J. Nonsmooth Anal. Optim. 5 (2024), 12235, doi: 10.46298/jnsao-2024-12235
Submitted: 2023-09-06, accepted: 2024-05-14

LOCAL PROPERTIES AND AUGMENTED LAGRANGIANS IN FULLY NONCONVEX COMPOSITE OPTIMIZATION

Alberto De Marchi* Patrick Mehlitz†

"*In fact the great watershed in optimization isn't between linearity and nonlinearity, but between convexity and nonconvexity.*"

— R. T. Rockafellar [51]

Abstract A broad class of optimization problems can be cast in composite form, that is, considering the minimization of the composition of a lower semicontinuous function with a differentiable mapping. This paper investigates the versatile template of composite optimization without any convexity assumptions. First- and second-order optimality conditions are discussed. We highlight the difficulties that stem from the lack of convexity when dealing with necessary conditions in a Lagrangian framework and when considering error bounds. Building upon these characterizations, a local convergence analysis is delineated for a recently developed augmented Lagrangian method, deriving rates of convergence in the fully nonconvex setting.

Keywords Augmented Lagrangian framework · Composite nonconvex optimization · Error bounds · Local convergence properties · Second-order variational analysis

MSC (2020) 49J52 · 49J53 · 65K10 · 90C30 · 90C33

1 INTRODUCTION

In this paper, we are concerned with finite-dimensional optimization problems of the form

$$\text{(P)} \qquad \underset{x}{\text{minimize}} \quad \varphi(x) := f(x) + g(c(x)),$$

where $x \in \mathbb{R}^n$ is the decision variable, $f\colon \mathbb{R}^n \to \mathbb{R}$ and $c\colon \mathbb{R}^n \to \mathbb{R}^m$ are smooth mappings, and $g\colon \mathbb{R}^m \to \overline{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$ is merely proper and lower semicontinuous. The data functions f and g are allowed to be nonconvex mappings, the nonsmooth cost g is not necessarily continuous nor real-valued, and the mapping c is potentially nonlinear. Such setting of nonsmooth nonconvex composite optimization was named *generalized* [54] (or *extended* [52]) nonlinear programming by Rockafellar, and it is well known that the model problem (P) covers numerous applications from signal processing, sparse optimization, compressed sensing, machine learning, and disjunctive programming. Let us mention that the situation where g is *convex* has been intensively studied in the literature, see e.g. [52, 53, 54]. We do *not* make such an assumption and consider the far more challenging general setup.

Our motivation behind the potential nonconvexity of g is driven by applications from sparse or low-rank optimization. Although the (convex) ℓ_1- and nuclear norm are known to promote sparse or low-rank behavior, solutions are often not sparse *enough* in certain settings. In order to overcome this

*University of the Bundeswehr Munich, Department of Aerospace Engineering, Institute of Applied Mathematics and Scientific Computing, 85577 Neubiberg, Germany. alberto.demarchi@unibw.de, orcid: 0000-0002-3545-6898.

†Philipps-Universität Marburg, Department of Mathematics and Computer Science, 35032 Marburg, Germany. mehlitz@uni-marburg.de, orcid: 0000-0002-9355-850X.

issue, one can rely on the ℓ_0-quasi-norm or the matrix rank as "regularizers", which are discontinuous functions. Intermediate choices for g like the ℓ_q-quasi-norm or the q-Schatten-quasi-norm for $q \in (0, 1)$ are nonconvex but uniformly continuous, and have turned out to work well in numerical practice. Another driving force behind this work comes from disjunctive programming, in particular from the observation that constraints can be naturally formulated in a function-in-set format whereby sets are nonconvex yet simple (to project onto). The template (P) lends itself to capture this scenario, taking full advantage of g as the indicator of a nonconvex set, comprising structures typical for, e.g., complementarity, switching, and vanishing constraints, see the classical monographs [40, 46] and the more recent contributions [23, 41].

