: Optimality conditions' hierarchy.

In the next section we will derive methods to obtain points satisfying 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 with stepsize $\frac{1}{L}$; the prox operator can be computed using Algorithm 1.

Algorithm 2: proximal gradient method

Input: $\mathbf{x}^0 \in \mathbb{R}^n$, $L > 0$.

repeat

1. $\mathbf{x}^{k+1} \in \operatorname{prox}_{\frac{1}{L}h}\left(T_L(\mathbf{x}^k)\right)$;
2. $k \leftarrow k + 1$;

Since f is lower bounded, the sufficient decrease lemma (Lemma 2.2) can be utilized in order 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. *Suppose that $f \in C^{1,1}_{L_f}$, and let $\{\mathbf{x}^k\}_{k\geq 0}$ be the sequence generated by the proximal gradient method with $L > L_f$. Then*

(a) $f(\mathbf{x}^k) + \lambda\|g(\mathbf{x}^k)\|_0 - f(\mathbf{x}^{k+1}) - \lambda\|g(\mathbf{x}^{k+1})\|_0 \geq \frac{L-L_f}{2}\left\|\mathbf{x}^{k+1} - \mathbf{x}^k\right\|^2$;

(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 2.2). To prove part (b), note that by part (a) the sequence of function values $\{f(\mathbf{x}^k)+\lambda\|g(\mathbf{x}^k)\|_0\}_{k\geq 0}$ is nonincreasing and in addition, by the standing 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 $\mathbf{x}^{k_i+1} \in \operatorname{prox}_{\frac{1}{L}h}(T_L(\mathbf{x}^{k_i}))$, by the definition of the prox operator we have

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

Taking the limit $i \to \infty$, and exploiting the continuity of T_L, and the lower semicontinuity of $\|\cdot\|_0$, yields

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

and consequently $\mathbf{x}^* \in \operatorname{prox}_{\frac{1}{L}h}(T_L(\mathbf{x}^*))$, meaning that $\mathbf{x}^*$ is an L-stationary point. □

5.2 Group Coordinate Descent Methods

We present two coordinate descent methods that obtain points that satisfy the coordinate-wise optimality conditions defined in Section 4.2. The convexity of $D_i \subseteq \mathbb{R}^{n_i}$ for any $i = 1, 2, \ldots, m$ is a prerequisite and thus will be assumed throughout this section.

5.2.1 Partial group coordinate descent

The *partial group coordinate descent (PGCD)* algorithm is designed to obtain an L-PCWO point, and under the assumption that $f \in C^{1,1}_{L_f}$, which is a required assumption for the necessity of the L_f-stationarity condition (Theorem 4.2), an L^G_f-PCWO point that is also L^G_f-stationary. Consequently, the PGCD method returns points satisfying a more restrictive optimality condition than that of the outputs of the proximal gradient method.

The PGCD algorithm moves between GSO points. At each iteration it examines the current GSO point and three other GSO points according to the PCWO condition, and if a better point is found, then it is chosen as the next point. Otherwise, the current point is an L-PCWO point and this point is returned.

As in the previous section, to simplify the discussion we will make the assumptions described in Remarks 4.10 and 4.11 regarding the ambiguity in choosing a GSO point given an index set and the choices of the indices that will enter or exist the support set.

Algorithm 3: partial group coordinate descent (PGCD)

Input: $L > 0$, $\mathbf{x}^0 \in B \cap C_s$.

1. initialize: $\mathbf{x} \in \mathcal{O}(I_1(\mathbf{x}^0))$;
2. compute: $i \in \underset{\ell \in I_1(\mathbf{x})}{\operatorname{argmin}} \{\omega(T_L(\mathbf{x}))_\ell\}$ and $j \in \underset{\ell \notin I_1(\mathbf{x})}{\operatorname{argmax}} \{\omega(T_L(\mathbf{x}))_\ell\}$;
3. compute:

$$\begin{aligned} \mathbf{x}^{i,-} &\in \mathcal{O}(J^i) \quad J^i = I_1(\mathbf{x}) \setminus \{i\}, \\ \mathbf{x}^{j,+} &\in \mathcal{O}(J_j), \quad J_j = I_1(\mathbf{x}) \cup \{j\}, \\ \mathbf{x}^{i,j} &\in \mathcal{O}(J^i_j), \quad J^i_j = (I_1(\mathbf{x}) \cup \{j\}) \setminus \{i\}; \end{aligned}$$

