Deutsche Forschungsgemeinschaft
Priority Programme 1962

# *An Augmented Lagrangian Method for Optimization Problems with Structured Geometric Constraints*

Xiaoxi Jia, Christian Kanzow, Patrick Mehlitz, Gerd Wachsmuth

SPP 1962
Non-smooth and Complementarity-based Distributed Parameter Systems: Simulation and Hierarchical Optimization

Preprint Number SPP1962-170

received on May 19, 2021

Edited by
SPP1962 at Weierstrass Institute for Applied Analysis and Stochastics (WIAS)
Leibniz Institute in the Forschungsverbund Berlin e.V.
Mohrenstraße 39, 10117 Berlin, Germany
E-Mail: spp1962@wias-berlin.de

World Wide Web: http://spp1962.wias-berlin.de/

# An Augmented Lagrangian Method for Optimization Problems with Structured Geometric Constraints *

Xiaoxi Jia† Christian Kanzow† Patrick Mehlitz‡ Gerd Wachsmuth‡

May 19, 2021

**Abstract.** This paper is devoted to the theoretical and numerical investigation of an augmented Lagrangian method for the solution of optimization problems with geometric constraints. Specifically, we study situations where parts of the constraints are nonconvex and possibly complicated, but allow for a fast computation of projections onto this nonconvex set. Typical problem classes which satisfy this requirement are optimization problems with disjunctive constraints (like complementarity or cardinality constraints) as well as optimization problems over sets of matrices which have to satisfy additional rank constraints. The key idea behind our method is to keep these complicated constraints explicitly in the constraints and to penalize only the remaining constraints by an augmented Lagrangian function. The resulting subproblems are then solved with the aid of a problem-tailored nonmonotone projected gradient method. The corresponding convergence theory allows for an inexact solution of these subproblems. Nevertheless, the overall algorithm computes so-called Mordukhovich-stationary points of the original problem under a mild asymptotic regularity condition, which is generally weaker than most of the respective available problem-tailored constraint qualifications. Extensive numerical experiments addressing complementarity- and cardinality-constrained optimization problems as well as a semidefinite reformulation of Maxcut problems visualize the power of our approach.

**Keywords.** Asymptotic Regularity, Augmented Lagrangian Method, Complementarity Constraints, Cardinality Constraints, Maxcut Problem, Mordukhovich-Stationarity, Nonmonotone Projected Gradient Method

**AMS subject classifications.** 49J53, 65K10, 90C22, 90C30, 90C33

*This research was supported by the German Research Foundation (DFG) within the priority program "Non-smooth and Complementarity-based Distributed Parameter Systems: Simulation and Hierarchical Optimization" (SPP 1962) under grant numbers KA 1296/24-2 and WA 3636/4-2.

†University of Würzburg, Institute of Mathematics, 97074 Würzburg, Germany, email: {xiaoxi.jia,kanzow}@mathematik.uni-wuerzburg.de,

‡Brandenburgische Technische Universität Cottbus-Senftenberg, Institute of Mathematics, 03046 Cottbus, Germany, email: {mehlitz,wachsmuth}@b-tu.de.

# 1 Introduction

We consider the program

$$\min_{w} \; f(w) \quad \text{s.t.} \quad G(w) \in C, \; w \in D, \tag{P}$$

where $\mathbb{W}$ and $\mathbb{Y}$ are Euclidean spaces, i.e., real and finite-dimensional Hilbert spaces, $f\colon \mathbb{W} \to \mathbb{R}$ and $G\colon \mathbb{W} \to \mathbb{Y}$ are continuously differentiable, $C \subset \mathbb{Y}$ is nonempty, closed, and convex, whereas the set $D \subset \mathbb{W}$ is only assumed to be nonempty and closed. This setting is very general and covers, amongst others, standard nonlinear programs, second-order cone and, more generally, conic optimization problems [11, 21], as well as several so-called disjunctive programming problems like mathematical programs with complementarity, vanishing, switching, or cardinality constraints, see [12, 13, 25, 49] for an overview and suitable references. Since $\mathbb{W}$ and $\mathbb{Y}$ are Euclidean spaces, our model also covers matrix optimization problems like semidefinite programs or low-rank approximation problems [46].

The aim of this paper is to apply a (structured) augmented Lagrangian technique to (P) in order to find suitable stationary points. The augmented Lagrangian or multiplier penalty method is a classical approach for the solution of nonlinear programs, see [14] as a standard reference. The more recent book [15] presents a slightly modified version of this classical augmented Lagrangian method, which uses a safeguarded update of the Lagrange multipliers and has stronger global convergence properties. In the meantime, this safeguarded augmented Lagrangian method has also been applied to a number of optimization problems with disjunctive constraints, see e.g. [4, 29, 38, 41, 54].

Since, to the best of our knowledge, augmented Lagrangian methods have not yet been applied to the general problem (P) with nonconvex $D$, and in order to get a better understanding of our contributions, let us add some comments regarding the existing results for the probably most prominent non-standard optimization problem, namely the class of mathematical programs with complementarity constraints (MPCCs). Due to the particular structure of the feasible set, the usual Karush–Kuhn–Tucker (KKT for short) conditions are typically not satisfied at a local minimum. Hence, other (weaker) stationary concepts have been proposed, like C- (abbreviating Clarke) and M- (for Mordukhovich) stationarity, with M-stationarity being the stronger concept. Most algorithms (regularization, penalty, augmented Lagrangian methods etc.) for the solution of MPCCs solve a sequence of standard nonlinear programs, and their limit points are typically C-stationary points only. Some approaches can identify M-stationary points if the underlying nonlinear programs are solved exactly, but they loose this desirable property if these programs are solved only inexactly, see the discussion in [42] for more details.

The authors are currently aware of only three approaches where convergence to M-stationary points for a general (nonlinear) MPCC is shown using inexact solutions of the corresponding subproblems, namely [7, 29, 54]. All three papers deal with suitable modifications of the (safeguarded) augmented Lagrangian method. The basic idea of reference [7] is to solve the subproblems such that both a first- and a second-order necessary optimality condition hold inexactly at each iteration, i.e., satisfaction

of the second-order condition is the central point here which, obviously, causes some overhead for the subproblem solver and usually excludes the application of this approach to large-scale problems. The paper [54] proves convergence to M-stationary points by solving some complicated subproblems, but for the latter no method is specified. Finally, the recent approach described in [29] provides an augmented Lagrangian technique for the solution of MPCCs where the complementarity constraints are kept as constraints, whereas the standard constraints are penalized. The authors present a technique which computes a suitable stationary point of these subproblems in such a way that the entire method generates M-stationary accumulation points for the original MPCC. Let us also mention that [32] suggests to solve (a discontinuous reformulation of) the M-stationarity system associated with an MPCC by means of a semismooth Newton-type method. Naturally, this approach should be robust with respect to (w.r.t.) an inexact solution of the appearing Newton-type equations although this issue is not discussed in [32].

The current paper universalizes the idea from [29] to the much more general problem (P). In fact, a closer look at the corresponding proofs shows that the technique from [29] can be generalized using some relatively small modifications. We are therefore able to skip some of the proofs, complete details can be found in an appendix of the preprint version [39] of this paper. This allows us to concentrate on some additional new contributions. In particular, we prove convergence to an M-type stationary point of the general problem (P) under a very weak sequential constraint qualification introduced recently in [47] for the general setting from (P). We further show that this sequential constraint qualification holds under the conditions for which convergence to M-stationary points of an MPCC is shown in [29]. Note that this is also the first algorithmic application of the general sequential stationarity and regularity concepts from [47].

The global convergence result for our method holds for the abstract problem (P) with geometric constraints without any further assumptions regarding the sets $C$ and, in particular, $D$. Conceptually, we are therefore able to deal with a very large class of optimization problems. On the other hand, we use a projected gradient-type method for the solution of the resulting subproblems. Since this requires projections onto the (usually nonconvex) set $D$, our method can be implemented efficiently only if $D$ is simple in the sense that projections onto $D$ are easy to compute. For this kind of "structured" geometric constraints (this explains the title of this paper), the entire method is then both an efficient tool and applicable to large-scale problems. In particular, we show that this is the case for MPCCs, optimization problems with cardinality constraints, and some rank-constrained matrix optimization problems.

The paper is organized as follows. We begin with restating some basic definitions from variational analysis in Section 2. There, we also relate the general regularity concept from [47] to the constraint qualification (the so-called relaxed constant positive linear dependence condition, RCPLD for short) used in the underlying paper [29] (as well as in many other related publications in this area). We then present the spectral gradient method for optimization problems over nonconvex sets in Section 3. This method is used to solve the resulting subproblems of our augmented Lagrangian method whose details are given in Section 4. Global convergence to M-type stationary

points is also shown in this section. Since, in our augmented Lagrangian approach, we penalize the seemingly easy constraints $G(w) \in C$, but keep the condition $w \in D$ explicitly in the constraints, we have to compute projections onto $D$. Section 5 therefore considers a couple of situations where this can be done in a numerically very efficient way. Extensive computational experiments for some of these situations are documented in Section 6. This includes MPCCs, cardinality-constrained (sparse) optimization problems, and a rank-constrained reformulation of the famous Maxcut problem. We close with some final remarks in Section 7.

Notation. The Euclidean inner product of two vectors $x, y \in \mathbb{R}^n$ will be denoted by $x^\top y$. More generally, $\langle x, y \rangle$ is used to represent the inner product of $x, y \in \mathbb{W}$ whenever $\mathbb{W}$ is some abstract Euclidean space. For brevity, we exploit $x + A := A + x := \{x + a \,|\, a \in A\}$ for arbitrary vectors $x \in \mathbb{W}$ and sets $A \subset \mathbb{W}$. The sets $\operatorname{cone} A$ and $\operatorname{span} A$ denote the smallest cone containing the set $A$ and the smallest subspace containing $A$, respectively. Whenever $L\colon \mathbb{W} \to \mathbb{Y}$ is a linear operator between Euclidean spaces $\mathbb{W}$ and $\mathbb{Y}$, $L^*\colon \mathbb{Y} \to \mathbb{W}$ denotes its adjoint. For some continuously differentiable mapping $\varphi\colon \mathbb{W} \to \mathbb{Y}$ and some point $w \in \mathbb{W}$, we use $\varphi'(w) \in \mathcal{L}(\mathbb{W}, \mathbb{Y})$ in order to denote the derivative of $\varphi$ at $w$. In the particular case $\mathbb{Y} := \mathbb{R}$, we set $\nabla\varphi(w) := \varphi'(w)^* 1 \in \mathbb{W}$ for brevity.

# 2 Preliminaries

We first recall some basic concepts from variational analysis in Section 2.1, and then introduce and discuss general stationarity and regularity concepts for the abstract problem (P) in Section 2.2.

## 2.1 Fundamentals of Variational Analysis

In this section, we comment on the tools of variational analysis which will be exploited in order to describe the geometry of the closed, convex set $C \subset \mathbb{Y}$ and the closed (but not necessarily convex) set $D \subset \mathbb{W}$ which appear in the formulation of (P).

The Euclidean projection $P_C\colon \mathbb{Y} \to \mathbb{Y}$ onto the closed, convex set $C$ is given by

$$P_C(y) := \operatorname*{argmin}_{z \in C} \|z - y\|.$$

Thus, the corresponding distance function $d_C\colon \mathbb{Y} \to \mathbb{R}$ can be written as

$$d_C(y) := \min_{z \in C} \|z - y\| = \|P_C(y) - y\|.$$

On the other hand, projections onto the potentially nonconvex set $D$ still exist, but are, in general, not unique. Therefore, we define the corresponding (usually set-valued) projection operator $\Pi_D\colon \mathbb{W} \rightrightarrows \mathbb{W}$ by

$$\Pi_D(x) := \operatorname*{argmin}_{z \in D} \|z - x\| \neq \varnothing.$$

Given $\bar{w} \in D$, the closed cone

$$\mathcal{N}_D^{\lim}(\bar{w}) := \limsup_{w \to \bar{w}} \big[\operatorname{cone}(w - \Pi_D(w))\big]$$

is referred to as the limiting normal cone to $D$ at $\bar{w}$, see [52, 57] for other representations and properties of this variational tool. Above, we used the notion of the outer (or upper) limit of a set-valued mapping at a certain point, see e.g. [57, Definition 4.1]. For $w \notin D$, we set $\mathcal{N}_D^{\lim}(w) := \varnothing$. Note that the limiting normal cone depends on the inner product of $\mathbb{W}$ and is stable in the sense that

$$\limsup_{w \to \bar{w}} \mathcal{N}_D^{\lim}(w) = \mathcal{N}_D^{\lim}(\bar{w}) \qquad \forall \bar{w} \in \mathbb{W} \tag{2.1}$$

holds. This stability property, which might be referred to as outer semicontinuity of the set-valued operator $\mathcal{N}_D^{\lim} \colon \mathbb{W} \rightrightarrows \mathbb{W}$, will play an essential role in our subsequent analysis. The limiting normal cone to the convex set $C$ coincides with the standard normal cone from convex analysis, i.e., for $\bar{y} \in C$, we have

$$\mathcal{N}_C^{\lim}(\bar{y}) = \mathcal{N}_C(\bar{y}) := \{\lambda \in \mathbb{Y} \,|\, \langle \lambda, y - \bar{y} \rangle \leq 0 \quad \forall y \in C\}\,.$$

For points $y \notin C$, we set $\mathcal{N}_C(y) := \varnothing$ for formal completeness. Note that the stability property (2.1) is also satisfied by the set-valued operator $\mathcal{N}_C \colon \mathbb{Y} \rightrightarrows \mathbb{Y}$.

## 2.2 Stationarity and Regularity Concepts

Noting that the abstract set $D$ is generally nonconvex in the exemplary settings we have in mind, the so-called concept of Mordukhovich-stationarity, which exploits limiting normals to $D$, is a reasonable concept of stationarity which addresses (P).

**Definition 2.1.** Let $\bar{w} \in \mathbb{W}$ be feasible for the optimization problem (P). Then $\bar{w}$ is called an *M-stationary point* (Mordukhovich-stationary point) of (P) if there exists a multiplier $\lambda \in \mathbb{Y}$ such that

$$0 \in \nabla f(\bar{w}) + G'(\bar{w})^* \lambda + \mathcal{N}_D^{\lim}(\bar{w}), \quad \lambda \in \mathcal{N}_C(G(\bar{w})).$$

Note that this definition coincides with the usual KKT conditions of (P) if the set $D$ is convex. An asymptotic counterpart of this definition is the following one, see [47].

**Definition 2.2.** Let $\bar{w} \in \mathbb{W}$ be feasible for the optimization problem (P). Then $\bar{w}$ is called an *AM-stationary point* (asymptotically M-stationary point) of (P) if there exist sequences $\{w^k\}, \{\varepsilon^k\} \subset \mathbb{W}$ and $\{\lambda^k\}, \{z^k\} \subset \mathbb{Y}$ such that $w^k \to \bar{w}$, $\varepsilon^k \to 0$, $z^k \to 0$, as well as

$$\varepsilon^k \in \nabla f(w^k) + G'(w^k)^* \lambda^k + \mathcal{N}_D^{\lim}(w^k), \quad \lambda^k \in \mathcal{N}_C(G(w^k) - z^k) \qquad \forall k \in \mathbb{N}.$$

The definition of an AM-stationary point is similar to the notion of an AKKT (asymptotic or approximate KKT) point, see [15], but requires some explanation: The meanings of the iterates $w^k$ and the Lagrange multiplier estimates $\lambda^k$ should be clear. The vector $\varepsilon^k$ measures the inexactness by which the stationary conditions are satisfied at $w^k$ and $\lambda^k$. The vector $z^k$ does not occur (at least not explicitly) in the context of standard nonlinear programs, but is required here for the following reason: The method to be considered in this paper generates a sequence $\{w^k\}$ satisfying $w^k \in D$, while the constraint $G(w) \in C$ gets penalized, hence, the condition $G(w^k) \in C$ will typically be violated. Consequently, the corresponding normal cone $\mathcal{N}_C(G(w^k))$ would be empty which is why we cannot expect to have $\lambda^k \in \mathcal{N}_C(G(w^k))$, though we hope that this holds asymptotically. In order to deal with this situation, we therefore have to introduce the sequence $\{z^k\}$.

Apart from this difference, the motivation of AM-stationarity is similar to the one of AKKT-stationarity: Suppose that the sequence $\{\lambda^k\}$ is bounded and, therefore, convergent along a subsequence. Then, taking the limit on this subsequence in the definition of an AM-stationary point while using the stability property (2.1) of the limiting normal cone shows that the corresponding limit point satisfies the M-stationarity conditions from Definition 2.1. In general, however, the Lagrange multiplier estimates $\{\lambda^k\}$ in the definition of AM-stationarity might be unbounded. Though this boundedness can be guaranteed under suitable (relatively strong) assumptions, the resulting convergence theory works under significantly weaker conditions.