The possibility to include constraints in the model problem (P) becomes apparent with a direct reformulation. Introducing an auxiliary variable $z \in \mathbb{R}^m$, (P) can be equivalently rewritten in the form

$$(\mathrm{P_S}) \qquad \underset{x,z}{\text{minimize}} \quad f(x) + g(z) \qquad \text{subject to} \quad c(x) - z = 0,$$

which has a separable objective function, without nontrivial compositions, and explicitly includes some equality constraints. An analogous template has been studied in [17], demonstrating its modeling versatility, mostly enabled by accepting potentially nonconvex g.

A fundamental technique for solving constrained optimization problems is the augmented Lagrangian (AL) framework, which can effortlessly handle nonsmooth objectives, see e.g. [12, 14, 16, 17, 20, 28, 29, 53, 54, 56] for some recent contributions, and [10, 11, 15] for some fundamental literature which addresses the setting of standard nonlinear programming. Particularly, Rockafellar extended the approach in [53, 54] to the broad setting of (P) with g convex, relying on some local duality to build a connection with the proximal point algorithm (PPA), see [49, 50]. Embracing the fully nonconvex setting, we are interested here in investigating AL methods for generalized nonlinear programming without any convexity assumption. Although the shifted-penalty approach underpinning the seminal "method of multipliers" still applies in our setting, it appears more difficult to leverage the perspective of PPA. Moreover, the nonconvexity of g leads to a lack of regularity, as its proximal mapping is potentially set-valued. Here, we seek a better understanding of the variational properties of (P) and the convergence guarantees of AL methods for this class of problems. Building upon the global characterization in [16], we will focus on second-order optimality conditions and local analysis, portraying a convergence theory for the fully nonconvex setting including rates-of-convergence results.

The following blanket assumptions are considered throughout, without further mention. Technical definitions are given in Section 2.

Assumption 1.1. The following hold in (P):

(i) $f\colon \mathbb{R}^n \to \mathbb{R}$ and $c\colon \mathbb{R}^n \to \mathbb{R}^m$ are twice continuously differentiable;

(ii) $g\colon \mathbb{R}^m \to \overline{\mathbb{R}}$ is proper, lower semicontinuous, and prox-bounded;

(iii) $\inf \varphi \in \mathbb{R}$.

The prox-boundedness assumption on g in Assumption 1.1(ii) is included to ensure that, for some suitable parameters, the proximal mapping of g is well-defined, and so is the overall numerical scheme considered in this work. However, such stipulation is not necessary. As suggested in [54], a "trust region" can be specified to localize the proximal minimization and to support its attainment. While avoiding the artificial unboundedness potentially introduced by relaxing the composition constraint, this localization would affect some algorithmic and global aspects, but not the local behavior and properties we are interested in here.

1.1 RELATED WORK AND CONTRIBUTIONS

Since its inception [30, 47, 49], the AL framework has been extensively investigated and developed [11, 15], also in infinite dimensions [38]. It was soon recognized that, in the convex setting, the method of multipliers can be associated to the PPA applied to a dual problem, see [50]. Following this pattern, local convexity enabled by some second-order optimality conditions allowed to reconcile the AL scheme with an application of the PPA, and thereby establishing convergence, beyond the convex setting [10, 53, 54]. However, when it comes to *local* convergence properties, available results remain confined to the case with g convex.

Contributions With this work, we extend recent results by Rockafellar from [53, 54], where composite optimization problems with convex g are considered, to the more general setting. In particular, we study the implicit AL method from [16] and characterize its local convergence behavior. Particularly, under suitable conditions, we show convergence of the full sequence with linear or superlinear rate in Theorem 4.12. To proceed, we make use of problem-tailored second-order conditions which have been developed recently in [8]. Moreover, the Lagrange multiplier has to be locally unique, see Lemma 3.13.

Sparsity-promoting terms and nonconvex constraint sets have turned out to work well in the AL framework—at least from a global perspective, e.g. in [17, 35]. We are also interested in local properties now, with a focus on the numerical method proposed in [16], which favorably avoids the use of slack variables, see [6] for a recent study.