4. if $f(\mathbf{x}) + \lambda \|g(\mathbf{x})\|_0 > \min\{f(\mathbf{y}) + \lambda \|g(\mathbf{y})\|_0 : \mathbf{y} \in \{\mathbf{x}^{i,-}, \mathbf{x}^{j,+}, \mathbf{x}^{i,j}\}\}$, then set $\mathbf{x} \in \operatorname{argmin}\{f(\mathbf{y}) + \lambda \|g(\mathbf{y})\|_0 : \mathbf{y} \in \{\mathbf{x}^{i,-}, \mathbf{x}^{j,+}, \mathbf{x}^{i,j}\}\}$, $k \leftarrow k + 1$ and goto 2.
5. **Return** $\mathbf{x}$.

The PGCD method is finite; since the update condition in step 4 dictates a strict decrease in the function value when moving from the current GSO point to the next and no group support is repeated, meaning that every group support is examined at most once. Therefore, the number of iterations is bounded by the number of possible group supports – at most 2^m (in the case $s = m$).

The properties of the output of the PGCD method are given next.

Lemma 5.2. *Let* $\mathbf{x}$ *be the output of the PGCD method with input* $(L, \mathbf{x}^0)$ *(*$L > 0, \mathbf{x}^0 \in B \cap C_s$*). Then*

1. $\mathbf{x}$ *is an L-PCWO point of* (P)*.*
2. *If* $f \in C^{1,1}$ *and* $L = L_f^G$*. Then* $\mathbf{x}$ *is an* L_f^G*-PCWO point of* (P) *and an* L_f^G*-stationary point of* (P)*.*

Proof. The first part is a direct result of the finiteness of the PGCD method together with the stopping criteria in step 4 (and choice of indices in step 2).

For the second part, suppose that $f \in C^{1,1}_{L_f}$ and that $L = L_f^G$. By the first part, $\mathbf{x}$ is an L_f^G-PCWO point of (P), and thus, by Theorem 4.16, it is an L_f^G-stationary point. □

5.2.2 Full group coordinate descent

The *full group coordinate-wise descent (FGCD)* algorithm given below generates a sequence of GSO points and terminates when a CWO point is obtained after a finite amount of steps.

Algorithm 4: full group coordinate descent (FGCD)

Input: $\mathbf{x}^0 \in B \cap C_s$.

1. initialize: $\mathbf{x} \in \mathcal{O}(I_1(\mathbf{x}^0))$
2. for $[i, j] \in I_1(\mathbf{x}) \times I_1(\mathbf{x})^c$ do:
   (a) compute:

$$\begin{array}{rll} \mathbf{x}^{i,-} & \in \mathcal{O}(J^i) & J^i = I_1(\mathbf{x}) \setminus \{i\}; \\ \mathbf{x}^{j,+} & \in \mathcal{O}(J_j), & J_j = I_1(\mathbf{x}) \cup \{j\}; \\ \mathbf{x}^{i,j} & \in \mathcal{O}(J^i_j), & J^i_j = (I_1(\mathbf{x}) \cup \{j\}) \setminus \{i\}; \end{array}$$

   (b) if $f(\mathbf{x}) + \lambda\|g(\mathbf{x})\|_0 > \min\{f(\mathbf{y}) + \lambda\|g(\mathbf{y})\|_0 : \mathbf{y} \in \{\mathbf{x}^{i,-}, \mathbf{x}^{j,+}, \mathbf{x}^{i,j}\}\}$, then
      (b.1) set: $\mathbf{x} \in \operatorname{argmin}\left\{f(\mathbf{y}) + \lambda\|g(\mathbf{y})\|_0 : \mathbf{y} \in \{\mathbf{x}^{i,-}, \mathbf{x}^{j,+}, \mathbf{x}^{i,j}\}\right\}$;
      (b.2) goto 2;

   end for.
3. **return** $\mathbf{x}$.