It is well known in optimization theory that a local minimizer of (P) is M-stationary only under validity of a suitable constraint qualification. In contrast, it has been pointed out in [47, Theorem 4.2, Section 5.1] that each local minimizer of (P) is AM-stationary. In order to infer that an AM-stationary point is already M-stationary, the presence of so-called asymptotic regularity is necessary, see [47, Definition 4.4].

**Definition 2.3.** A feasible point $\bar{w} \in \mathbb{W}$ of (P) is called *AM-regular* (asymptotically Mordukhovich-regular) whenever the condition

$$\limsup_{w \to \bar{w},\, z \to 0} \mathcal{M}(w, z) \subset \mathcal{M}(\bar{w}, 0)$$

holds, where $\mathcal{M}\colon \mathbb{W} \times \mathbb{Y} \rightrightarrows \mathbb{W}$ is the set-valued mapping defined via

$$\mathcal{M}(w, z) := G'(w)^* \mathcal{N}_C(G(w) - z) + \mathcal{N}_D^{\lim}(w).$$

The concept of AM-regularity has been inspired by the notion of AKKT-regularity (sometimes referred to as cone continuity property), which became popular as one of the weakest constraint qualifications for standard nonlinear programs or MPCCs, see e.g. [5, 6, 54], and can be generalized to a much higher level of abstractness. In this regard, we would like to point the reader's attention to the fact that AM-stationarity and -regularity from Definitions 2.2 and 2.3 are referred to as *decoupled* asymptotic Mordukhovich-stationarity and -regularity in [47] since these are already refinements of more general concepts. For the sake of a concise notation, however, we omit the term *decoupled* here.

It has been shown in [47, Section 5.1] that validity of AM-regularity at a feasible point $\bar{w} \in \mathbb{W}$ of (P) is implied by

$$0 \in G'(\bar{w})^* \lambda + \mathcal{N}_D^{\lim}(\bar{w}), \quad \lambda \in \mathcal{N}_C(G(\bar{w})) \quad \Longrightarrow \quad \lambda = 0. \tag{2.2}$$

The latter is known as NNAMCQ (no nonzero abnormal multiplier constraint qualification) or GMFCQ (generalized Mangasarian–Fromovitz constraint qualification) in the literature. Indeed, in the setting where we fix $C := \mathbb{R}^{m_1}_- \times \{0\}^{m_2}$ and $D := \mathbb{W}$, (2.2) boils down to the classical Mangasarian–Fromovitz constraint qualification from standard nonlinear programming. The latter choice for $C$ will be of particular interest, which is why we formalize this setting below.

**Setting 2.4.** Given $m_1, m_2 \in \mathbb{N}$, we set $m := m_1 + m_2$, $\mathbb{Y} := \mathbb{R}^m$, and $C := \mathbb{R}^{m_1}_- \times \{0\}^{m_2}$. No additional assumptions are postulated on the set $D$. We denote the component functions of $G$ by $G_1, \ldots, G_m \colon \mathbb{W} \to \mathbb{R}$. Thus, the constraint $G(w) \in C$ encodes the constraint system

$$G_i(w) \leq 0 \quad i = 1, \ldots, m_1, \qquad G_i(w) = 0 \quad i = m_1 + 1, \ldots, m$$

of standard nonlinear programming. For our analysis, we exploit the index sets

$$I(\bar{w}) := \{i \in \{1, \ldots, m_1\} \,|\, G_i(\bar{w}) = 0\}, \qquad J := \{m_1 + 1, \ldots, m\},$$

whenever $\bar{w} \in D$ satisfies $G(\bar{w}) \in C$ in the present situation.

Let us emphasize that we did not make any assumptions regarding the structure of the set $D$ in Setting 2.4. Thus, it still covers numerous interesting problem classes like complementarity-, vanishing-, or switching-constrained programs. These so-called disjunctive programs of special type are addressed in the setting mentioned below which provides a refinement of Setting 2.4.

**Setting 2.5.** Let $\mathbb{X}$ be another Euclidean space, let $X \subset \mathbb{X}$ be a closed, convex set of simple structure, and let $T \subset \mathbb{R}^2$ be the union of two polyhedrons $T_1, T_2 \subset \mathbb{R}^2$. For functions $g \colon \mathbb{X} \to \mathbb{R}^{m_1}$, $h \colon \mathbb{X} \to \mathbb{R}^{m_2}$, and $p, q \colon \mathbb{X} \to \mathbb{R}^{m_3}$, we consider the constraint system given by

$$\begin{aligned} g_i(x) &\leq 0 && i = 1, \ldots, m_1, \\ h_i(x) &= 0 && i = 1, \ldots, m_2, \\ \big(p_i(x), q_i(x)\big) &\in T && i = 1, \ldots, m_3, \\ x &\in X. \end{aligned}$$

Setting $\mathbb{W} := \mathbb{X} \times \mathbb{R}^{m_3} \times \mathbb{R}^{m_3}$, $\mathbb{Y} := \mathbb{R}^{m_1} \times \mathbb{R}^{m_2} \times \mathbb{R}^{m_3} \times \mathbb{R}^{m_3}$,

$$G(x, u, v) := \big(g(x), h(x), p(x) - u, q(x) - v\big),$$

and

$$C := \mathbb{R}^{m_1}_- \times \{0\}^{m_2 + 2m_3}, \qquad D := X \times \{(u, v) \,|\, (u_i, v_i) \in T \quad \forall i \in \{1, \ldots, m_3\}\},$$

we can handle this situation in the framework of this paper.

Constraint regions as characterized in Setting 2.4 can be tackled with a recently introduced version of RCPLD (relaxed constant positive linear dependence constraint qualification), see [60, Definition 1.1].

**Definition 2.6.** Let $\bar{w} \in \mathbb{W}$ be a feasible point of the optimization problem (P) in Setting 2.4. Then $\bar{w}$ is said to satisfy RCPLD whenever the following conditions hold:

(i) the family $(\nabla G_i(w))_{i\in J}$ has constant rank on a neighborhood of $\bar{w}$,

(ii) there exists an index set $S \subset J$ such that the family $(\nabla G_i(\bar{w}))_{i\in S}$ is a basis of the subspace $\operatorname{span}\{\nabla G_i(\bar{w}) \,|\, i \in J\}$, and

(iii) for each index set $I \subset I(\bar{w})$, each set of multipliers $\lambda_i \geq 0$ $(i \in I)$ and $\lambda_i \in \mathbb{R}$ $(i \in S)$, not all vanishing at the same time, and each vector $\eta \in \mathcal{N}_D^{\lim}(\bar{w})$ which satisfy

$$0 \in \sum_{i\in I\cup S} \lambda_i \nabla G_i(\bar{w}) + \eta,$$

we find neighborhoods $U$ of $\bar{w}$ and $V$ of $\eta$ such that for all $w \in U$ and $\tilde{\eta} \in \mathcal{N}_D^{\lim}(w) \cap V$, the vectors from

$$(\nabla G_i(w))_{i\in I\cup S},\ \tilde{\eta}$$

are linearly dependent.

RCPLD has been introduced for standard nonlinear programs (i.e., $D := \mathbb{W} = \mathbb{R}^n$ in Setting 2.4) in [3]. Some extensions to complementarity-constrained programs can be found in [23, 30].

In case where $D$ is a set of product structure, condition (iii) in Definition 2.6 can be slightly weakened in order to obtain a reasonable generalization of the classical relaxed constant positive linear dependence constraint qualification, see [60, Remark 1.1] for details. Observing that GMFCQ from (2.2) takes the particular form

$$0 \in \sum_{i\in I(\bar{w})\cup J} \lambda_i \nabla G_i(\bar{w}) + \mathcal{N}_D^{\lim}(\bar{w}), \quad \lambda_i \geq 0\,(i \in I) \quad \Longrightarrow \quad \lambda_i = 0\,(i \in I(\bar{w}) \cup J)$$

in Setting 2.4, it is obviously sufficient for RCPLD. The subsequently stated result generalizes related observations from [5, 54].

**Lemma 2.7.** *Let $\bar{w} \in \mathbb{W}$ be a feasible point for the optimization problem (P) in Setting 2.4 where RCPLD holds. Then $\bar{w}$ is AM-regular.*

**Proof:** Fix some $\xi \in \limsup_{w\to\bar{w},\, z\to 0} \mathcal{M}(w, z)$. Then we find $\{w^k\}, \{\xi^k\} \subset \mathbb{W}$ and $\{z^k\} \subset \mathbb{R}^m$ which satisfy $w^k \to \bar{w}$, $\xi^k \to \xi$, $z^k \to 0$, and $\xi^k \in \mathcal{M}(w^k, z^k)$ for all $k \in \mathbb{N}$. Particularly, there are sequences $\{\lambda^k\}$ and $\{\eta^k\}$ satisfying $\lambda^k \in \mathcal{N}_C(G(w^k) - z^k)$, $\eta^k \in \mathcal{N}_D^{\lim}(w^k)$, and $\xi^k = G'(w^k)^*\lambda^k + \eta_k$ for each $k \in \mathbb{N}$. From $G(w^k) - z^k \to G(\bar{w})$ and the special structure of $C$, we find $G_i(w^k) - z_i^k < 0$ for all $i \in \{1, \ldots, m_1\} \setminus I(\bar{w})$ and all sufficiently large $k \in \mathbb{N}$, i.e.,

$$\lambda_i^k \begin{cases} = 0 & i \in \{1, \ldots, m_1\} \setminus I(\bar{w}), \\ \geq 0 & i \in I(\bar{w}) \end{cases}$$

for sufficiently large $k \in \mathbb{N}$. Thus, we may assume without loss of generality that

$$\xi^k = \sum_{i \in I(\bar{w}) \cup J} \lambda_i^k \nabla G_i(w^k) + \eta^k$$

holds for all $k \in \mathbb{N}$. By definition of RCPLD, $(\nabla G_i(w^k))_{i \in S}$ is a basis of the subspace $\operatorname{span}\{\nabla G_i(w^k) \,|\, i \in J\}$ for all sufficiently large $k \in \mathbb{N}$. Hence, there exist scalars $\mu_i^k$ $(i \in S)$ such that

$$\xi^k = \sum_{i \in I(\bar{w})} \lambda_i^k \nabla G_i(w^k) + \sum_{i \in S} \mu_i^k \nabla G_i(w^k) + \eta^k$$

holds for all sufficiently large $k \in \mathbb{N}$. On the other hand, [3, Lemma 1] yields the existence of an index set $I^k \subset I(\bar{w})$ and multipliers $\hat{\mu}_i^k > 0$ $(i \in I^k)$, $\hat{\mu}_i^k \in \mathbb{R}$ $(i \in S)$, and $\sigma_k \geq 0$ such that

$$\xi^k = \sum_{i \in I^k \cup S} \hat{\mu}_i^k \nabla G_i(w^k) + \sigma_k \, \eta^k$$

and

$$\begin{aligned} \sigma_k > 0 \quad &\Longrightarrow \quad (\nabla G_i(w^k))_{i \in I^k \cup S},\ \eta^k \text{ linearly independent,} \\ \sigma_k = 0 \quad &\Longrightarrow \quad (\nabla G_i(w^k))_{i \in I^k \cup S} \text{ linearly independent.} \end{aligned}$$

Since there are only finitely many subsets of $I(\bar{w})$, there needs to exist $I \subset I(\bar{w})$ such that $I^k = I$ holds along a whole subsequence. Along such a particular subsequence (without relabeling), we furthermore may assume $\sigma_k > 0$ (otherwise, the proof will be easier) and, thus, may set $\hat{\eta}^k := \sigma_k \eta^k \in \mathcal{N}_D^{\lim}(w^k)$. From above, we find linear independence of

$$(\nabla G_i(w^k))_{i \in I \cup S},\ \hat{\eta}^k.$$

Furthermore, we have

$$\xi^k = \sum_{i \in I \cup S} \hat{\mu}_i^k \nabla G_i(w^k) + \hat{\eta}^k. \tag{2.3}$$

Suppose that the sequence $\{((\hat{\mu}_i^k)_{i \in I \cup S}, \hat{\eta}^k)\}$ is not bounded. Dividing (2.3) by the norm of $((\hat{\mu}_i^k)_{i \in I \cup S}, \hat{\eta}^k)$, taking the limit $k \to \infty$, and respecting boundedness of $\{\xi^k\}$, continuity of $G'$, and outer semicontinuity of the limiting normal cone yield the existence of a non-vanishing multiplier $((\hat{\mu}_i)_{i \in I \cup S}, \hat{\eta})$ which satisfies $\hat{\mu}_i \geq 0$ $(i \in I)$, $\hat{\eta} \in \mathcal{N}_D^{\lim}(\bar{w})$, and

$$0 = \sum_{i \in I \cup S} \hat{\mu}_i \nabla G_i(\bar{w}) + \hat{\eta}.$$

Obviously, the multipliers $\hat{\mu}_i$ $(i \in I \cup S)$ do not vanish at the same time since, otherwise, $\hat{\eta} = 0$ would follow from above which yields a contradiction. Now, validity of RCPLD guarantees that the vectors

$$(\nabla G_i(w^k))_{i \in I \cup S},\ \hat{\eta}^k$$

need to be linearly dependent for sufficiently large $k \in \mathbb{N}$. However, we already have shown above that these vectors are linearly independent, a contradiction.

Thus, the sequence $\{((\hat{\mu}_i^k)_{i\in I\cup S}, \hat{\eta}^k)\}$ is bounded and, therefore, possesses a convergent subsequence with limit $((\bar{\mu}_i)_{i\in I\cup S}, \bar{\eta})$. Taking the limit in (2.3) while respecting $\xi^k \to \xi$, the continuity of $G'$, and the outer semicontinuity of the limiting normal cone, we come up with $\bar{\mu}_i \geq 0$ $(i \in I)$, $\bar{\eta} \in \mathcal{N}_D^{\lim}(\bar{w})$, and

$$\xi = \sum_{i\in I\cup S} \bar{\mu}_i \nabla G_i(\bar{w}) + \bar{\eta}.$$

Finally, we set $\bar{\mu}_i := 0$ for all $i \in \{1, \ldots, m\} \setminus (I \cup S)$. Then we have $(\bar{\mu}_i)_{i=1,\ldots,m} \in \mathcal{N}_C(G(\bar{w}))$ from $I \subset I(\bar{w})$, i.e.,

$$\xi \in G'(\bar{w})^* \mathcal{N}_C(G(\bar{w})) + \mathcal{N}_D^{\lim}(\bar{w}) = \mathcal{M}(\bar{w}, 0).$$

This shows that $\bar{w}$ is AM-regular. □

Another popular situation, where AM-regularity is inherently satisfied, is described in the following lemma which follows from [48, Theorem 3.10].

**Lemma 2.8.** *Let $\bar{w} \in \mathbb{W}$ be a feasible point for the optimization problem (P) where $G$ is affine, $C$ is a polyhedron, and $D$ is the union of finitely many polyhedrons. Then $\bar{w}$ is AM-regular.*

The above considerations underline that AM-regularity is a comparatively weak constraint qualification for (P).

# 3 A Spectral Gradient Method for Nonconvex Sets

In this section, we discuss a solution method for constrained optimization problems which applies whenever projections onto the feasible set are easy to find. Particularly, our method can be used in situations where the feasible set has a complicated nonconvex structure.

To motivate the method, first consider the unconstrained optimization problem

$$\min_w \ \varphi(w) \quad \text{s.t.} \quad w \in \mathbb{R}^n$$

with a continuously differentiable objective function $\varphi\colon \mathbb{R}^n \to \mathbb{R}$, and let $w^k$ be a current estimate for a solution of this problem. Computing the next iterate $w^{k+1}$ as the unique minimizer of the local quadratic model

$$\min_w \ \varphi(w^k) + \nabla\varphi(w^k)^\top (w - w^k) + \frac{\gamma_k}{2}\|w - w^k\|^2$$

for some $\gamma_k > 0$ leads to the explicit expression

$$w^{k+1} := w^k - \frac{1}{\gamma_k}\nabla\varphi(w^k),$$

i.e., we get a steepest descent method with stepsize $t_k := 1/\gamma_k$. Classical approaches compute $t_k$ using a suitable stepsize rule such that $\varphi(w^{k+1}) < \varphi(w^k)$. On the other hand, one can view the update formula as a special instance of a quasi-Newton scheme

$$w^{k+1} := w^k - B_k^{-1}\nabla\varphi(w^k)$$

with the very simple quasi-Newton matrix $B_k := \gamma_k I$ as an estimate of the (not necessarily existing) Hessian $\nabla^2\varphi(w^k)$. Then the corresponding quasi-Newton equation

$$B_{k+1}s^k = y^k \quad \text{with } s^k := w^{k+1} - w^k,\ y^k := \nabla\varphi(w^{k+1}) - \nabla\varphi(w^k),$$

see [24], reduces to the linear system $\gamma_{k+1}s^k = y^k$. Solving this overdetermined system in a least squares sense, we then obtain the stepsize

$$\gamma_{k+1} := (s^k)^\top y^k/(s^k)^\top s^k$$

introduced by Barzilai and Borwein [8]. This stepsize often leads to very good numerical results, but may not yield a monotone decrease in the function value. A convergence proof for general nonlinear programs is therefore difficult, even if the choice of $\gamma_k$ is safeguarded in the sense that it is projected onto some box $[\gamma_{\min}, \gamma_{\max}]$ for suitable constants $0 < \gamma_{\min} < \gamma_{\max}$.