Local fast convergence of an AL method in composite optimization has been considered from the viewpoint of variational analysis in the recent paper [28] in the context where g is a continuous, piecewise quadratic, convex function. This allows for a unified analysis as the standard second-order sufficient condition already gives the necessary error bound condition (due to the result from Lemma 2.7 and the analysis in [42, 55]). A recent analysis of the local convergence of AL methods for (P) is that in [29], restricted to convex g, which considers second-order sufficient conditions and establishes Q-linear convergence of the primal-dual sequence, without assuming any constraint qualification (CQ).

Our analysis also took inspiration from [13, 36, 57] where, among other things, the local analysis of AL methods for (smooth) optimization problems with geometric constraints of type $c(x) \in K$ for some closed, convex set K is considered in a Banach space setting. We, at least roughly, follow the arguments in [57] and (apart from the fact that we are working in a fully finite-dimensional setting) generalize the findings therein to nonsmooth composite problems.

Let us point out that desirable local convergence properties of AL methods in standard nonlinear programming can be guaranteed with no more than a second-order sufficient condition, i.e., no additional CQ is necessary, see [22], and the second-order sufficient condition can even be replaced by a weaker noncriticality assumption on the involved multipliers, as shown later in [33]. One reason for this behavior is the inherent (convex) polyhedrality of the involved constraint set, see Example 3.4 below, which also gives (convex) polyhedrality of the associated set of Lagrange multipliers. The fact that polyhedrality comes along with certain stability properties (in the sense of error bounds) is well known from the seminal papers [48, 59]. In the general, nonpolyhedral situation, such a result is not likely to hold, see [36], and an additional CQ might be necessary. Exemplary, this has been visualized in the papers [27, 37] where AL methods in second-order cone programming have been investigated. In order to obtain convergence rates, the authors do not only postulate the validity of a second-order condition, but make use of an addition assumption. In [37], the authors exploit the strict Robinson condition (which guarantees uniqueness of the underlying Lagrange multiplier) while in [27], a certain multiplier mapping is assumed to be calm while, at the point of interest, uniqueness of the Lagrange multiplier is also needed.

Roadmap The remainder of the paper is organized as follows. Section 2 provides some preliminary results from variational analysis and generalized differentiation. Section 3 is dedicated to the investigation of first-order necessary and second-order sufficient optimality conditions in nonconvex composite optimization. Furthermore, we comment on a reasonable choice for an AL function and investigate error bounds for a system of necessary optimality conditions associated with (P). In Section 4, we first introduce the AL method of our interest before presenting some global convergence results which complement the analysis provided in [16]. Then, local convergence results are presented. We start by clarifying the existence and convergence of minimizers for the associated AL subproblems before focusing on the derivation of rates-of-convergence results. Section 5 illustrates our findings by means of two exemplary settings: sparsity-promoting nonlinear optimization and complementarity-constrained optimization. The paper closes with some concluding remarks in Section 6.

2 PRELIMINARIES

This section provides notation, preliminaries, and known facts based on [43, 55], with some additional basic results.

With $\mathbb{R}$ and $\overline{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$ we indicate the real and extended-real line, respectively. The set of natural numbers is denoted by $\mathbb{N}$. We equip the appearing Euclidean spaces possessing the standard Euclidean inner product $\langle\cdot,\cdot\rangle$ with the associated Euclidean norm $\|\cdot\|$. In product spaces, we make use of the associated maximum norm. With $\mathbb{B}_r(x)$ we indicate the closed ball centered at $x \in \mathbb{R}^n$ with radius $r > 0$. Given a set $A \subseteq \mathbb{R}^n$, we use $x + A := A + x := \{a + x \in \mathbb{R}^n \mid a \in A\}$ for brevity. The notation $\{a^k\}_{k\in K}$ represents a sequence indexed by elements of the set $K \subseteq \mathbb{N}$, and we write $\{a^k\}_{k\in K} \subseteq A$ to indicate that $a^k \in A$ for all indices $k \in K$. Whenever clear from context, we may simply write $\{a^k\}$ to indicate $\{a^k\}_{k\in\mathbb{N}}$. Notation $a^k \to_K x$ ($a^k \to x$) is used to express convergence of $\{a^k\}_{k\in K}$ (of $\{a^k\}$) to x. If $n = 1$, we use $\{a_k\}_{k\in K}$ and $\{a_k\}$ to emphasize that we are dealing with sequences of scalars. We will adopt the *little-o* notation for asymptotics: given sequences $\{a_k\}$ and $\{\varepsilon_k\} \subset (0,\infty)$, we write $a_k \in o(\varepsilon_k)$ to indicate that $\lim_{k\to\infty} |a_k|/\varepsilon_k = 0$.