Remark 5.3. We assume that the order by which the indices in step 3 are chosen is given.

The FGCD method is finite as the update in step 2(b) dictates a strict decrease in the function value when moving from the current GSO point to the next, thus no group support is repeated. Since there are at most 2^m possible group supports, the number of iterations is finite.

The FGCD method obviously returns a CWO point.

Theorem 5.4. *Let* $\mathbf{x}$ *be the output of the FGCD. Then* $\mathbf{x}$ *is a CWO point of problem* (P)*.*

6 Numerical Illustrations

6.1 Investment Problems

In many investment problems such as portfolio optimization or index tracking (see [26]), the decision variables are stocks that are already partitioned into disjoint groups according to their activity sector, such as transportation or retail trade. In some cases, one of the objectives is to bound the number of different sectors the investor wishes to invests in.

To illustrate the results obtained for the different optimality conditions derived in this paper, we experimented on a portfolio optimization problem. We assume that we are given the following parameters: $\boldsymbol{\mu} \in \mathbb{R}^n$ - mean return vector, $\mathbf{C} \in \mathbb{R}^{n\times n}$ - positive semidefinite covariance matrix, and the parameter $\gamma > 0$ that penalizes the variance with respect to the mean return. The set $\Delta_k = \{\mathbf{x} \in \mathbb{R}^k : \mathbf{x} \geq \mathbf{0}, \sum_{i=1}^k x_i = 1\}$ is the unit-simplex. The portfolio optimization problem we consider minimizes a weighted sum of the variance and minus the expected return subject to budget constraints and a limit on the number of invested sector. The mathematical formulation is as follows:

$$\begin{aligned} \min \quad & -\boldsymbol{\mu}^T\mathbf{x} + \gamma \cdot \mathbf{x}^T\mathbf{C}\mathbf{x} \\ \text{s.t.} \quad & \mathbf{x} \in C_s \cap B, \end{aligned}$$

where $B = \prod_{i=1}^m (\Delta_{n_i} \cup \{\mathbf{0}\})$. The objective function belongs to the class of $C^{1,1}$ functions with parameters $L_f = 2\gamma\lambda_{\max}(\mathbf{C})$ and L_f^G computed by (2.6).

The experiment's data was the stock returns of the members of the SP500 in the consecutive trading days between March 1st 2016 to December 30 2016. In purpose of demonstrating the difference between the optimality conditions, two tests were conducted: (1) counting the number of points satisfying each optimality condition; (2) randomizing 100 starting points (uniformly over the simplex for each group by the method described in [22, Algorithm 2.5.3]), computing the percentage that the proximal gradient method reached the optimal solution (unique in this problem) and then computing the chances that the PGCD algorithm obtained the optimal solution after it was employed on the output of the proximal gradient method. This means that if the proximal gradient method obtained the optimal solution, then the PGCD algorithm obtained it as well (as it started from it). Table 1 summarizes the results of both experiments for several values of γ and s.

In all instances, the number of GSO points was equal to the number of possible different supports of size s; in almost all instances, any L_f-stationary point was also an L_f^G-stationary point, and therefore we only present the number of L_f^G-stationary points. In the PGCD algorithm we used $L = L_f^G$.

Table description:

- GSO = number of different GSO points;
- L_f^G-stat. = number of $L_G(f)$-stationary points;
- L_f^G-PCWO = number of $L_G(f)$-PCWO points;
- CWO = number of CWO points;

- OPT = number of optimal points;
- PG = % the proximal gradient method reached the optimal solution from a random point;
- PG+PGCD= % the PGCD algorithm reached the optimal solution from the output of the proximal gradient method.