Raydan [55] then suggested to control this nonmonotone behavior by combining the Barzilai–Borwein stepsize with the nonmonotone linesearch strategy introduced by Grippo et al. [28]. This, in particular, leads to a global convergence theory for general unconstrained optimization problems.

This idea was then generalized by Birgin et al. [16] to constrained optimization problems

$$\min_w\ \varphi(w) \quad \text{s.t.} \quad w \in W$$

with a nonempty, closed, and convex set $W \subset \mathbb{R}^n$ and is called the *nonmonotone spectral gradient method*. Here, we extend their approach to minimization problems

$$\min_w\ \varphi(w) \quad \text{s.t.} \quad w \in D \tag{3.1}$$

with a continuously differentiable function $\varphi\colon \mathbb{W} \to \mathbb{R}$ and some nonempty, closed set $D \subset \mathbb{W}$, where $\mathbb{W}$ is an arbitrary Euclidean space. Let us emphasize that neither $\varphi$ nor $D$ need to be convex in our subsequent considerations. A detailed description of the corresponding generalized spectral gradient method is as follows.

**Algorithm 3.1.** *(General Spectral Gradient Method)*

(S.0) *Choose* $\tau > 1, \sigma \in (0,1), 0 < \gamma_{\min} \le \gamma_{\max} < \infty, m \in \mathbb{N}, w^0 \in D$, *and set* $k := 0$.

(S.1) *If a suitable termination criterion holds at iteration* $k$*: STOP.*

(S.2) *Set* $m_k := \min(k, m)$ *and choose* $\gamma_k^0 \in [\gamma_{\min}, \gamma_{\max}]$. *For* $i = 0, 1, \ldots$, *compute a solution* $w^{k,i}$ *of*

$$\min_w\ \varphi(w^k) + \langle\nabla\varphi(w^k), w - w^k\rangle + \frac{\gamma_{k,i}}{2}\|w - w^k\|^2 \quad \text{s.t.} \quad w \in D \qquad (\mathrm{Q}(k,i))$$

*with* $\gamma_{k,i} = \tau^i \gamma_k^0$, *until the acceptance criterion*

$$\varphi(w^{k,i}) \leq \max_{j=0,1,\dots,m_k} \varphi(w^{k-j}) + \sigma \varphi'(w^k)(w^{k,i} - w^k) \tag{3.2}$$

*holds. Denote by* $i_k := i$ *the terminal value and set* $\gamma_k := \gamma_{k,i_k}$ *and* $w^{k+1} = w^{k,i_k}$.

(S.3) *Set* $k \leftarrow k+1$, *and go to* (S.1).

Particular instances of this approach with nonconvex sets $D$ can already be found in [10, 29]. Note that all iterates belong to the set $D$, that the subproblems (Q$(k,i)$) are always solvable, and that we have to compute only one solution, although their solutions are not necessarily unique. We would like to emphasize that $\nabla\varphi(w^k)$ was used in the formulation of (Q$(k,i)$) in order to underline that Algorithm 3.1 is a projected gradient method. Indeed, simple calculations reveal that the global solutions of (Q$(k,i)$) correspond to the projections of $w^k - \gamma_{k,i}^{-1}\nabla\varphi(w^k)$ onto $D$. Note also that the acceptance criterion in (S.2) is the nonmonotone Armijo rule introduced by Grippo et al. [28].

We stress that the previous generalization of existing spectral gradient methods plays a fundamental role in order to apply our subsequent augmented Lagrangian technique to several interesting and difficult optimization problems, but the convergence analysis of Algorithm 3.1 can be carried out similar to the one given in [29]. We therefore skip the corresponding proofs in this section, but for the readers' convenience, we present complete proofs in the appendix of the preprint version [39] of this paper.

Throughout, we assume implicitly that Algorithm 3.1 generates an infinite sequence, i.e., we neglect the termination criterion in (S.1). The next result then shows that the inner loop in (S.2) is always finite as long as $w^k$ is not already an M-stationary point of the given optimization problem (3.1). To this end, recall that $w^k$ is an M-stationary point of (3.1) if

$$0 \in \nabla\varphi(w^k) + \mathcal{N}_D^{\lim}(w^k)$$

holds. Similarly, $w^{k+1}$ solves the subproblem

$$\min_w \ \varphi(w^k) + \langle \nabla\varphi(w^k), w - w^k\rangle + \frac{\gamma_k}{2}\|w - w^k\|^2 \quad \text{s.t.} \quad w \in D \tag{3.3}$$

and satisfies the corresponding stationarity condition

$$0 \in \nabla\varphi(w^k) + \gamma_k\big(w^{k+1} - w^k\big) + \mathcal{N}_D^{\lim}(w^{k+1}). \tag{3.4}$$

Let us point the reader's attention to the fact that strong stationarity, where the limiting normal cone is replaced by the smaller regular normal cone in the stationarity system, provides a more restrictive necessary optimality condition for (3.1) and the surrogate (3.3), see [57, Definition 6.3, Theorem 6.12]. It is well known that the limiting normal cone is the outer limit of the regular normal cone. In contrast to the limiting normal cone, the regular one is not robust in the sense of (2.1), and since we are interested in taking limits later on, one either way ends up with a stationarity

systems in terms of limiting normals at the end. Thus, we will rely on the limiting normal cone and the associated concept of M-stationarity.

Coming back to (3.4), if $w^{k+1} = w^k$, then $w^k$ would already be an M-stationary point of the given optimization problem. Otherwise, we have the following result.

**Proposition 3.2.** *Consider a fixed iteration $k$ and assume that $w^k$ is not an M-stationary point of (3.1). Then the inner loop in (S.2) of Algorithm 3.1 is finite, i.e., we have $\gamma_k = \gamma_{k,i_k}$ for some finite index $i_k \in \{0, 1, 2, \ldots\}$.*

Let $w^0 \in D$ be the starting point from Algorithm 3.1, and let

$$\mathcal{L}_\varphi(w^0) := \big\{w \in D \,\big|\, \varphi(w) \leq \varphi(w^0)\big\}$$

denote the corresponding (feasible) sublevel set. Then the following observation holds, see [28, 39, 59] for the details.

**Proposition 3.3.** *Let $\{w^k\}$ be a sequence generated by Algorithm 3.1. Assume that $\varphi$ is bounded from below and uniformly continuous on $\mathcal{L}_\varphi(w^0)$. Then $\|w^{k+1} - w^k\| \to 0$ holds as $k \to \infty$.*

The previous result allows to prove the following main convergence result for Algorithm 3.1, see, again, [39] for a complete proof.

**Theorem 3.4.** *Let $\{w^k\}$ be a sequence generated by Algorithm 3.1. Assume that $\varphi$ is bounded from below and uniformly continuous on $\mathcal{L}_\varphi(w^0)$. Suppose that $\bar{w}$ is an accumulation point of $\{w^k\}$, i.e., $w^k \to_K \bar{w}$ along a subsequence $K$. Then $\bar{w}$ is an M-stationary point of the optimization problem (3.1), and we have $\gamma_k \left(w^{k+1} - w^k\right) \to_K 0$.*

Observe that (3.4) implies

$$\gamma_{k-1}\big(w^{k-1} - w^k\big) + \nabla\varphi(w^k) - \nabla\varphi(w^{k-1}) \in \nabla\varphi(w^k) + \mathcal{N}_D^{\lim}(w^k)$$

and this justifies the termination criterion (evaluated in iterations $k > 0$)

$$\|\gamma_{k-1}\big(w^{k-1} - w^k\big) + \nabla\varphi(w^k) - \nabla\varphi(w^{k-1})\| \leq \varepsilon_{\text{tol}} \tag{3.5}$$

with $\varepsilon_{\text{tol}} > 0$ for Algorithm 3.1, since the condition $0 \in \nabla\varphi(w^k) + \mathcal{N}_D^{\lim}(w^k)$ encodes M-stationarity of $w^k$ for (3.1). Thus, (3.5) means that $w^k$ is approximately M-stationary. Moreover, Proposition 3.3 and Theorem 3.4 imply that this condition is satisfied for $k \in K$ large enough since $w^k, w^{k-1} \to_K \bar{w}$ and $\nabla\varphi\colon \mathbb{W} \to \mathbb{W}$ is continuous.

Finally, we mention that it may happen that some iterate $w^k$ is already M-stationary. However, this is not easy to detect algorithmically. Then Proposition 3.2 does not apply and the inner iteration might not terminate. To circumvent this pathological situation, one could add the termination criterion

$$\|\gamma_{k,i}\big(w^{k,i} - w^k\big) + \nabla\varphi(w^k) - \nabla\varphi(w^{k,i})\| \leq \varepsilon_{\text{tol}} \tag{3.6}$$

to the inner loop in (S.2). If this condition is satisfied, the entire Algorithm 3.1 terminates with the approximately M-stationary point $w^{k,i}$. The proof of Proposition 3.2 shows that any sufficiently large $i$ satisfies (3.2) or (3.6), even if $w_k$ is M-stationary.

# 4 An Augmented Lagrangian Approach for Structured Geometric Constraints

Section 4.1 contains a detailed statement of our augmented Lagrangian method applied to the general class of problems (P) together with several explanations. The convergence theory is then presented in Section 4.2.

## 4.1 Statement of the Algorithm

We now consider the optimization problem (P) under the given smoothness and convexity assumptions stated there (recall that $D$ is not necessarily convex). This section presents a safeguarded augmented Lagrangian approach for the solution of (P). The method penalizes the constraints $G(w) \in C$, but leaves the possibly complicated condition $w \in D$ explicitly in the constraints. Hence, the resulting subproblems that have to be solved in the augmented Lagrangian framework have exactly the structure of the (simplified) optimization problems discussed in Section 3.

To be specific, consider the (partially) augmented Lagrangian

$$\mathcal{L}_\rho(w,\lambda) := f(w) + \frac{\rho}{2} d_C^2\left(G(w) + \frac{\lambda}{\rho}\right) \tag{4.1}$$

of (P), where $\rho > 0$ denotes the penalty parameter. Note that the squared distance function of a convex set is always smooth which yields that $\mathcal{L}_\rho(\cdot,\lambda)$ is a continuously differentiable mapping. Using the definition of the distance, we can alternatively write this (partially) augmented Lagrangian as

$$\mathcal{L}_\rho(w,\lambda) = f(w) + \frac{\rho}{2}\left\| G(w) + \frac{\lambda}{\rho} - P_C\left(G(w) + \frac{\lambda}{\rho}\right)\right\|^2.$$

In order to control the update of the penalty parameter, we also introduce the auxiliary function

$$V_\rho(w,\lambda) := \left\| G(w) - P_C\left(G(w) + \frac{\lambda}{\rho}\right)\right\| \tag{4.2}$$

which may be viewed as a kind of composite measure of feasibility and complementarity (i.e., $G(w) \in C$ and $\lambda \in \mathcal{N}_C(G(w))$) at the current point. The overall method then is as follows.

**Algorithm 4.1.** *(Safeguarded Augmented Lagrangian Method for Geometric Constraints)*

(S.0) *Choose $\rho_0 > 0$, $\beta > 1$, $\eta \in (0,1)$, $w^0 \in D$, and a nonempty, bounded set $U \subset \mathbb{Y}$. Set $k := 0$.*

(S.1) *If $w^k$ satisfies a suitable termination criterion: STOP.*

(S.2) *Choose $u^k \in U$ and compute an approximate M-stationary point $w^{k+1}$ of the subproblem*

$$\min_w \ \mathcal{L}_{\rho_k}(w,u^k) \quad s.t. \quad w \in D \tag{4.3}$$

*satisfying*

$$\varepsilon^{k+1} \in \nabla_w \mathcal{L}_{\rho_k}(w^{k+1}, u^k) + \mathcal{N}_D^{\lim}(w^{k+1})$$

*for some suitable (sufficiently small) vector* $\varepsilon^{k+1} \in \mathbb{W}$.

(S.3) *Set*

$$\lambda^{k+1} := \rho_k \left[ G(w^{k+1}) + \frac{u^k}{\rho_k} - P_C\left( G(w^{k+1}) + \frac{u^k}{\rho_k} \right) \right]. \tag{4.4}$$

(S.4) *If* $k = 0$ *or*

$$V_{\rho_k}(w^{k+1}, u^k) \leq \eta V_{\rho_{k-1}}(w^k, u^{k-1}),$$

*then set* $\rho_{k+1} := \rho_k$, *else update* $\rho_{k+1} := \beta \rho_k$.

(S.5) *Set* $k \leftarrow k+1$, *and go to* (S.1).

Throughout our convergence analysis, we assume implicitly that Algorithm 4.1 does not stop after finitely many iterations. Since we will prove (under suitable assumptions) convergence to M-stationary points of the optimization problem (P), this means that the termination criterion in (S.1) could check whether the M-stationarity condition holds (at least approximately). Another reasonable approach would be to check approximate feasibility w.r.t. the penalized constraint $G(w) \in C$. Step (S.2), in general, contains the main computational effort since we have to "solve" a constrained nonlinear program at each iteration. Due to the nonconvexity of this subproblem, we only require to compute an M-stationary point of this program. In fact, we allow the computation of an inexact M-stationary point, with the vector $\varepsilon^{k+1}$ measuring the degree of inexactness. The choice $\varepsilon^{k+1} = 0$ corresponds to an exact M-stationary point. Note that the subproblems arising in (S.2) have precisely the structure of the problem investigated in Section 3, hence, the spectral gradient method discussed there is a canonical candidate for the solution of these subproblems (note also that the objective function $\mathcal{L}_{\rho_k}(\cdot, u^k)$ is once, but usually not twice continuously differentiable).

Note that Algorithm 4.1 is called a safeguarded augmented Lagrangian method due to the appearance of the auxiliary sequence $\{u^k\}$. In fact, if we would replace $u^k$ by $\lambda^k$ in (S.2) (and the corresponding subsequent formulas), we would obtain the classical augmented Lagrangian method. However, the safeguarded version has superior global convergence properties, see [15] for a general discussion and [43] for an explicit (counter-) example. In practice, $u^k$ is typically chosen to be equal to $\lambda^k$ as long as this vector belongs to the set $U$, otherwise $u^k$ is taken as the projection of $\lambda^k$ onto this set. In situations where $\mathbb{Y}$ is equipped with some (partial) order relation $\lesssim$, a typical choice for $U$ is given by the box $[u_{\min}, u_{\max}] := \{u \in \mathbb{Y} \,|\, u_{\min} \lesssim u \lesssim u_{\max}\}$ where $u_{\min}, u_{\max} \in \mathbb{Y}$ are given bounds satisfying $u_{\min} \lesssim u_{\max}$.

In order to understand the update of the Lagrange multiplier estimate in (S.3), recall that the augmented Lagrangian is differentiable, with its derivative given by

$$\nabla_w \mathcal{L}_\rho(w, \lambda) = \nabla f(w) + \rho G'(w)^* \left[ G(w) + \frac{\lambda}{\rho} - P_C\left( G(w) + \frac{\lambda}{\rho} \right) \right].$$

Hence, if we denote the usual (partial) Lagrangian of (P) by

$$\mathcal{L}(w, \lambda) := f(w) + \langle \lambda, G(w) \rangle,$$

we obtain from (S.3) that

$$\nabla_w \mathcal{L}_{\rho_k}(w^{k+1}, u^k) = \nabla f(w^{k+1}) + G'(w^{k+1})^* \lambda^{k+1} = \nabla_w \mathcal{L}(w^{k+1}, \lambda^{k+1}). \tag{4.5}$$

This formula is actually the motivation for the precise update used in (S.3).

The particular updating rule in (S.4) is quite common, but other formulas might also be possible. In particular, one can use a different norm in the definition (4.2) of $V_\rho$. Exemplary, we exploited the maximum-norm for our experiments in Section 6 where $\mathbb{W}$ is a space of real vectors or matrices. Let us emphasize that increasing the penalty parameter $\rho_k$ based on a pure infeasibility measure does not work in Algorithm 4.1. One usually has to take into account both the infeasibility of the current iterate (w.r.t. the constraint $G(w) \in C$) and a kind of complementarity condition (i.e., $\lambda \in \mathcal{N}_C(G(w))$).

## 4.2 Convergence

Like all penalty-type methods, augmented Lagrangian methods suffer from the drawback that they generate accumulation points which are not necessarily feasible for the given optimization problem (P). The following (standard) result therefore presents some conditions under which it is guaranteed that limit points are feasible.