A function $f\colon \mathbb{R}^n\times\mathbb{R}^m \to \overline{\mathbb{R}}$ with values $f(x,z)$ is *level-bounded in x locally uniformly in z* if for each $\alpha \in \mathbb{R}$ and $\bar z \in \mathbb{R}^m$ there exists $\varepsilon > 0$ such that the set $\{(x,z) \in \mathbb{R}^n \times \mathbb{R}^m \mid f(x,z) \le \alpha, \|z - \bar z\| \le \varepsilon\}$ is bounded. The *effective domain* of a function $h\colon \mathbb{R}^m \to \overline{\mathbb{R}}$ is denoted by $\operatorname{dom} h := \{z \in \mathbb{R}^m \mid h(z) < \infty\}$. The set $\operatorname{epi} h := \{(z,\alpha) \in \mathbb{R}^m \times \mathbb{R} \mid \alpha \ge h(z)\}$ is called the *epigraph* of h. We say that h is *proper* if $\operatorname{dom} h \neq \emptyset$ and *lower semicontinuous* if $h(\bar z) \le \liminf_{z\to\bar z} h(z)$ for all $\bar z \in \mathbb{R}^m$. Note that h is lower semicontinuous if and only if $\operatorname{epi} h$ is closed. Given a point $\bar z \in \operatorname{dom} h$, we may avoid to assume h continuous and instead appeal to *h-attentive* convergence of a sequence $\{z^k\}$, denoted as $z^k \xrightarrow{h} \bar z$ and given by $z^k \to \bar z$ with $h(z^k) \to h(\bar z)$. For some real number $\lambda \ge 1$, we refer to h as *positively homogeneous of degree* λ if $h(\alpha y) = \alpha^\lambda h(y)$ holds for all $y \in \mathbb{R}^m$ and real numbers $\alpha > 0$. The *conjugate function* $h^*\colon \mathbb{R}^m \to \overline{\mathbb{R}}$ associated with (proper and lower semicontinuous) h is defined by

$$h^*(y) := \sup_z \{\langle y, z\rangle - h(z)\}.$$

We note that h^* is a convex function by definition since it is a supremum of affine functions.

For a proper and lower semicontinuous function $h\colon \mathbb{R}^m \to \overline{\mathbb{R}}$, a point $\bar z \in \mathbb{R}^m$ is called *feasible* if $\bar z \in \operatorname{dom} h$. A feasible point $\bar z \in \mathbb{R}^m$ is said to be *locally optimal*, or called a *local minimizer*, if there exists $r > 0$ such that $h(\bar z) \le h(z)$ holds for all feasible $z \in \mathbb{B}_r(\bar z)$. Additionally, if this inequality holds for all feasible $z \in \mathbb{R}^m$, then $\bar z$ is said to be *(globally) optimal*.

We use the notation $\Gamma\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ to indicate a point-to-set function $\Gamma\colon \mathbb{R}^n \to 2^{\mathbb{R}^m}$. The set $\operatorname{gph}\Gamma := \{(x,y) \in \mathbb{R}^n\times\mathbb{R}^m \mid y \in \Gamma(x)\}$ is called the *graph* of Γ. The set-valued mapping $\Gamma^{-1}\colon \mathbb{R}^m \rightrightarrows \mathbb{R}^n$ given by $\operatorname{gph}\Gamma^{-1} := \{(y,x) \in \mathbb{R}^m \times \mathbb{R}^n \mid (x,y) \in \operatorname{gph}\Gamma\}$ is referred to as the *inverse* of Γ. The set

$\ker\Gamma := \{x \in \mathbb{R}^n \mid 0 \in \Gamma(x)\}$ is the *kernel* of Γ. Recall that Γ is said to be a *polyhedral* mapping if $\operatorname{gph}\Gamma$ can be represented as the union of finitely many convex polyhedral sets.