Main observations:

1. The number of PCWO points is significantly smaller than the number of L_f^G-stationary points for almost any value of γ and s, which suggests that obtaining an optimal solution using the PGCD method is more likely compared to the proximal gradient algorithm.
2. There is only one CWO point (which is also the optimal solution) for any value of γ and s. This means that the FGCD is guaranteed to converge to the optimal solution in all problem instances that were explored in this experiment.
3. The PGCD method was able to improve the chances for obtaining the optimal solution in many situations.

6.2 Binary Decision Variables

An interesting instance of problem (P) is the optimization over binary decision variables already described in Example 1.4. Many combinatorial optimization problems can be formulated as binary constrained problems, see for example [21] and reference therein. Here we consider the maximum weight clique problem as described in [12, Theorem 2.2], with the additional constraint of bound on the number of chosen vertices. Let $G = (V, E)$ be an undirected graph composed of the vertices set $V = \{1, 2, \ldots, n\}$ and the edges set $E \subseteq V \times V$. Each vertex $i \in V$ is associated with a positive weight, collected in the weights vector $\mathbf{w} \in \mathbb{R}^n_{++}$. The maximum weight clique problem is given by (see [12, Theorem 2.4] in which the equivalence to the independent set problem is also discussed)

$$\begin{aligned} \min \quad & \mathbf{x}^T \mathbf{Q} \mathbf{x} \\ \text{s.t.} \quad & \textstyle\sum_{i=1}^n x_i \leq s, \ \mathbf{x} \in \{0,1\}^n, \end{aligned}$$

where

$$\mathbf{Q}_{i,j} = \begin{cases} -w_i, & i = j, \\ \frac{1}{2}(w_i + w_j), & (i,j) \notin E, \\ 0, & (i,j) \in E. \end{cases}$$

We generated 100 graphs for 6 possible graph sizes ($n = 10, 12, 14, 16, 18, 20$) and computed the number of points satisfying each optimality condition. For each graph instance, the edges set E was randomly generated (each edge exists in probability $\frac{1}{2}$). The weights vector for the set V was generated uniformly over the set $\{1, 2, \ldots, 5\}^n$, and the bound on the clique size was chosen as $s = 5$. Table 2 summarizes the results by depicting the mean number of points satisfying each optimality condition per size $n = |V|$.

γ	**s**	**GSO**	L_f^G**-stat.**	L_f^G**-PCWO**	**CWO**	**OPT**	**PG**	**PG+PGCD**
0.02	8	165	5	3	1	1	97%	100%
0.04	8	165	20	2	1	1	83%	100%
0.06	8	165	57	22	1	1	75%	100%
0.08	8	165	84	2	1	1	79%	100%
0.24	8	165	129	42	1	1	52%	52%
0.48	8	165	165	55	1	1	20%	20%
0.96	8	165	165	48	1	1	5%	11%
1.92	8	165	165	41	1	1	1%	4%
0.02	5	462	3	3	1	1	100%	100%
0.04	5	462	10	10	1	1	100%	100%
0.06	5	462	48	1	1	1	100%	100%
0.08	5	462	336	133	1	1	100%	100%
0.24	5	462	336	133	1	1	48%	53%
0.48	5	462	462	131	1	1	7%	34%
0.96	5	462	462	63	1	1	2%	8%
1.92	5	462	462	49	1	1	1%	7%
0.02	3	165	2	1	1	1	100%	100%
0.04	3	165	5	1	1	1	100%	100%
0.06	3	165	7	7	1	1	100%	100%
0.08	3	165	9	9	1	1	95%	100%
0.24	3	165	129	35	1	1	1%	1%
0.48	3	165	165	39	1	1	3%	3%
0.96	3	165	165	21	1	1	2%	6%
1.92	3	165	165	19	1	1	2%	25%

Table 1: Number of points satisfying each optimality condition, and the percentage of reaching the optimal solution from a random point by the prox-grad, and the percentage of reaching the optimal from the output of the prox-grad by the PGCD.

n	**GSO**	L_f**-stat.**	L_f^G**-stat.**	L_f^G**-PCWO**	**CWO**	**OPT**
10	637	522.3	450.59	30.19	3.66	1.38
12	1585	1393.04	1172.67	51.65	4.48	1.42
14	3472	3241.72	2690.19	83.79	5.82	1.5
16	6884	6703.56	5678.19	131.53	7.05	1.51
18	12615	12539.61	10732.97	178.2	8.98	1.63
20	21699	21678.67	18988.51	254.97	11.16	1.68

Table 2: Mean number of points satisfying each optimality condition per n.