**Proposition 4.2.** *Each accumulation point $\bar{w}$ of a sequence $\{w^k\}$ generated by Algorithm 4.1 is feasible for the optimization problem (P) if one of the following conditions holds:*

(a) *$\{\rho_k\}$ is bounded, or*

(b) *there exists some $B \in \mathbb{R}$ such that $\mathcal{L}_{\rho_k}(w^{k+1}, u^k) \leq B$ holds for all $k \in \mathbb{N}$.*

**Proof:** (a) Since $\{\rho_k\}$ is bounded, (S.4) implies that $V_{\rho_k}(w^{k+1}, u^k) \to 0$ for $k \to \infty$. This implies

$$d_C(G(w^{k+1})) \leq \left\| G(w^{k+1}) - P_C\left(G(w^{k+1}) + \frac{u^k}{\rho_k}\right) \right\| = V_{\rho_k}(w^{k+1}, u^k) \to 0.$$

Now, let $\bar{w}$ be an arbitrary accumulation point and, say, $\{w^{k+1}\}_K$ a corresponding subsequence with $w^{k+1} \to_K \bar{w}$. A continuity argument yields $d_C(G(\bar{w})) = 0$. Since $C$ is a closed set, this implies $G(\bar{w}) \in C$. Furthermore, by construction, we have $w^{k+1} \in D$ for all $k \in \mathbb{N}$, so that the closedness of $D$ also yields $\bar{w} \in D$. Altogether, this shows that $\bar{w}$ is feasible for the optimization problem (P).

(b) In view of part (a), it suffices to consider the situation where $\rho_k \to \infty$. By assumption, we have

$$f(w^{k+1}) + \frac{\rho_k}{2} d_C^2\left(G(w^{k+1}) + \frac{u^k}{\rho_k}\right) \leq B \quad \forall k \in \mathbb{N}.$$

Rearranging terms yields

$$d_C^2\left(G(w^{k+1}) + \frac{u^k}{\rho_k}\right) \leq \frac{2(B - f(w^{k+1}))}{\rho_k} \quad \forall k \in \mathbb{N}. \tag{4.6}$$

Let $\bar{w}$ be once again an accumulation point and $\{w^{k+1}\}_K$ be a convergent subsequence with limit $\bar{w}$. Then, taking the limit $k \to_K \infty$ in (4.6) and using the boundedness of $\{u^k\}$, we obtain

$$d_C^2\big(G(\bar{w})\big) = \lim_{k \to_K \infty} d_C^2 \left( G(w^{k+1}) + \frac{u^k}{\rho_k} \right) = 0$$

by a continuity argument. Similar to part (a), this implies feasibility of $\bar{w}$. □

The two conditions in (a) and (b) of Proposition 4.2 are, of course, difficult to check a priori. Nevertheless, in the situation where each iterate $w^{k+1}$ is actually a global minimizer of the subproblem in (S.2) and $w$ denotes any feasible point of the optimization problem (P), we have

$$\mathcal{L}_{\rho_k}(w^{k+1}, u^k) \le \mathcal{L}_{\rho_k}(w, u^k) \le f(w) + \frac{\|u^k\|^2}{2\rho_k} \le f(w) + \frac{\|u^k\|^2}{2\rho_0} \le B$$

for some suitable constant $B$ due to the boundedness of the sequence $\{u^k\}$. The same argument also works if $w^{k+1}$ is only an inexact global minimizer.

The next result shows that, even in the case where a limit point is not necessarily feasible, it still contains some useful information in the sense that it is at least a stationary point for the constraint violation. In general, this is the best that one can expect.

**Proposition 4.3.** *Suppose that the sequence $\{\varepsilon^k\}$ in Algorithm 4.1 is bounded. Then each accumulation point $\bar{w}$ of a sequence $\{w^k\}$ generated by Algorithm 4.1 is an M-stationary point of the so-called feasibility problem*

$$\min_w \tfrac{1}{2} d_C^2(G(w)) \quad s.t. \quad w \in D. \tag{4.7}$$

**Proof:** In view of Proposition 4.2, if $\{\rho_k\}$ is bounded, then each accumulation point is a global minimum of the feasibility problem (4.7) and, therefore, an M-stationary point of this problem.

Hence, it remains to consider the case where $\{\rho_k\}$ is unbounded, i.e., we have $\rho_k \to \infty$ as $k \to \infty$. In view of (S.2) and (S.3), see also (4.5), we have

$$\varepsilon^{k+1} \in \nabla f(w^{k+1}) + G'(w^{k+1})^* \lambda^{k+1} + \mathcal{N}_D^{\lim}(w^{k+1})$$

with $\lambda^{k+1}$ as in (4.4). Dividing this inclusion by $\rho_k$ and using the fact that $\mathcal{N}_D^{\lim}(w^{k+1})$ is a cone, we therefore get

$$\frac{\varepsilon^{k+1}}{\rho_k} \in \frac{\nabla f(w^{k+1})}{\rho_k} + G'(w^{k+1})^* \left[ G(w^{k+1}) + \frac{u^k}{\rho_k} - P_C \left( G(w^{k+1}) + \frac{u^k}{\rho_k} \right) \right] + \mathcal{N}_D^{\lim}(w^{k+1}).$$

Now, let $\bar{w}$ be an accumulation point and $\{w^{k+1}\}_K$ be a subsequence satisfying $w^{k+1} \to_K \bar{w}$. Then the sequences $\{\varepsilon^{k+1}\}_K$, $\{u^k\}_K$, and $\{\nabla f(w^{k+1})\}_K$ are bounded. Thus, taking the limit $k \to_K \infty$ yields

$$0 \in G'(\bar{w})^* \big[ G(\bar{w}) - P_C(G(\bar{w})) \big] + \mathcal{N}_D^{\lim}(\bar{w})$$

by the outer semicontinuity of the limiting normal cone. Since we also have $\bar{w} \in D$ and due to

$$\nabla\big(\tfrac{1}{2}d_C^2 \circ G\big)(\bar{w}) = G'(\bar{w})^*\big[G(\bar{w}) - P_C(G(\bar{w}))\big],$$

it follows that $\bar{w}$ is an M-stationary point of the feasibility problem (4.7). □

We next investigate suitable properties of feasible limit points. The following may be viewed as the main observation in that respect and shows that any such accumulation point is automatically an AM-stationary point in the sense of Definition 2.2.

**Theorem 4.4.** *Suppose that the sequence $\{\varepsilon^k\}$ in Algorithm 4.1 satisfies $\varepsilon^k \to 0$. Then each feasible accumulation point $\bar{w}$ of a sequence $\{w^k\}$ generated by Algorithm 4.1 is an AM-stationary point.*

**Proof:** Let $\{w^{k+1}\}_K$ denote a subsequence such that $w^{k+1} \to_K \bar{w}$. Define

$$s^{k+1} := P_C\left(G(w^{k+1}) + \frac{u^k}{\rho_k}\right) \quad \text{and} \quad z^{k+1} := G(w^{k+1}) - s^{k+1}$$

for each $k \in \mathbb{N}$. We claim that the four (sub-) sequences $\{w^{k+1}\}_K, \{z^{k+1}\}_K$, $\{\varepsilon^{k+1}\}_K$, and $\{\lambda^{k+1}\}_K$ generated by Algorithm 4.1 or defined in the above way satisfy the properties from Definition 2.2 and therefore show that $\bar{w}$ is an AM-stationary point. By construction, we have $w^{k+1} \to_K \bar{w}$ and $\varepsilon^{k+1} \to_K 0$. Further, from (S.2) and (4.5), we obtain

$$\varepsilon^{k+1} \in \nabla_w \mathcal{L}_{\rho_k}(w^{k+1}, u^k) + \mathcal{N}_D^{\lim}(w^{k+1}) = \nabla f(w^{k+1}) + G'(w^{k+1})^*\lambda^{k+1} + \mathcal{N}_D^{\lim}(w^{k+1}).$$

Since $\mathcal{N}_C(s^{k+1})$ is a cone, the relation between $P_C$ and $\mathcal{N}_C$ together with the definitions of $s^{k+1}, \lambda^{k+1}$, and $z^{k+1}$ yield

$$\lambda^{k+1} = \rho_k\left[G(w^{k+1}) + \frac{u^k}{\rho_k} - s^{k+1}\right] \in \mathcal{N}_C(s^{k+1}) = \mathcal{N}_C(G(w^{k+1}) - z^{k+1}).$$

Hence, it remains to show $z^{k+1} \to_K 0$. To this end, we consider two cases, namely whether $\{\rho_k\}$ stays bounded or is unbounded. In the bounded case, (S.4) implies that $V_{\rho_k}(w^{k+1}, u^k) \to 0$ for $k \to \infty$. The corresponding definitions therefore yield

$$\|z^{k+1}\| = \|G(w^{k+1}) - s^{k+1}\| = V_{\rho_k}(w^{k+1}, u^k) \to 0 \quad \text{for } k \to_K \infty.$$

On the other hand, if $\{\rho_k\}$ is unbounded, we have $\rho_k \to \infty$. Since $\{u^k\}$ is bounded by construction, the continuity of the projection operator together with the assumed feasibility of $\bar{w}$ implies

$$s^{k+1} = P_C\left(G(w^{k+1}) + \frac{u^k}{\rho_k}\right) \to P_C(G(\bar{w})) = G(\bar{w}) \quad \text{for } k \to_K \infty.$$

Consequently, we obtain $z^{k+1} = G(w^{k+1}) - s^{k+1} \to_K 0$ also in this case. Altogether, this implies that $\bar{w}$ is AM-stationary. □

Recalling that, by definition, each AM-stationary point of (P) which is AM-regular must already be M-stationary, we obtain the following corollary.

**Corollary 4.5.** *Suppose that the sequence* $\{\varepsilon^k\}$ *in Algorithm 4.1 satisfies* $\varepsilon^k \to 0$. *Then each feasible and AM-regular accumulation point* $\bar{w}$ *of a sequence* $\{w^k\}$ *generated by Algorithm 4.1 is an M-stationary point.*

This result generalizes [29, Theorem 3] which addresses a similar MPCC-tailored augmented Lagrangian method and exploits an MPCC-tailored version of RCPLD, see Lemma 2.7 as well.

# 5 Realizations

Recall that we need to solve the subproblem

$$\min_w \ \mathcal{L}_{\rho_k}(w^k, u^k) + \langle \nabla_w \mathcal{L}_{\rho_k}(w^k, u^k), w - w^k \rangle + \frac{\gamma_k}{2}\|w - w^k\|^2 \quad \text{s.t.} \quad w \in D$$

with some given $\gamma_k > 0$ in the iterations of Algorithm 3.1, which is a promising candidate to solve the ALM-subproblem (4.3) in Algorithm 4.1. As pointed out in Section 3, the above problem possesses the same solutions as

$$\min_w \ \left\| w - \left( w^k - \frac{1}{\gamma_k} \nabla_w \mathcal{L}_{\rho_k}(w^k, u^k) \right) \right\|^2 \quad \text{s.t.} \quad w \in D,$$

i.e., we need to be able to compute elements of the (possibly multi-valued) projection $\Pi_D\big(w^k - \frac{1}{\gamma_k}\nabla_w \mathcal{L}_{\rho_k}(w^k, u^k)\big)$. Boiling this requirement down to its essentials, we have to be in position to find projections of arbitrary points onto the set $D$ in an efficient way. Subsequently, this will be discussed in the context of several practically relevant settings.

## 5.1 The Disjunctive Programming Case

We consider (P) in the special Setting 2.5 with $\mathbb{X} := \mathbb{R}^n$ and $X := [\ell, u]$ where $\ell, u \in \mathbb{R}^n$ satisfy $-\infty \le \ell_i < u_i \le \infty$ for $i = 1, \ldots, n$. Recall that the set $D$ is given by

$$D = \{(x, y, z) \in \mathbb{R}^n \times \mathbb{R}^{m_3} \times \mathbb{R}^{m_3} \,|\, x \in [\ell, u],\ (y_i, z_i) \in T \quad \forall i \in \{1, \ldots, m_3\}\} \tag{5.1}$$

in this situation. For given $\bar{w} = (\bar{x}, \bar{y}, \bar{z}) \in \mathbb{R}^n \times \mathbb{R}^{m_3} \times \mathbb{R}^{m_3}$, we want to characterize the elements of $\Pi_D(\bar{w})$. Therefore, we consider the optimization problem

$$\min_w \ \tfrac{1}{2}\|w - \bar{w}\|^2 \quad \text{s.t.} \quad w = (x, y, z) \in D. \tag{5.2}$$

We observe that the latter can be decomposed into the $n$ one-dimensional optimization problems

$$\min_{x_i} \ \tfrac{1}{2}(x_i - \bar{x}_i)^2 \quad \text{s.t.} \quad x_i \in [\ell_i, u_i],$$

$i = 1, \ldots, n$, possessing the respective solution $P_{[\ell_i, u_i]}(\bar{x}_i)$, as well as into $m_3$ two-dimensional optimization problems

$$\min_{y_i, z_i} \ \tfrac{1}{2}(y_i - \bar{y}_i)^2 + \tfrac{1}{2}(z_i - \bar{z}_i)^2 \quad \text{s.t.} \quad (y_i, z_i) \in T, \tag{5.3}$$

$i = 1, \dots, m_3$. Due to $T = T_1 \cup T_2$, each of these problems on its own can be decomposed into the two two-dimensional subproblems

$$\min_{y_i, z_i} \ \tfrac{1}{2}(y_i - \bar{y}_i)^2 + \tfrac{1}{2}(z_i - \bar{z}_i)^2 \quad \text{s.t.} \quad (y_i, z_i) \in T_j, \qquad (R(i,j))$$

$j = 1, 2$. In most of the popular settings from disjunctive programming, $(R(i,j))$ can be solved with ease. By a simple comparison of the associated objective function values, we find the solutions of (5.3). Putting the solutions of the subproblems together, we find the solutions of (5.2), i.e., the elements of $\Pi_D(\bar{w})$.

In the remainder of this section, we consider a particularly interesting instance of this setting where $T$ is given by

$$T := \{(s,t) \,|\, s \in [\sigma_1, \sigma_2],\ t \in [\tau_1, \tau_2],\ st = 0\}. \tag{5.4}$$

Here, $-\infty \leq \sigma_1, \tau_1 \leq 0$ and $0 < \sigma_2, \tau_2 \leq \infty$ are given constants. Particularly, we find the decomposition

$$T_1 := [\sigma_1, \sigma_2] \times \{0\}, \qquad T_2 := \{0\} \times [\tau_1, \tau_2]$$

of $T$ in this case. Due to the geometrical shape of the set $T$, one might be tempted to refer to this setting as "box-switching constraints". Note that it particularly covers

- switching constraints ($\sigma_1 = \tau_1 := -\infty$, $\sigma_2 = \tau_2 := \infty$), see [40, 50],
- complementarity constraints ($\sigma_1 = \tau_1 := 0$, $\sigma_2 = \tau_2 := \infty$), see [45, 53], and
- relaxed reformulated cardinality constraints ($\sigma_1 := -\infty$, $\sigma_2 := \infty$, $\tau_1 := 0$, $\tau_2 := 1$), see [18, 20].

We refer the reader to Figure 5.1 for a visualization of these types of constraints.

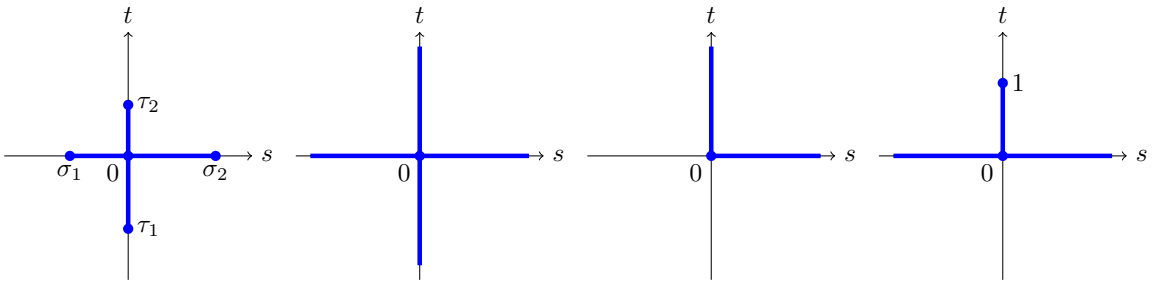


Figure 5.1: Geometric illustrations of box-switching, switching, complementarity, and relaxed reformulated cardinality constraints (from left to right), respectively.

One can easily check that the solutions of $(\mathrm{R}(i,1))$ and $(\mathrm{R}(i,2))$ are given by $(P_{[\sigma_1,\sigma_2]}(\bar{y}_i), 0)$ and $(0, P_{[\tau_1,\tau_2]}(\bar{z}_i))$, respectively. This yields the following result.