2.1 PROXIMAL MAPPINGS

Let $h\colon \mathbb{R}^m \to \overline{\mathbb{R}}$ be proper and lower semicontinuous. Given a parameter value $\mu > 0$, the *proximal* mapping $\operatorname{prox}_{\mu h}\colon \mathbb{R}^m \rightrightarrows \mathbb{R}^m$ is defined by

$$\operatorname{prox}_{\mu h}(z) := \operatorname*{arg\,min}_{z'} \left\{ h(z') + \frac{1}{2\mu}\|z' - z\|^2 \right\}. \tag{2.1}$$

We say that h is *prox-bounded* if $h + \|\cdot\|^2/(2\mu)$ is bounded below on $\mathbb{R}^m$ for some $\mu > 0$, see [55, Def. 1.23]. The supremum of all such μ is the threshold μ_h of prox-boundedness for h. In particular, if h is bounded below by an affine function, then $\mu_h = \infty$. For any $\mu \in (0, \mu_h)$, the proximal mapping $\operatorname{prox}_{\mu h}$ is locally bounded as well as nonempty- and compact-valued, see [55, Thm 1.25]. The value function of the minimization problem defining the proximal mapping is the *Moreau envelope* with parameter $\mu \in (0, \mu_h)$, denoted $h^\mu\colon \mathbb{R}^m \to \mathbb{R}$, namely

$$h^\mu(z) := \inf_{z'} \left\{ h(z') + \frac{1}{2\mu}\|z' - z\|^2 \right\}.$$

The *projection* mapping $\Pi_\Omega\colon \mathbb{R}^m \rightrightarrows \mathbb{R}^m$ and the *distance* function $\operatorname{dist}_\Omega\colon \mathbb{R}^m \to \mathbb{R}$ of a nonempty set $\Omega \subseteq \mathbb{R}^m$ are defined by

$$\Pi_\Omega(z) := \operatorname*{arg\,min}_{z' \in \Omega} \|z' - z\|, \qquad \operatorname{dist}(z, \Omega) := \inf_{z' \in \Omega} \|z' - z\|.$$

The former is a set-valued mapping whenever Ω is nonconvex, whereas the latter is always single-valued.

The following technical lemmas are used later on.

Lemma 2.1. *Let $h\colon \mathbb{R}^m \to \overline{\mathbb{R}}$ be proper, lower semicontinuous, and prox-bounded. Let $\operatorname{dom} h$ be closed and fix $\bar z \in \mathbb{R}^m$. Then*

$$\lim_{z' \to \bar z,\, \mu \downarrow 0} \inf_{z \in \operatorname{dom} h} \left\{ \mu h(z) + \frac{1}{2}\|z - z'\|^2 \right\} = \frac{1}{2}\operatorname{dist}^2(\bar z, \operatorname{dom} h).$$

Proof. As $\operatorname{dom} h$ is nonempty and closed, we find $\tilde z \in \operatorname{dom} h$ such that $\operatorname{dist}(\bar z, \operatorname{dom} h) = \|\tilde z - \bar z\|$. For every $z' \in \mathbb{R}^m$ and $\mu > 0$, this gives

$$\inf_{z \in \operatorname{dom} h} \left\{ \mu h(z) + \frac{1}{2}\|z - z'\|^2 \right\} \le \mu h(\tilde z) + \frac{1}{2}\|\tilde z - z'\|^2,$$

and taking the upper limit, we find

$$\limsup_{z' \to \bar z,\, \mu \downarrow 0} \inf_{z \in \operatorname{dom} h} \left\{ \mu h(z) + \frac{1}{2}\|z - z'\|^2 \right\} \le \frac{1}{2}\|\tilde z - \bar z\|^2 = \frac{1}{2}\operatorname{dist}^2(\bar z, \operatorname{dom} h). \tag{2.2}$$