Evidently, there is a very large gap between the number of points satisfying each optimality condition in all problems, a gap which significantly increases as the number of vertices increases.

References

[1] L. Baldassarre, N. Bhan, V. Cevher, A. Kyrillidis, and S. Satpathi. Group-sparse model selection: Hardness and relaxations. *IEEE Transactions on Information Theory*, 62(11):6508–6534, 2016.

[2] R.G. Baraniuk, V. Cevher, M.F. Duarte, and C. Hegde. Model-based compressive sensing. *IEEE Transactions on Information Theory*, 56(4):1982–2001, 2010.

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

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

[5] A. Beck and N. Hallak. Proximal mapping for symmetric penalty and sparsity.

[6] 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.

[7] 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.

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

[9] T. Blumensath and M. E. Davies. Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis*, 27(3):265–274, 2009.

[10] T. Blumensath, M. E. Davies, and E. Mike. Sampling theorems for signals from the union of finite-dimensional linear subspaces. *IEEE Transactions on Information Theory*, 55(4):1872–1882, 2009.

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

[12] I.M. Bomze, M. Budinich, P.M. Pardalos, and M. Pelillo. The maximum clique problem. In *Handbook of combinatorial optimization*, pages 1–74. Springer, 1999.

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

[14] M.F Duarte, V. Cevher, and R.G. Baraniuk. Model-based compressive sensing for signal ensembles. In *Communication, Control, and Computing, 2009. Allerton 2009. 47th Annual Allerton Conference on*, pages 244–250. IEEE, 2009.

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

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

[17] Y.C. Eldar, P. Kuppinger, and H. Bolcskei. Block-sparse signals: Uncertainty relations and efficient recovery. *IEEE Transactions on Signal Processing*, 58(6):3042–3054, 2010.

[18] Y.C. Eldar and M. Mishali. Robust recovery of signals from a structured union of subspaces. *IEEE Transactions on Information Theory*, 55(11):5302–5316, 2009.

[19] P. Jain, N. Rao, and I.S. Dhillon. Structured sparse regression via greedy hard thresholding. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 1516–1524. Curran Associates, Inc., 2016.

[20] R. Jenatton, J. Audibert, and F. Bach. Structured variable selection with sparsity-inducing norms. *Journal of Machine Learning Research*, 12(Oct):2777–2824, 2011.

[21] G. Kochenberger, J. Hao, F. Glover, M. Lewis, Z. Lü, H. Wang, and Y. Wang. The unconstrained binary quadratic programming problem: a survey. *Journal of Combinatorial Optimization*, 28(1):58–81, 2014.

[22] D.P. Kroese and R.Y. Rubinstein. *Simulation and the Monte Carlo method*. Wiley New York, 2008.

[23] L. Meier, S. Van De Geer, and P. Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71, 2008.

[24] Jean-Jacques Moreau. Proximité et dualité dans un espace hilbertien. *Bulletin de la Société mathématique de France*, 93:273–299, 1965.

[25] Y. Nardi and A. Rinaldo. On the asymptotic properties of the group lasso estimator for linear models. *Electronic Journal of Statistics*, 2:605–633, 2008.

[26] J.L. Prigent. *Portfolio optimization and performance analysis*. CRC Press, 2007.

[27] S. Sra, S. Nowozin, and S.J. Wright. *Optimization for machine learning*. Mit Press, 2012.

[28] M. Stojnic, F. Parvaresh, and B. Hassibi. On the reconstruction of block-sparse signals with an optimal number of measurements. *IEEE Transactions on Signal Processing*, 57(8):3075–3085, 2009.

[29] 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.

[30] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.