**Proposition 5.1.** *Consider the set $D$ from (5.1) where $T$ is given as in (5.4). For given $\bar{w} = (\bar{x}, \bar{y}, \bar{z}) \in \mathbb{R}^n \times \mathbb{R}^{m_3} \times \mathbb{R}^{m_3}$, we have $\hat{w} := (\hat{x}, \hat{y}, \hat{z}) \in \Pi_D(\bar{w})$ if and only if $\hat{x} = P_{[\ell,u]}(\bar{x})$ and*

$$(\hat{y}_i, \hat{z}_i) \in \begin{cases} \{(P_{[\sigma_1,\sigma_2]}(\bar{y}_i), 0)\} & \text{if } \phi_s(\bar{y}_i, \bar{z}_i) < \phi_t(\bar{y}_i, \bar{z}_i), \\ \{(0, P_{[\tau_1,\tau_2]}(\bar{z}_i))\} & \text{if } \phi_s(\bar{y}_i, \bar{z}_i) > \phi_t(\bar{y}_i, \bar{z}_i), \\ \{(P_{[\sigma_1,\sigma_2]}(\bar{y}_i), 0), (0, P_{[\tau_1,\tau_2]}(\bar{z}_i))\} & \text{if } \phi_s(\bar{y}_i, \bar{z}_i) = \phi_t(\bar{y}_i, \bar{z}_i) \end{cases}$$

*for all* $i = 1, \ldots, m_3$, *where we used*

$$\phi_s(a,b) := (P_{[\sigma_1,\sigma_2]}(a) - a)^2 + b^2, \qquad \phi_t(a,b) := a^2 + (P_{[\tau_1,\tau_2]}(b) - b)^2.$$

Particularly, it turns out that in order to compute the projections onto the set $D$ under consideration, one basically needs to compute $n + 2m_3$ projections onto real intervals. In the specific setting of complementarity-constrained programming, this already has been observed in [29, Section 4].

Let us briefly mention that other popular instances of disjunctive programs like vanishing- and or-constrained optimization problems, see e.g. [1, 48], where $T$ is given by

$$T := \{(s,t) \,|\, st \leq 0,\, t \geq 0\} \quad \text{or} \quad T := \{(s,t) \,|\, \min(s,t) \leq 0\},$$

respectively, can be treated in an analogous fashion.

## 5.2 The Sparsity-Constrained Case

We fix $\mathbb{W} := \mathbb{R}^n$ and some $\kappa \in \mathbb{N}$ with $1 \leq \kappa \leq n-1$. Consider the set

$$S_\kappa := \left\{ w \in \mathbb{R}^n \;\middle|\; \|w\|_0 \leq \kappa \right\}$$

with $\|w\|_0$ being the number of nonzero entries of the vector $w$. This set plays a prominent role in sparse optimization and for problems with cardinality constraints. Since $S_\kappa$ is nonempty and closed, projections of some vector $w \in \mathbb{R}^n$ (w.r.t. the Euclidean norm) onto this set exist (but may not be unique), and are known to consist of those vectors $y \in \mathbb{R}^n$ such that the nonzero entries of $y$ are precisely the $\kappa$ largest (in absolute value) components of $w$ (which may not be unique), see e.g. [9, Proposition 3.6].

Hence, within our augmented Lagrangian framework, we may take $D := S_\kappa$ and then get an explicit formula for the solutions of the corresponding subproblems arising within the spectral gradient method. However, typical implementations of augmented Lagrangian methods (like `ALGENCAN`, see [2]) do not penalize box constraints, i.e., they leave the box constraints explicitly as constraints when solving the corresponding subproblems. Hence, let us assume that we have some lower and upper bounds satisfying $-\infty \leq \ell_i < u_i \leq \infty$ for all $i = 1, \ldots, n$. We are then forced to compute projections onto the set

$$D := S_\kappa \cap [\ell, u]. \tag{5.5}$$

It turns out that there exists an explicit formula for this projection. Before presenting the result, let us first assume, for notational simplicity, that

$$0 \in [\ell_i, u_i] \quad \forall i = 1, \ldots, n. \tag{5.6}$$

We mention that this assumption is not restrictive. Indeed, let us assume that, e.g., $0 \notin [\ell_1, u_1]$. Then the first component of $w \in D$ cannot be zero, and this shows

$$D = S_\kappa \cap [\ell, u] = [\ell_1, u_1] \times \left( \hat{S}_{\kappa-1} \cap [\hat{\ell}, \hat{u}] \right), \tag{5.7}$$

where $\hat{S}_{\kappa-1} := \{w \in \mathbb{R}^{n-1} \mid \|w\|_0 \leq \kappa - 1\}$ and the vectors $\hat{\ell}, \hat{u} \in \mathbb{R}^{n-1}$ are obtained from $\ell, u$ by dropping the first component, respectively. For the computation of the projection onto $S_\kappa$, we can now exploit the product structure (5.7). Similarly, we can remove all remaining components $i = 2, \ldots, n$ with $0 \notin [\ell_i, u_i]$ from $D$. Thus, we can assume (5.6) without loss of generality.

We begin with a simple observation.

**Lemma 5.2.** *Let $w \in \mathbb{R}^n$ be arbitrary. Then, for each $y \in \Pi_D(w)$, where $D$ is the set from (5.5), we have*

$$y_i \in \big\{0, P_{[\ell_i,u_i]}(w_i)\big\} \qquad \forall i = 1, \ldots, n.$$

**Proof:** To the contrary, assume that $y_i \neq 0$ and $y_i \neq P_{[\ell_i,u_i]}(w_i)$ hold for some index $i \in \{1, \ldots, n\}$. Define the vector $q \in \mathbb{R}^n$ by $q_j := y_j$ for $j \neq i$ and $q_i := P_{[\ell_i,u_i]}(w_i)$. Due to $y_i \neq 0$, we have $\|q\|_0 \leq \|y\|_0 \leq \kappa$, i.e., $q \in S_\kappa$. Additionally, $q \in [\ell, u]$ is clear from $y \in [\ell, u]$ and $q_i = P_{[\ell_i,u_i]}(w_i)$. Thus, we find $q \in D$. Furthermore, $\|q - w\| < \|y - w\|$ since $q_i = P_{[\ell_i,u_i]}(w_i) \neq y_i$. This contradicts the fact that $y$ is a projection of $w$ onto $D$. □

Due to the above lemma, we only have two choices for the value of the components associated with projections to $D$ from (5.5). Thus, for an arbitrary index set $I \subset \{1, \ldots, n\}$ and an arbitrary vector $w \in \mathbb{R}^n$, we define $p^I(w) \in \mathbb{R}^n$ via

$$p_i^I(w) := \begin{cases} P_{[\ell_i,u_i]}(w_i) & \text{if } i \in I, \\ 0 & \text{otherwise} \end{cases} \qquad \forall i = 1, \ldots, n.$$

It remains to characterize those index sets $I$ which ensure that $p^I(w)$ is a projection of $w$ onto $D$. To this end, we define an auxiliary vector $d(w) \in \mathbb{R}^n$ via

$$d_i(w) := w_i^2 - \big(P_{[\ell_i,u_i]}(w_i) - w_i\big)^2 \quad \forall i = 1, \ldots, n.$$

Note that this definition directly yields

$$\|p^I(w) - w\|^2 = \|w\|^2 - \sum_{i \in I} d_i(w). \tag{5.8}$$

We state the following simple observation.

**Lemma 5.3.** *Fix $w \in \mathbb{R}^n$ and assume that (5.6) is valid. Then the following statements hold:*

(a) $d_i(w) \geq 0$ *for all* $i = 1, \ldots, n$,

(b) $d_i(w) = 0 \iff P_{[\ell_i,u_i]}(w_i) = 0$.

**Proof:** (a) Since $0 \in [\ell_i, u_i]$, we obtain

$$d_i(w) = (w_i - 0)^2 - \big(w_i - P_{[\ell_i,u_i]}(w_i)\big)^2 \geq 0$$

by definition of the (one-dimensional) projection.
(b) If $P_{[\ell_i,u_i]}(w_i) = 0$ holds, we immediately obtain $d_i(w) = 0$. Conversely, let $d_i(w) = 0$. Then

$$0 = w_i^2 - \big(w_i - P_{[\ell_i,u_i]}(w_i)\big)^2 = P_{[\ell_i,u_i]}(w_i)\big(2w_i - P_{[\ell_i,u_i]}(w_i)\big).$$

Hence, we find $P_{[\ell_i,u_i]}(w_i) = 0$ or $P_{[\ell_i,u_i]}(w_i) = 2w_i$. In the first case, we are done. In the second case, we have $\{0, 2w_i\} \subset [\ell_i, u_i]$. By convexity, this gives $w_i \in [\ell_i, u_i]$. Consequently, $w_i = P_{[\ell_i,u_i]}(w_i) = 2w_i$. This implies $P_{[\ell_i,u_i]}(w_i) = 0$. □

Observe that the second assertion of the above lemma implies

$$\|p^I(w)\|_0 = \big|\{i \in I \mid P_{[\ell_i,u_i]}(w_i) \neq 0\}\big| = \big|\{i \in I \mid d_i(w) \neq 0\}\big| \quad \forall w \in \mathbb{R}^n. \tag{5.9}$$

This can be used to characterize the set of projections onto the set $D$ from (5.5).

**Proposition 5.4.** *Let $D$ be the set from (5.5) and assume that (5.6) holds. Then, for each $w \in \mathbb{R}^n$, $y \in \Pi_D(w)$ holds if and only if there exists an index set $I \subset \{1, \dots, n\}$ with $|I| = \kappa$ such that*

$$d_i(w) \geq d_j(w) \quad \forall i \in I,\ \forall j \notin I \tag{5.10}$$

*and $y = p^I(w)$ hold.*

**Proof:** If $y \in \Pi_D(w)$ holds, then $y = p^J(w)$ is valid for some index set $J$, see Lemma 5.2. Thus, it remains to check that $p^J(w)$ is a projection onto $D$ if and only if $p^J(w) = p^I(w)$ holds for some index set $I$ satisfying $|I| = \kappa$ and (5.10).

Note that $p^J(w)$ is a projection if and only if $J$ minimizes $\|p^I(w) - w\|$ over all $I \subset \{1, \dots, n\}$ satisfying $\|p^I(w)\|_0 \leq \kappa$. This can be reformulated via $d(w)$ by using (5.8) and (5.9). In particular, $p^J(w)$ is a projection if and only if $J$ solves

$$\max_I \quad \sum_{i \in I} d_i(w) \quad \text{s.t.} \quad I \subset \{1, \dots, n\}, \quad \big|\{i \in I \mid d_i(w) \neq 0\}\big| \leq \kappa. \tag{5.11}$$

It is clear that index sets $I$ with $|I| = \kappa$ and (5.10) are solutions of this problem. This shows the direction $\Longleftarrow$.

To prove the converse direction $\Longrightarrow$, let $p^J(w)$ be a projection. Thus, $J$ solves (5.11). We note that the solutions of this problem are invariant under addition and removal of indices $i$ with $d_i(w) = 0$. Due to Lemma 5.3 (b), these operations also do not alter the associated $p^I(w)$. Thus, for each projection $p^J(w)$, we can add or remove indices $i$ with $d_i(w) = 0$, to obtain a set $I$ with $p^I(w) = p^J(w)$ and $|I| = \kappa$. It is also clear that (5.10) holds for such a choice of $I$. □

Below, we comment on the result of Proposition 5.4.

*Remark* 5.5. (a) Let $y = p^I(w)$ be a projection of $w \in \mathbb{R}^n$ onto $D$ from (5.5) such that (5.6) holds. Observe that $y_i = 0$ may also hold for some indices $i \notin I$.

(b) In the unconstrained case $[\ell, u] = \mathbb{R}^n$, we find $d_i(w) = w_i^2$ for each $w \in \mathbb{R}^n$ and all $i = 1, \dots, n$. Thus, Proposition 5.4 recovers the well-known characterization of the projection onto the set $S_\kappa$ which can be found in [9, Proposition 3.6].

We want to close this section with some brief remarks regarding the variational geometry of $D$ from (5.5). Observing that the sets $S_\kappa$ and $[\ell, u]$ are both polyhedral in the sense that they can be represented as the union of finitely many polyhedrons, the normal cone intersection rule

$$\mathcal{N}_D^{\lim}(w) = \mathcal{N}_{S_\kappa \cap [\ell,u]}^{\lim}(w) \subset \mathcal{N}_{S_\kappa}^{\lim}(w) + \mathcal{N}_{[\ell,u]}^{\lim}(w) = \mathcal{N}_{S_\kappa}^{\lim}(w) + \mathcal{N}_{[\ell,u]}(w)$$

applies for each $w \in D$ by means of [34, Corollary 4.2] and [56, Proposition 1]. While the evaluation of $\mathcal{N}_{[\ell,u]}(w)$ is standard, a formula for $\mathcal{N}_{S_\kappa}^{\lim}(w)$ can be found in [9, Theorem 3.9].

## 5.3 Low-Rank Approximation

### 5.3.1 General Low-Rank Approximations

For natural numbers $m, n \in \mathbb{N}$ with $m, n \geq 2$, we fix $\mathbb{W} := \mathbb{R}^{m \times n}$. Equipped with the standard Frobenius inner product, $\mathbb{W}$ indeed is a Euclidean space. Now, for fixed $\kappa \in \mathbb{N}$ satisfying $1 \leq \kappa \leq \min(m, n) - 1$, let us investigate the set

$$D := \{W \in \mathbb{W} \,|\, \operatorname{rank} W \leq \kappa\}.$$

Constraint systems involving rank constraints of type $W \in D$ can be used to model numerous practically relevant problems in computer vision, machine learning, computer algebra, signal processing, or model order reduction, see [46, Section 1.3] for an overview. Nowadays, one of the most popular applications behind low-rank constraints is the so-called low-rank matrix completion, particularly, the "Netflix-problem", see [19] for details.

Observe that the variational geometry of $D$ has been explored recently in [36]. Particularly, a formula for the limiting normal cone to this set can be found in [36, Theorem 3.1]. Using the singular value decomposition of a given matrix $\overline{W} \in \mathbb{W}$, one can easily construct an element of $\Pi_D(\overline{W})$ by means of the so-called Eckart–Young–Mirsky theorem, see e.g. [46, Theorem 2.23].

**Proposition 5.6.** *For a given matrix $\overline{W} \in \mathbb{W}$, let $\overline{W} = U\Sigma V^\top$ be its singular value decomposition with orthogonal matrices $U \in \mathbb{R}^{m \times m}$ and $V \in \mathbb{R}^{n \times n}$ as well as a diagonal matrix $\Sigma \in \mathbb{R}^{m \times n}$ whose diagonal entries are in non-increasing order. Let $\widehat{U} \in \mathbb{R}^{m \times \kappa}$ and $\widehat{V} \in \mathbb{R}^{n \times \kappa}$ be the matrices resulting from $U$ and $V$ by deleting the last $m - \kappa$ and $n - \kappa$ columns, respectively. Furthermore, let $\widehat{\Sigma} \in \mathbb{R}^{\kappa \times \kappa}$ be the top left $\kappa \times \kappa$ block of $\Sigma$. Then we have $\widehat{U}\widehat{\Sigma}\widehat{V}^\top \in \Pi_D(\overline{W})$.*

Note that the projection formulas from the previous sections allow a very efficient computation of the corresponding projections, which is in contrast to the projection provided by Proposition 5.6. Though the formula given there is conceptually very simple, its realization requires to compute the singular value decomposition of the given matrix.

### 5.3.2 Symmetric Low-Rank Approximation

Given $n \in \mathbb{N}$ with $n \geq 2$, we consider the set of symmetric matrices $\mathbb{W} := \mathbb{R}^{n\times n}_{\mathrm{sym}}$, still equipped with the Frobenius inner product. Now, for fixed $\kappa \in \mathbb{N}$ satisfying $1 \leq \kappa \leq n$, let us investigate the set

$$D := \{W \in \mathbb{W} \mid W \succeq 0,\ \operatorname{rank} W \leq \kappa\}.$$

Above, the constraint $W \succeq 0$ is used to abbreviate that $W$ has to be positive semidefinite. Constraint systems involving rank constraints of type $W \in D$ arise frequently in several different mathematical models of data science, see [44] for an overview, and Section 6.3 for an application. Note that $\kappa := n$ covers the setting of pure semidefiniteness constraints.

Exploiting the eigenvalue decomposition of a given matrix $\overline{W} \in \mathbb{W}$, one can easily construct an element of $\Pi_D(\overline{W})$.

**Proposition 5.7.** *For a given matrix $\overline{W} \in \mathbb{W}$, we denote by $\overline{W} = \sum_{i=1}^{n} \lambda_i v_i v_i^\top$ its (orthonormal) eigenvalue decomposition with non-increasingly ordered eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ and associated pairwise orthonormal eigenvectors $v_1, \ldots, v_n$. Then we have $\widehat{W} := \sum_{i=1}^{\kappa} \max(\lambda_i, 0) v_i v_i^\top \in \Pi_D(\overline{W})$.*

**Proof:** We define the positive and negative part $\overline{W}^{\pm} := \sum_{i=1}^{n} \max(\pm\lambda_i, 0) v_i v_i^\top$. This yields $\overline{W} = \overline{W}^{+} - \overline{W}^{-}$ and $\langle \overline{W}^{+}, \overline{W}^{-} \rangle = \operatorname{trace}(\overline{W}^{+}\overline{W}^{-}) = 0$. Thus, for each positive semidefinite $B \in \mathbb{W}$, we have

$$\|\overline{W} - B\|^2 = \|\overline{W}^{+} - B\|^2 + \|\overline{W}^{-}\|^2 + 2\langle \overline{W}^{-}, B \rangle \geq \|\overline{W}^{+} - B\|^2 + \|\overline{W}^{-}\|^2.$$

Since the singular value decomposition of $\overline{W}^{+}$ coincides with the eigenvalue decomposition, the right-hand side is minimized by $B = \widehat{W}$, see Proposition 5.6 while noting that we have $\widehat{W} = \overline{W}^{+}$ in case $\kappa = n$. Due to $\langle \overline{W}^{-}, \widehat{W} \rangle = 0$, $B = \widehat{W}$ also minimizes the left-hand side. □

It is clear that the computation of the $\kappa$ largest eigenvalues of $\overline{W} \in \mathbb{W}$ is sufficient to compute an element from the projection $\Pi_D(\overline{W})$. This can be done particularly efficient for small $\kappa$ (note that $\kappa = 1$ holds in our application from Section 6.3).

## 5.4 Extension to Nonsmooth Objectives

For some lower semicontinuous functional $q\colon \mathbb{W} \to \mathbb{R}$, we consider the optimization problem

$$\min_{w}\ f(w) + q(w) \quad \text{s.t.} \quad G(w) \in C. \tag{5.12}$$

Particularly, we do not assume that $q$ is continuous. Exemplary, let us mention the special cases where $q$ is the indicator function of a closed set, counts the nonzero entries of the argument vector (in case $\mathbb{W} := \mathbb{R}^n$), or encodes the rank of the argument matrix (in case $\mathbb{W} := \mathbb{R}^{m\times n}$). In this regard, (5.12) can be used to model real-world applications from e.g. image restoration or signal processing. Necessary optimality

conditions and constraint qualifications addressing (5.12) can be found in [31]. In [22], the authors suggest to handle (5.12) numerically with the aid of an augmented Lagrangian method (without safeguarding) based on the (partially) augmented Lagrangian function (4.1) and the subproblems

$$\min_{w} \ \mathcal{L}_{\rho_k}(w, \lambda^k) + q(w) \quad \text{s.t.} \quad w \in \mathbb{W}$$

which are solved with a nonmonotone proximal gradient method inspired by [59]. In this regard, the solution approach to (5.12) described in [22] possesses some parallels to our strategy for the numerical solution of (P). The authors in [22] were able to prove convergence of their method to reasonable stationary points of (5.12) under a variant of the basic constraint qualification and RCPLD. Let us mention that the authors in [22,31] only considered standard inequality and equality constraints, but the theory in these papers can be easily extended to the more general constraints considered in (5.12) doing some nearby adjustments.

We note that (P) can be interpreted as a special instance of (5.12) where $q$ plays the role of the indicator function of the set $D$. Then the nonmonotone proximal gradient method from [22] reduces to the spectral gradient method from Section 3. However, the authors in [22] did not challenge their method with discontinuous functionals $q$ and, thus, cut away some of the more reasonable applications behind the model (P). Furthermore, we would like to mention that (5.12) can be reformulated (by using the epigraph $\operatorname{epi} q := \{(w, \alpha) \mid q(w) \leq \alpha\}$ of $q$) as

$$\min_{w,\alpha} \ f(w) + \alpha \quad \text{s.t.} \quad G(w) \in C, \ (w, \alpha) \in \operatorname{epi} q \tag{5.13}$$

which is a problem of type (P). One can easily check that (5.12) and (5.13) are equivalent in the sense that $\bar{w} \in \mathbb{W}$ is a local/global minimizer of (5.12) if and only if $(\bar{w}, q(\bar{w}))$ is a local/global minimizer of (5.13). Problem (5.13) can be handled with Algorithm 4.1 as soon as the computation of projections onto $D := \operatorname{epi} q$ is possible in an efficient way. Our result from Corollary 4.5 shows that Algorithm 4.1 applied to (5.13) computes M-stationary points of (5.12) under a problem-tailored version of AM-regularity, i.e., we are in position to find points satisfying

$$0 \in \nabla f(\bar{w}) + \partial q(\bar{w}) + G'(\bar{w})^* \mathcal{N}_C(G(\bar{w}))$$

under a very mild condition which enhances [22, Theorem 3.1]. Here, we used the limiting subdifferential of $q$ given by

$$\partial q(w) := \{\xi \in \mathbb{W} \,|\, (\xi, -1) \in \mathcal{N}^{\lim}_{\operatorname{epi} q}(w, q(w))\}.$$

# 6 Numerical Results

We implemented Algorithm 4.1, based on the underlying subproblem solver Algorithm 3.1, in MATLAB and tested it on three classes of difficult problems which are discussed in Sections 6.1 to 6.3. All test runs use the following parameters:

$$\tau := 2, \ \sigma := 10^{-4}, \ m := 10, \ \beta := 10, \ \eta := 0.8.$$

While we use $\gamma_{\min} := 10^{-10}$ and $\gamma_{\max} := 10^{10}$ for our experiments in Sections 6.1 and 6.2, $\gamma_{\min} := 10^{-3}$ and $\gamma_{\max} := 10^{3}$ are exploited in Section 6.3. In iteration $k$ of Algorithm 4.1, we terminate Algorithm 3.1 at the iteration $i$ if the inner iterates $w^i$ satisfy

$$\|\gamma_{i-1}(w^{i-1} - w^i) + \nabla\varphi(w^i) - \nabla\varphi(w^{i-1})\|_\infty \leq \frac{10^{-4}}{\sqrt{k+1}},$$

where $\|\cdot\|_\infty$ stands for the maximum-norm for both $\mathbb{W}$ equal to $\mathbb{R}^n$ and equal to $\mathbb{R}^{n\times n}_{\text{sym}}$ (other Euclidean spaces do not occur in the subsequent applications), see (3.5). The outer iteration stops as soon as the infeasibility of the current iterate is less than $10^{-4}$ (in the infinity norm). This is motivated by the fact that a stationary point of the subproblem, which is also feasible for the given optimization problem, is already a stationary point of the original problem. Similarly, we use the infinity norm in the definition (4.2) of $V_\rho$.

Given an arbitrary (possibly random) starting point $w^0$, we note that we first project this point onto the set $D$ and then use this projected point as the true starting point, so that all iterates $w^k$ generated by Algorithm 4.1 belong to $D$. The choice of the initial penalty parameter is similar to the rule in [15, p. 153] and given by

$$\rho_0 := P_{[10^{-3},10^3]}\left(10\frac{\max(1, f(w^0))}{\max\left(1, \frac{1}{2}d_C^2(G(w^0))\right)}\right).$$

In all our examples, the space $\mathbb{Y}$ is given by $\mathbb{R}^m$ as in Setting 2.4. This allows us to choose the safeguarded multiplier estimate $u^k$ as the projection of the current value $\lambda^k$ onto a given box $[u_{\min}, u_{\max}]$, where this box is (in componentwise fashion) chosen to be $[-10^{20}, 10^{20}]$ for all equality constraints and $[0, 10^{20}]$ for all inequality constraints. In this way, we basically guarantee that the safeguarded augmented Lagrangian method from Algorithm 4.1 coincides with the classical approach as long as bounded multiplier estimates $\lambda^k$ are generated.

## 6.1 MPCC Examples

The specification of Algorithm 4.1 is essentially the method discussed in [29], where extensive numerical results (including comparisons with other methods) are presented. We therefore keep this section short and consider only two particular examples in order to illustrate certain aspects of our method.

**Example 6.1.** Here, for $w := (y, z) \in \mathbb{R}^2$, we consider the two-dimensional MPCC given by

$$\min_w \tfrac{1}{2}(y-1)^2 + \tfrac{1}{2}(z-1)^2 \quad \text{s.t.} \quad y + z \leq 2,\ y \geq 0,\ z \geq 0, yz = 0,$$

which is essentially the example from [58] with an additional (inactive) inequality constraint in order to have at least one standard constraint, so that Algorithm 4.1 does not automatically reduce to the spectral gradient method. The problem possesses two global minimizers at $(0, 1)$ and $(1, 0)$ which are M-stationary (in fact, they are even

strongly stationary in the MPCC-terminology). Moreover, it has a local maximizer at $(0,0)$ which is a point of attraction for many MPCC solvers since it can be shown to be C-stationary, see e.g. [35] for the corresponding definitions and some convergence results to C- and M-stationary points.

In view of our convergence theory, Algorithm 4.1 should not converge to the origin. To verify this statement numerically, we generated 1000 random starting points (uniformly distributed) from the box $[-10,10]^2$ and then applied Algorithm 4.1 to the above example. As expected, the method converges for all 1000 starting points to one of the two minima. Moreover, we can even start our method at the origin, and the method still converges to the point $(1,0)$ or $(0,1)$. The limit point itself depends on our choice of the projection which is not unique for iterates $(y^k, z^k)$ with $y^k = z^k > 0$.

The next example is used to illustrate a limitation of our approach which is based on the fact that we use the spectral gradient method as a subproblem solver. There are examples where this spectral gradient method reduces the number of iterations even for two-dimensional problems from more than 100000 to just a few iterations. Nevertheless, in the end, the spectral gradient method is a projected gradient method, which exploits a different stepsize selection, but which eventually reduces to a standard projected gradient method if there are a number of consecutive iterations with very small progress. This situation typically happens for problems which are ill-conditioned, and we illustrate this observation by the following example.

**Example 6.2.** We consider the optimal control of a discretized obstacle problem as investigated in [32, Section 7.4]. Using $w := (x, y, z)$, in our notation, the problem is given by

$$\begin{aligned} \min_{w} \quad & f(w) := \tfrac{1}{2}\|x\|^2 - e^\top y + \tfrac{1}{2}\|y\|^2 \\ \text{s.t.} \quad & x \geq 0,\ -Ay - x + z = 0,\ y \geq 0,\ z \geq 0,\ y^\top z = 0. \end{aligned}$$

Here, $A$ is a tridiagonal matrix which arises from a discretization of the Laplace operator in one dimension, i.e., $a_{ii} = 2$ for all $i$ and $a_{ij} = -1$ for all $i = j \pm 1$. Furthermore, $e$ denotes the all-one vector of appropriate size. We note that $\bar{w} := 0$ is the global minimizer as well as an M-stationary point of this program. Viewing the constraint $x \geq 0$ as a box constraint, taking a moderate discretization with $A \in \mathbb{R}^{64\times 64}$, and using the all-one vector as a starting point, we obtain the results from Table 6.1. The method terminates after 12 outer iterations, which is a reasonable number, especially taking into account that the final penalty parameter $\rho_k$ is relatively large, so that several subproblems with different values of $\rho_k$ have to be solved in the intermediate steps. On the other hand, the number of inner iterations $i$ (at each outer iteration $k$) is very large. In the final step, the method requires more than one million inner iterations. This is a typical behavior of gradient-type methods and indicates that the underlying subproblems are ill-conditioned. This is also reflected by the fact that the stepsize $t_k := 1/\gamma_k$ tends to zero.

Hence, there are two types of difficulties in Example 6.2: there are challenging constraints (the complementarity constraints), and there is an ill-conditioning. The

| $k$ | $i$ | $f(w^k)$ | feasibility | $t_k$ | $\rho_k$ |
|---|---|---|---|---|---|
| 0 | 0 | 32.0000000000 | — | — | 320 |
| 1 | 4846 | -30.2322093464 | 0.0178853 | 0.00020734 | 320 |
| 2 | 2993 | -29.5693121774 | 0.0107722 | 0.00020962 | 320 |
| 3 | 2951 | -29.1713706474 | 0.0083671 | 0.00038935 | 320 |
| 4 | 2748 | -28.8787590641 | 0.0070769 | 0.00019331 | 3200 |
| 5 | 16344 | -27.6160748446 | 0.0038450 | 0.00002223 | 3200 |
| 6 | 16170 | -26.8702060815 | 0.0026747 | 0.00001993 | 3200 |
| 7 | 17782 | -26.4929699435 | 0.0024367 | 0.00001961 | 32000 |
| 8 | 130226 | -25.3129109259 | 0.0023566 | 0.00000198 | 320000 |
| 9 | 602944 | -13.1312397442 | 0.0008675 | 0.00000020 | 320000 |
| 10 | 755631 | -5.3024300847 | 0.0003160 | 0.00000020 | 320000 |
| 11 | 908277 | -2.0002209108 | 0.0001151 | 0.00000029 | 320000 |
| 12 | 1084222 | -0.7376636823 | 0.0000419 | 0.00000020 | 320000 |

Table 6.1: Numerical results for Example 6.2.

difficult constraints are treated by Algorithm 4.1 successfully, but the ill-conditioning causes some problems when solving the resulting subproblems. In principle, this difficulty can be circumvented by using another subproblem solver (like a semismooth Newton method), but then it is no longer guaranteed that we obtain M-stationary points at the limit.

Despite the fact that the ill-conditioning causes some difficulties, we stress again that each iteration of the spectral gradient method is extremely cheap. Moreover, for all test problems in the subsequent sections, we put an upper bound of 50000 inner iterations (as a safeguard), and this upper bound was not reached in any of these examples.

## 6.2 Cardinality-Constrained Problems

We first consider an artificial example to illustrate the convergence behavior of Algorithm 4.1 for cardinality-constrained problems.

**Example 6.3.** Consider the example

$$\min_w \ f(w) := \tfrac{1}{2} w^\top Q w + c^\top w \quad \text{s.t.} \quad e^\top w \le 8, \ \|w\|_0 \le 2,$$

where $Q := E + I$ with $E \in \mathbb{R}^{5\times 5}$ being the all one matrix, $I \in \mathbb{R}^{5\times 5}$ the identity matrix, and $c := -(3, 2, 3, 12, 5)^\top \in \mathbb{R}^5$. This is a minor modification of an example from [10], to which we added an (inactive) inequality constraint for the same reason as in Example 6.1. Taking into account that there are $\binom{5}{2}$ possibilities to choose two possibly nonzero components of $w$, an elementary calculation shows that there are exactly 10 M-stationary points $w^1, \ldots, w^{10}$ which are given in Table 6.2 together with the corresponding function values. It follows that $w^6$ is the global minimizer. The points $w^3, w^8, w^{10}$ have function values which are not too far away from $f(w^6)$, whereas all other M-stationary points have significantly larger function values. We

then took 1000 random starting points from the box $[-10, 10]^5$ (uniformly distributed) and applied Algorithm 4.1 to this example. Surprisingly, the method converged, for all 1000 starting points, to the global minimizer $w^6$. We then changed the example by putting an upper bound $u_4 := 0$ to the fourth component. This excludes the four most interesting points $w^3, w^6, w^8$, and $w^{10}$. Among the remaining points, the three vectors $w^4, w^7$, and $w^9$ have identical function values. Running our program again using 1000 randomly generated starting points, we obtain convergence to $w^4$ in 609 cases, convergence to $w^7$ in 331 situations, whereas in 60 instances only we observe convergence to the non-optimal point $w^2$.

| $w^i$ | $f(w^i)$ | $w^i$ | $f(w^i)$ |
|---|---|---|---|
| $w^1 := (4/3, 1/3, 0, 0, 0)^\top$ | $-2.33$ | $w^6 := (0, -8/3, 0, 22/3, 0)^\top$ | $-41.33$ |
| $w^2 := (1, 0, 1, 0, 0)^\top$ | $-3.00$ | $w^7 := (0, -1/3, 0, 0, 8/3)^\top$ | $-6.33$ |
| $w^3 := (-2, 0, 0, 7, 0)^\top$ | $-39.00$ | $w^8 := (0, 0, -2, 7, 0)^\top$ | $-39.00$ |
| $w^4 := (1/3, 0, 0, 0, 7/3)^\top$ | $-6.33$ | $w^9 := (0, 0, 1/3, 0, 7/3)^\top$ | $-6.33$ |
| $w^5 := (0, 1/3, 4/3, 0, 0)^\top$ | $-2.33$ | $w^{10} := (0, 0, 0, 19/3, -2/3)^\top$ | $-36.33$ |

Table 6.2: M-stationary points and corresponding function values for Example 6.3.

We next consider a class of cardinality-constrained problems of the form

$$\min_w \ \tfrac{1}{2} w^\top Q w \ \text{ s.t. } \ \mu^\top w \geq \varrho, \ e^\top w = 1, \ 0 \leq w \leq u, \ \|w\|_0 \leq \kappa. \tag{6.1}$$

This is a classical portfolio optimization problem, where $Q$ and $\mu$ denote the covariance matrix and the mean of $n$ possible assets, respectively, while $\varrho$ is some lower bound for the expected return. Furthermore, $u$ provides an upper bound for the individual assets within the portfolio. The data $Q, \mu, \varrho, u$ were randomly created by the test problem collection [26], which is available from the webpage `http://www.di.unipi.it/optimize/Data/MV.html`. Here, we used all 30 test instances of dimension $n := 200$ and three different values $\kappa \in \{5, 10, 20\}$ for each problem. We apply three different methods:

(a) Algorithm 4.1 with starting point $w^0 := 0$,

(b) a boosted version of Algorithm 4.1, and

(c) a CPLEX solver [37] to a reformulation of the portfolio optimization problem as a mixed integer quadratic program.

The CPLEX solver is used to (hopefully) identify the global optimum of the optimization problem (6.1). Note that we put a time limit of 0.5 hours for each test problem. Method (a) applies our augmented Lagrangian method to (6.1) using the set $D := \{w \in [0, u] \mid \|w\|_0 \leq \kappa\}$. Projections onto $D$ are computed using the analytic formula from Proposition 5.4. Finally, the boosted version of Algorithm 4.1 is the following: We first delete the cardinality constraint from the portfolio optimization problem. The resulting quadratic program is then convex and can therefore be solved easily. Afterwards, we apply Algorithm 4.1 to a sequence of relaxations of (6.1) in

which the cardinality is recursively decreased by 10 in each step (starting with $n-10$) as long as the desired value $\kappa \in \{5, 10, 20\}$ is not undercut. For $\kappa := 5$, a final call of Algorithm 4.1 with the correct cardinality is necessary. In each outer iteration, the projection of the solution of the previous iteration onto the set $D$ is used as a starting point.

The corresponding results are summarized in Figure 6.1 for the three different values $\kappa \in \{5, 10, 20\}$. This figure compares the optimal function values obtained by the above three methods for each of the thirty test problems. The optimal function values produced by CPLEX are used here as a reference value in order to judge the quality of the results obtained by the other approaches. The main observations are the following: The optimal function value computed by CPLEX is (not surprisingly) always the best one. On the other hand, the corresponding values computed by method (a) are usually not too far away from the optimal ones. Moreover, for all test problems, the boosted version (b) generates even better function values which are usually very close to the ones computed by CPLEX. Of course, if $\kappa$ is taken smaller, the problems are getting more demanding and are therefore more difficult to solve (in general). Hence, the gap between the optimal function values obtained by CPLEX and the other two methods are usually larger. Nevertheless, also for $\kappa := 5$, especially the boosted algorithm still computes rather good points. In this context, one should also note that our methods always terminate with a (numerically) feasible point, hence, the final iterate computed by our method can actually be used as a (good) approximation of the global minimizer. We also would like to mention that our MATLAB implementation of Algorithm 4.1 typically requires, on a Lenovo T490s ThinkPad with an Intel Core i7-8565U processor, only a CPU time of about 0.1 seconds for each of the test problems, whereas the boosted version requires slightly less than a second CPU time in average.

## 6.3 Maxcut Problems

This section considers the famous Maxcut problem as an application of our algorithm to problems with rank constraints. To this end, let $G = (V, E)$ be an undirected graph with vertex set $V = \{1, \dots, n\}$ and edges $e_{ij}$ between vertices $i, j \in V$. We assume that we have a weighted graph, with $a_{ij} = a_{ji}$ denoting the nonnegative weights of the edge $e_{ij}$. Since we allow zero weights, we can assume without loss of generality that $G$ is a complete graph. Now, given a subset $S \subset V$ with complement $S^c$, the *cut* defined by $S$ is the set $\delta(S) := \{e_{ij} \,|\, i \in S, j \in S^c\}$ of all edges such that one end point belongs to $S$ and the other one to $S^c$. The corresponding weight of this cut is defined by

$$w(S) := \sum_{e_{ij} \in \delta(S)} a_{ij}.$$

The Maxcut problem looks for the maximum cut, i.e., a cut with maximum weight. This graph-theoretical problem is known to be NP-hard, thus very difficult to solve.

Let $A := (a_{ij})$ and define $L := \operatorname{diag}(Ae) - A$. Then it is well known, see e.g. [27],

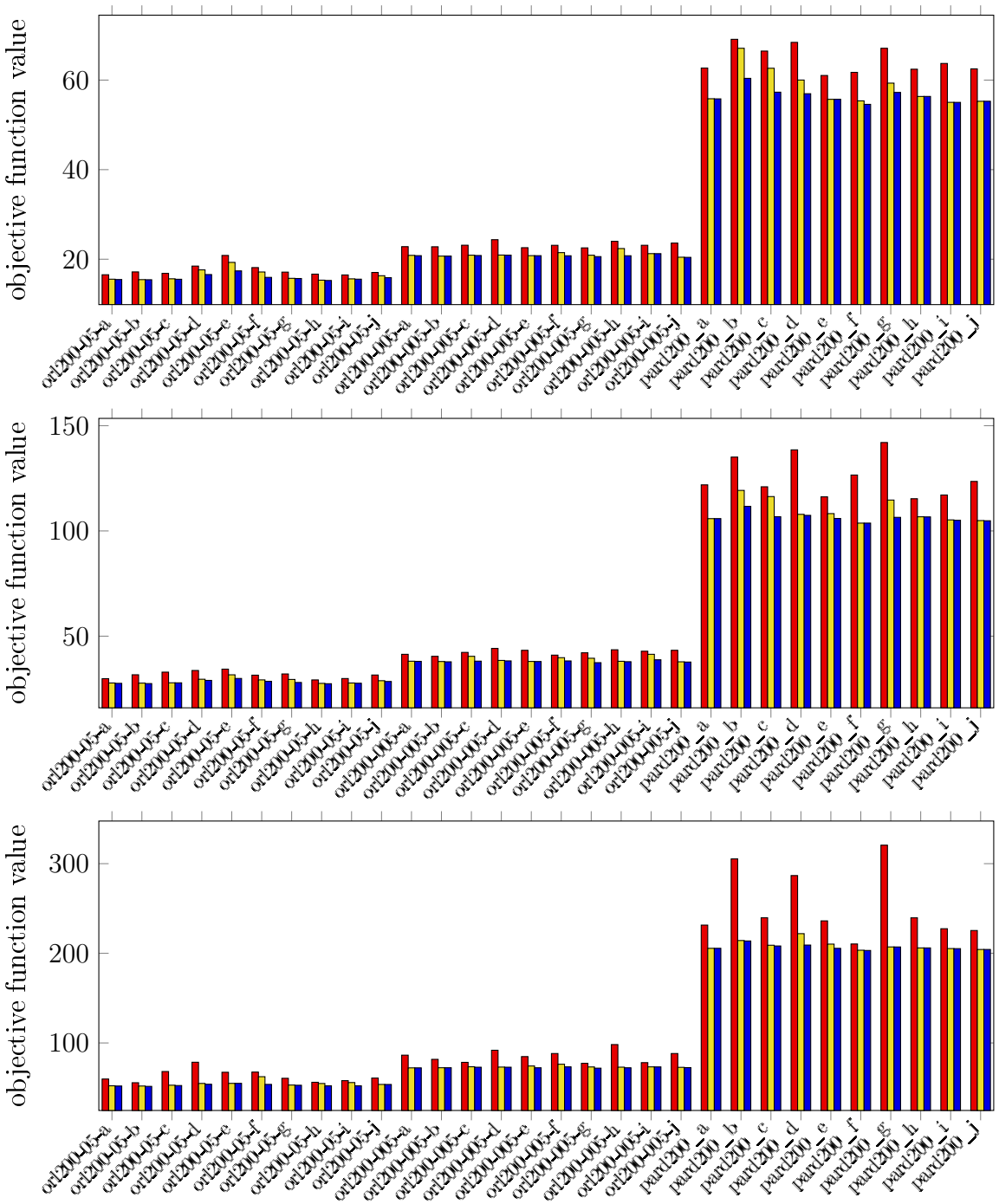


Figure 6.1: Optimal function values obtained by Algorithm 4.1 (red), Algorithm 4.1 with boosting technique (yellow), and CPLEX (blue), applied to the portfolio optimization problem (6.1) with cardinality $\kappa = 20$, $\kappa = 10$, and $\kappa = 5$ (top to bottom).

that the Maxcut problem can be reformulated as

$$\max_W \ \tfrac{1}{4}\operatorname{trace}(LW) \quad \text{s.t.} \quad \operatorname{diag} W = e,\ W \succeq 0,\ \operatorname{rank} W = 1, \tag{6.2}$$

where the variable $W$ is chosen from the space $\mathbb{W} := \mathbb{R}^{n\times n}_{\text{sym}}$. Due to the linear constraint $\operatorname{diag} W = e$, it follows that this problem is equivalent to

$$\max_W \ \tfrac{1}{4}\operatorname{trace}(LW) \quad \text{s.t.} \quad \operatorname{diag} W = e,\ W \succeq 0,\ \operatorname{rank} W \leq 1. \tag{6.3}$$

Deleting the difficult rank constraint, one gets the (convex) relaxation

$$\max_W \ \tfrac{1}{4}\operatorname{trace}(LW) \quad \text{s.t.} \quad \operatorname{diag} W = e,\ W \succeq 0, \tag{6.4}$$

which is a famous test problem for semidefinite programs.

Here, we directly deal with (6.3) by taking $D := \{W \in \mathbb{W} \,|\, W \succeq 0,\ \operatorname{rank} W \leq 1\}$ as the complicated set. Projections onto $D$ can be calculated via Proposition 5.7: Let $W \in \mathbb{W}$ denote an arbitrary symmetric matrix with maximum eigenvalue $\lambda$ and corresponding (normalized) eigenvector $v$ (note that $\lambda$ and $v$ are not necessarily unique), then $\max(\lambda, 0)vv^\top$ is a projection of $W$ onto $D$. In particular, the computation of this projection does not require the full spectral decomposition. Note that it is not clear whether a projection onto $C$ can be computed efficiently. Consequently, we penalize the linear constraint $\operatorname{diag} W = e$ by the augmented Lagrangian approach.

Throughout this section, we take the zero matrix as the starting point. In order to illustrate the performance of our method, we begin with the simple graph from Figure 6.2. Algorithm 4.1 applied to this example using the reformulation (6.3) (more precisely, the corresponding minimization problem) together with the previous specifications yields the iterations shown in Table 6.3.

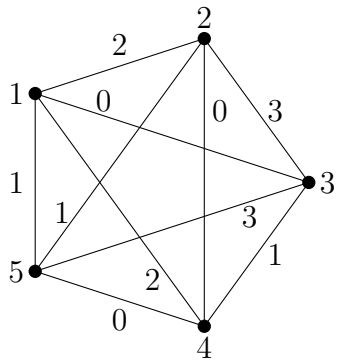


Figure 6.2: Example of a complete graph for the Maxcut problem.

The number of (outer) iterations is denoted by $k$, $i$ is the number of inner iterations, $i_{\text{cum}}$ the accumulated number of inner iterations, $f$-ev. provides the number of function evaluations (note that, due to the Armijo rule, we might have several function evaluations in a single inner iteration, hence, $f$-ev. is always an upper bound for $i_{\text{cum}}$), $f(W^k)$ denotes the current function value, the column titled "feasibility" is

| $k$ | $i$ | $i_{\text{cum}}$ | $f$-ev. | $f(W^k)$ | feasibility | $t_k$ | $\rho_k$ |
|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 1 | 0.00000000 | — | — | 4 |
| 1 | 10 | 10 | 14 | 19.66916669 | 0.839204 | 0.68928 | 4 |
| 2 | 8 | 18 | 24 | 12.03958568 | 0.027375 | 1.24551 | 4 |
| 3 | 5 | 23 | 31 | 12.00975494 | 0.006357 | 1.25001 | 4 |
| 4 | 3 | 26 | 35 | 12.00238060 | 0.001553 | 0.62522 | 4 |
| 5 | 3 | 29 | 39 | 12.00054208 | 0.000382 | 0.62504 | 4 |
| 6 | 3 | 32 | 43 | 12.00015307 | 0.000097 | 0.62502 | 4 |

Table 6.3: Numerical results for Maxcut associated to the graph from Figure 6.2.

the feasibility measure (in the current situation, this is $\| \operatorname{diag} W^k - e\|_\infty$), $t_k := 1/\gamma_k$ is the stepsize, and $\rho_k$ denotes the penalty parameter at iteration $k$. Note that this penalty parameter stays constant for this example. The feasibility measure tends to zero, and we terminate at iteration $k = 6$ since this measure becomes less than $10^{-4}$, i.e., we stop successfully. The associated function value is (approximately) 12 which actually corresponds to the maximum cut $S := \{1, 3\}$ for the graph from Figure 6.2, i.e., our method is able to solve the Maxcut problem for this particular instance.

We next apply our method to two test problem collections that can be downloaded from `http://biqmac.uni-klu.ac.at/biqmaclib.html`, namely the `rudy` and the `ising` collection. The first class of problems consists of 130 instances, whereas the second one includes 48 problems. The optimal function value $f_{\text{opt}}$ of all these examples is known. The details of the corresponding results obtained by our method are given in [39]. Here, we summarize the main observations.

All $130 + 48$ test problems were solved successfully by our method since the standard termination criterion was satisfied after finitely many iterations, i.e., we stop with an iterate $W^k$ which is feasible (within the given tolerance). Hence, the corresponding optimal function value $f_{\text{ALM}}$ is a lower bound for the optimal value $f_{\text{opt}}$. For the sake of completeness, we also solved the (convex) relaxed problem from (6.4), using again our augmented Lagrangian method with $D := \{W \in \mathbb{W} \,|\, W \succeq 0\}$. The corresponding function value is denoted by $f_{\text{SDP}}$. Since the feasible set of (6.4) is larger than the one of (6.3), we have the inequalities $f_{\text{ALM}} \leq f_{\text{opt}} \leq f_{\text{SDP}}$. The corresponding details for the solution of the SDP-relaxation are provided in [39] for the `rudy` collection.

The bar charts from Figures 6.3 and 6.4 summarize the results for the `rudy` and `ising` collections, respectively, in a very condensed way. They basically show that the function value $f_{\text{ALM}}$ obtained by our method is very close to the optimal value $f_{\text{opt}}$. More precisely, the interpretation is as follows: For each test problem, we take the quotient $f_{\text{ALM}}/f_{\text{opt}} \in [0, 1]$. If this quotient is equal to, say, 0.91, we count this example as one where we reach 91% of the optimal function value. Figure 6.3 then says that all 130 test problems were solved with at least 88% of the optimal function value. There are still 101 test examples which are solved with a precision of at least 95%. One third of the test examples, namely 44 problems, are even solved with an accuracy of at least 99%. For three examples (pm1d_80.9, pm1s_100.6, and pw01_100.8), we actually get the exact global maximum.

Figure 6.4 has a similar meaning for the `ising` collection: Though there is no example which is solved exactly, one half of the problems reaches an accuracy of at least 99%, and even in the worst case, we obtain a precision of 95%.

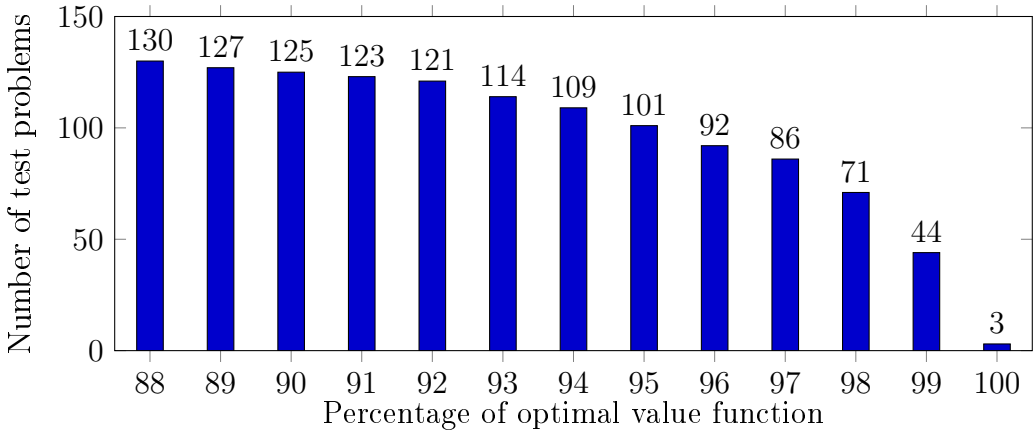


Figure 6.3: Summary of the results from the `rudy` collection.

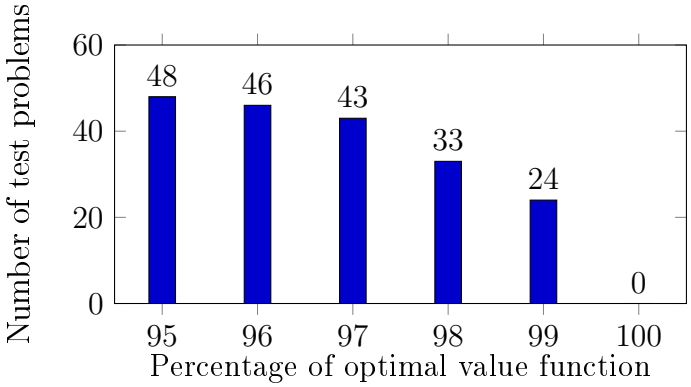


Figure 6.4: Summary of the results from the `ising` collection.

Altogether, this shows that we obtain a very good lower bound for the optimal function value. Moreover, since we are always feasible (in particular, all iterates are matrices of rank one), the final matrix can be used to create a cut through the given graph, i.e., the method provides a constructive way to create cuts which seem to be close to the optimal cuts. Note that this is in contrast to the semidefinite relaxation (6.4) which gives an upper bound, but the solution associated with this upper bound is usually not feasible for the Maxcut problem since the rank constraint is violated (the results in [39] show that the solutions of the relaxed programs for the `rudy` collection are matrices of rank between 4 and 7). In particular, these matrices can, in general, not be used to compute a cut for the graph and, therefore, are less constructive than the outputs of our method. Moreover, it is interesting to observe that $f_{\mathrm{ALM}}$ is usually much closer to $f_{\mathrm{opt}}$ than $f_{\mathrm{SDP}}$. In any case, both techniques together might be useful tools in a branch-and-bound-type method for solving Maxcut problems.

# 7 Concluding Remarks

In this paper, we demonstrated how M-stationary points of optimization problems with structured geometric constraints can be computed with the aid of an augmented Lagrangian method. The fundamental idea was to keep the complicated constraints out of the augmented Lagrangian function and to treat them directly in the associated subproblems which are solved by means of a nonmonotone projected gradient method. This way, the handling of challenging variational structures is encapsulated within the efficient computation of projections. This also puts a natural limit for the applicability. In contrast to several other approaches from the literature, the convergence guarantees for our method, which are valid in the presence of a comparatively weak asymptotic constraint qualification, remain true if the appearing subproblems are solved inexactly. Extensive numerical experiments visualized the quantitative qualities of this approach.

Despite our observations in Example 6.2, it might be interesting to think about extensions of these ideas to infinite-dimensional situations. In [17], an augmented Lagrangian method for the numerical solution of (P) in the context of Banach spaces has been considered where the set $D$ was assumed to be convex, and the subproblems in the resulting algorithm are of the same type as in our paper. Furthermore, convergence of the method to KKT points was shown under validity of a problem-tailored version of asymptotic regularity. As soon as $D$ becomes nonconvex, one has to face some uncomfortable properties of the appearing limiting normal cone which turns out to be comparatively large since weak-$*$-convergence is used for its definition as a set limit in the dual space, see [33, 51]. That it why the associated M-stationarity conditions are, in general, too weak in order to yield a reasonable stationarity condition. However, this issue might be surpassed by investigating the smaller strong limiting normal cone which is based on strong convergence in the dual space but possesses very limited calculus. It remains open whether reasonable asymptotic regularity conditions w.r.t. this variational object can be formulated. Furthermore, in order to exploit the smallness of the strong limiting normal cone in the resulting algorithm, one has to make sure (amongst others) that the (primal) sequence $\{w^k\}$ possesses strong accumulation points while the (dual) measures of inexactness $\{\varepsilon^k\}$ need to be strongly convergent as well. This might be restrictive. Furthermore, it has to be clarified how the subproblems can be solved to approximate strong M-stationarity.

# References


[1] W. Achtziger and C. Kanzow. Mathematical programs with vanishing constraints: optimality conditions and constraint qualifications. *Mathematical Programming, Series A*, 114(1):69–99, 2008. doi:10.1007/s10107-006-0083-3.

[2] R. Andreani, E. G. Birgin, J. M. Martínez, and M. L. Schuverdt. On augmented Lagrangian methods with general lower-level constraints. *SIAM Journal on Optimization*, 18(4):1286–1309, 2008. doi:10.1137/060654797.

[3] R. Andreani, G. Haeser, M. L. Schuverdt, and P. J. S. Silva. A relaxed constant positive linear dependence constraint qualification and applications. *Mathematical Programming*, 135(1):255–273, 2012. doi:10.1007/s10107-011-0456-0.

[4] R. Andreani, G. Haeser, L. D. Secchin, and P. J. S. Silva. New sequential optimality conditions for mathematical programs with complementarity constraints and algorithmic consequences. *SIAM Journal on Optimization*, 29(4):3201–3230, 2019. doi:10.1137/18M121040X.

[5] R. Andreani, J. M. Martínez, A. Ramos, and P. J. S. Silva. A cone-continuity constraint qualification and algorithmic consequences. *SIAM Journal on Optimization*, 26(1):96–110, 2016. doi:10.1137/15M1008488.

[6] R. Andreani, J. M. Martínez, A. Ramos, and P. J. S. Silva. Strict constraint qualifications and sequential optimality conditions for constrained optimization. *Mathematics of Operations Research*, 43(3):693–717, 2018. doi:10.1287/moor.2017.0879.

[7] R. Andreani, L. D. Secchin, and P. Silva. Convergence properties of a second order augmented Lagrangian method for mathematical programs with complementarity constraints. *SIAM Journal on Optimization*, 28(3):2574–2600, 2018. doi:10.1137/17m1125698.

[8] J. Barzilai and J. M. Borwein. Two-point step size gradient methods. *IMA Journal of Numerical Analysis*, 8(1):141–148, 1988. doi:10.1093/imanum/8.1.141.

[9] H. H. Bauschke, D. R. Luke, H. M. Phan, and X. Wang. Restricted normal cones and sparsity optimization with affine constraints. *Foundations of Computational Mathematics*, 14:63–83, 2013. doi:10.1007/s10208-013-9161-0.

[10] A. Beck and Y. C. Eldar. Sparsity constrained nonlinear optimization: optimality conditions and algorithms. *SIAM Journal on Optimization*, 23(3):1480–1509, 2013. doi:10.1137/120869778.

[11] A. Ben-Tal and A. Nemirovski. *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*. SIAM, Philadelphia, 2001. doi:10.1137/1.9780898718829.

[12] M. Benko, M. Červinka, and T. Hoheisel. Sufficient conditions for metric subregularity of constraint systems with applications to disjunctive and ortho-disjunctive programs. *Set-Valued and Variational Analysis*, 2021. doi:10.1007/s11228-020-00569-7.

[13] M. Benko and H. Gfrerer. New verifiable stationarity concepts for a class of mathematical programs with disjunctive constraints. *Optimization*, 67(1):1–23, 2018. doi:10.1080/02331934.2017.1387547.

[14] D. P. Bertsekas. *Constrained Optimization and Lagrange Multiplier Methods*. Academic Press, New York, 1982. doi:10.1016/C2013-0-10366-2.

[15] E. G. Birgin and J. M. Martínez. *Practical Augmented Lagrangian Methods for Constrained Optimization*. SIAM, Philadelphia, 2014. doi:10.1137/1.9781611973365.

[16] E. G. Birgin, J. M. Martínez, and M. Raydan. Nonmonotone spectral projected gradient methods on convex sets. *SIAM Journal on Optimization*, 10(4):1196–1211, 2000. doi:10.1137/s1052623497330963.

[17] E. Börgens, C. Kanzow, P. Mehlitz, and G. Wachsmuth. New constraint qualifications for optimization problems in Banach spaces based on asymptotic KKT conditions. *SIAM Journal on Optimization*, 30(4):2956–2982, 2020. doi:10.1137/19M1306804.

[18] O. P. Burdakov, C. Kanzow, and A. Schwartz. Mathematical programs with cardinality constraints: reformulation by complementarity-type conditions and a regularization method. *SIAM Journal on Optimization*, 26(1):397–425, 2016. doi:10.1137/140978077.

[19] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9:717, 2009. doi:10.1007/s10208-009-9045-5.

[20] M. Červinka, C. Kanzow, and A. Schwartz. Constraint qualifications and optimality conditions for optimization problems with cardinality constraints. *Mathematical Programming*, 160(1):353–377, 2016. doi:10.1007/s10107-016-0986-6.

[21] J.-S. Chen. *SOC Functions and their Applications*. Springer, Singapure, 2019. doi:10.1007/978-981-13-4077-2.

[22] X. Chen, L. Guo, Z. Lu, and J. J. Ye. An augmented Lagrangian method for non-Lipschitz nonconvex programming. *SIAM Journal on Numerical Analysis*, 55(1):168–193, 2017. doi:10.1137/15M1052834.

[23] N. H. Chieu and G. M. Lee. A relaxed constant positive linear dependence constraint qualification for mathematical programs with equilibrium constraints. *Journal of Optimization Theory and Applications*, 158(1):11–32, 2013. doi:10.1007/s10957-012-0227-y.

[24] J. E. Dennis Jr and R. B. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. SIAM, Philadelphia, 1996. doi:10.1137/1.9781611971200.

[25] M. L. Flegel, C. Kanzow, and J. V. Outrata. Optimality conditions for disjunctive programs with application to mathematical programs with equilibrium constraints. *Set-Valued Analysis*, 15(2):139–162, 2007. doi:10.1007/s11228-006-0033-5.

[26] A. Frangioni and C. Gentile. SDP diagonalizations and perspective cuts for a class of nonseparable MIQP. *Operations Research Letters*, 35(2):181–185, 2007. doi:10.1016/j.orl.2006.03.008.

[27] M. X. Goemans and D. P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM*, 42(6):1115–1145, 1995. doi:10.1145/227683.227684.

[28] L. Grippo, F. Lampariello, and S. Lucidi. A nonmonotone line search technique for Newton's method. *SIAM Journal on Numerical Analysis*, 23(4):707–716, 1986. doi:10.1137/0723046.

[29] L. Guo and Z. Deng. A new augmented Lagrangian method for MPCCs - theoretical and numerical comparison with existing augmented Lagrangian methods. *Mathematics of Operations Research*, pages 1–38, 2021. accepted for publication.

[30] L. Guo and G.-H. Lin. Notes on some constraint qualifications for mathematical programs with equilibrium constraints. *Journal of Optimization Theory and Applications*, 156:600–616, 2013. doi:10.1007/s10957-012-0084-8.

[31] L. Guo and J. J. Ye. Necessary optimality conditions and exact penalization for non-Lipschitz nonlinear programs. *Mathematical Programming*, 168:571–598, 2018. doi:10.1007/s10107-017-1112-0.

[32] F. Harder, P. Mehlitz, and G. Wachsmuth. Reformulation of the M-stationarity conditions as a system of discontinuous equations and its solution by a semismooth Newton method. *SIAM Journal on Optimization*, 2021. URL `https://arxiv.org/abs/2002.10124`. accepted for publication.

[33] F. Harder and G. Wachsmuth. The limiting normal cone of a complementarity set in Sobolev spaces. *Optimization*, 67(10):1579–1603, 2018. doi:10.1080/02331934.2018.1484467.

[34] R. Henrion, A. Jourani, and J. V. Outrata. On the calmness of a class of multifunctions. *SIAM Journal on Optimization*, 13(2):603–618, 2002. doi:10.1137/S1052623401395553.

[35] T. Hoheisel, C. Kanzow, and A. Schwartz. Theoretical and numerical comparison of relaxation schemes for mathematical programs with complementarity constraints. *Mathematical Programming*, 137:257–288, 2013. doi:10.1007/s10107-011-0488-5.

[36] S. Hosseini, D. R. Luke, and A. Uschmajew. Tangent and normal cones for low-rank matrices. In S. Hosseini, B. S. Mordukhovich, and A. Uschmajew, editors, *Nonsmooth Optimization and Its Applications*, pages 45–53. Springer, Cham, 2019. doi:10.1007/978-3-030-11370-4_3.

[37] IBM ILOG CPLEX V12.1. User's Manual for CPLEX. International Business Machines Corporation, 2009.

[38] A. F. Izmailov, M. V. Solodov, and E. I. Uskov. Global convergence of augmented Lagrangian methods applied to optimization problems with degenerate constraints, including problems with complementarity constraints. *SIAM Journal on Optimization*, 22(4):1579–1606, 2012. doi:10.1137/120868359.

[39] X. Jia, C. Kanzow, P. Mehlitz, and G. Wachsmuth. An augmented Lagrangian method for optimization problems with structured geometric constraints. Technical report, preprint arXiv, 2021. URL https://arxiv.org/abs/2105.08317.

[40] C. Kanzow, P. Mehlitz, and D. Steck. Relaxation schemes for mathematical programmes with switching constraints. *Optimization Methods and Software*, pages 1–36, 2019. doi:10.1080/10556788.2019.1663425.

[41] C. Kanzow, A. B. Raharja, and A. Schwartz. An augmented Lagrangian method for cardinality-constrained optimization. *Journal of Optimization Theory and Applications*, pages 1–21, 2021. doi:10.1007/s10957-021-01854-7.

[42] C. Kanzow and A. Schwartz. The price of inexactness: convergence properties of relaxation methods for mathematical programs with equilibrium constraints revisited. *Mathematics of Operations Research*, 40(2):253–275, 2015. doi:10.1287/moor.2014.0667.

[43] C. Kanzow and D. Steck. An example comparing the standard and safeguarded augmented Lagrangian methods. *Operations Research Letters*, 45(6):598–603, 2017. doi:10.1016/j.orl.2017.09.005.

[44] A. Lemon, A. M.-C. So, and Y. Ye. Low-rank semidefinite programming: theory and applications. *Foundations and Trends in Optimization*, 2(1-2):1–156, 2016. doi:10.1561/2400000009.

[45] Z.-Q. Luo, J.-S. Pang, and D. Ralph. *Mathematical Programs with Equilibrium Constraints*. Cambridge University Press, Cambridge, 1996. doi:10.1017/cbo9780511983658.

[46] I. Markovsky. *Low Rank Approximation: Algorithms, Implementation, Applications*. Communications and Control Engineering. Springer, London, 2012. doi:10.1007/978-3-319-89620-5.

[47] P. Mehlitz. Asymptotic stationarity and regularity for nonsmooth optimization problems. *Journal of Nonsmooth Analysis and Optimization*, 1:6575, 2020. doi:10.46298/jnsao-2020-6575.

[48] P. Mehlitz. A comparison of solution approaches for the numerical treatment of or-constrained optimization problems. *Computational Optimization and Applications*, 76(1):233–275, 2020. doi:10.1007/s10589-020-00169-z.

[49] P. Mehlitz. On the linear independence constraint qualification in disjunctive programming. *Optimization*, 69(10):2241–2277, 2020. doi:10.1080/02331934.2019.1679811.

[50] P. Mehlitz. Stationarity conditions and constraint qualifications for mathematical programs with switching constraints. *Mathematical Programming*, 181(1):149–186, 2020. doi:10.1007/s10107-019-01380-5.

[51] P. Mehlitz and G. Wachsmuth. The limiting normal cone to pointwise defined sets in Lebesgue spaces. *Set-Valued and Variational Analysis*, 26(3):449–467, 2018. doi:10.1007/s11228-016-0393-4.

[52] B. S. Mordukhovich. *Variational Analysis and Applications*. Springer, Cham, 2018. doi:10.1007/978-3-319-92775-6.

[53] J. V. Outrata, M. Kočvara, and J. Zowe. *Nonsmooth Approach to Optimization Problems with Equilibrium Constraints*. Kluwer Academic, Dordrecht, 1998. doi:10.1007/978-1-4757-2825-5.

[54] A. Ramos. Mathematical programs with equilibrium constraints: a sequential optimality condition, new constraint qualifications and algorithmic consequences. *Optimization Methods and Software*, 36(1):45–81, 2021. doi:10.1080/10556788.2019.1702661.

[55] M. Raydan. The Barzilai and Borwein gradient method for the large scale unconstrained minimization problem. *SIAM Journal on Optimization*, 7(1):26–33, 1997. doi:10.1137/s1052623494266365.

[56] S. M. Robinson. Some continuity properties of polyhedral multifunctions. In H. König, B. Korte, and K. Ritter, editors, *Mathematical Programming at Oberwolfach*, pages 206–214. Springer, Berlin, 1981. doi:10.1007/bfb0120929.

[57] R. T. Rockafellar and R. J.-B. Wets. *Variational Analysis*, volume 317. Springer Science & Business Media, Berlin, 2009. doi:10.1007/978-3-642-02431-3.

[58] S. Scholtes. Convergence properties of a regularization scheme for mathematical programs with complementarity constraints. *SIAM Journal on Optimization*, 11(4):918–936, 2001. doi:10.1137/S1052623499361233.

[59] S. J. Wright, R. D. Nowak, and M. A. T. Figueiredo. Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing*, 57(7):2479–2493, 2009. doi:10.1109/tsp.2009.2016892.

[60] M. Xu and J. J. Ye. Relaxed constant positive linear dependence constraint qualification and its application to bilevel programs. *Journal of Global Optimization*, 78:181–205, 2020. doi:10.1007/s10898-020-00907-x.