Next, we define $\psi\colon \mathbb{R}^m \times [0,\infty) \to \mathbb{R} \cup \{-\infty\}$ by means of

$$\forall z' \in \mathbb{R}^m,\ \forall \mu \in [0,\infty)\colon \quad \psi(z', \mu) := \inf_{z \in \operatorname{dom} h} \left\{ \mu h(z) + \frac{1}{2}\|z - z'\|^2 \right\}.$$

As h is prox-bounded, ψ takes finite values for all sufficiently small μ, and these finite values are attained, see [55, Thm 1.25]. Suppose that there are sequences $\{z^k\}, \{\bar z^k\} \subseteq \mathbb{R}^m$ and $\{\mu_k\} \subseteq [0,\infty)$ with

$z^k \to \bar z$ and $\mu_k \to 0$, $\psi(z^k,\mu_k) = \mu_k h(\bar z^k) + \frac{1}{2}\|\bar z^k - z^k\|^2$, and $\|\bar z^k\| \to \infty$. On the one hand, boundedness of $\{z^k\}$ and $\{\mu_k\}$ gives the existence of a constant $C > 0$ such that

$$\forall k \in \mathbb{N}: \quad \psi(z^k,\mu_k) \le \mu_k h(\tilde z) + \frac{1}{2}\|\tilde z - z^k\|^2 \le C. \tag{2.3}$$

On the other hand, the prox-boundedness of h implies that h is minorized by a quadratic function, see [55, Ex. 1.24]. Hence, there are constants $c_1, c_2, c_3 > 0$ such that, for sufficiently large $k \in \mathbb{N}$,

$$\begin{aligned}\psi(z^k,\mu_k) &\ge -\mu_k c_1\|\bar z^k\|^2 - \mu_k c_2\|\bar z^k\| - \mu_k c_3 + \frac{1}{2}\|\bar z^k - z^k\|^2\\ &\ge \left(\frac{1}{2} - \mu_k c_1\right)\|\bar z^k\|^2 - (\mu_k c_2 + \|z^k\|)\|\bar z^k\| - \mu_k c_3.\end{aligned}$$

Boundedness of $\{z^k\}$ and $\mu_k \to 0$ thus yield $\psi(z^k,\mu_k) \to \infty$ since $\|\bar z^k\| \to \infty$. This, however, is a contradiction to (2.3). Hence, we can choose a compact set $C \subseteq \mathbb{R}^m$ such that, for each $\mu \ge 0$ small enough and each z' sufficiently close to $\bar z$, we have

$$\psi(z',\mu) = \inf_{z \in C\cap \operatorname{dom} h}\left\{\mu h(z) + \frac{1}{2}\|z - z'\|^2\right\}.$$

Thus, due to the lower semicontinuity of h, we can apply [1, Thm 4.2.1(1)] in order to obtain

$$\liminf_{z'\to\bar z,\, \mu\downarrow 0}\ \inf_{z\in\operatorname{dom} h}\left\{\mu h(z) + \frac{1}{2}\|z - z'\|^2\right\} \ge \psi(\bar z, 0) = \frac{1}{2}\operatorname{dist}^2(\bar z, \operatorname{dom} h).$$

Together with (2.2), the assertion follows. □

Lemma 2.2. *Let $h\colon \mathbb{R}^m \to \overline{\mathbb{R}}$ be proper, lower semicontinuous, and prox-bounded. Fix $\bar z \in \mathbb{R}^m$ as well as sequences $\{\mu_k\} \subseteq (0,\infty)$, $\{z^k\} \subseteq \operatorname{dom} h$, and $\{\tilde z^k\} \subseteq \mathbb{R}^m$ such that $\mu_k \downarrow 0$, $\{\tilde z^k\}$ is bounded, and*

$$\limsup_{k\to\infty}\left(\mu_k h(z^k) + \frac{1}{2}\|z^k - \tilde z^k\|^2\right) \le 0. \tag{2.4}$$

Then $\mu_k h(z^k) \to 0$ and $\|z^k - \tilde z^k\| \to 0$.

Proof. As in the proof of Lemma 2.1, we use [55, Ex. 1.24] to find constants $c_1, c_2, c_3 > 0$ such that $h(z) \ge -c_1\|z\|^2 - c_2\|z\| - c_3$ holds for all $z \in \mathbb{R}^m$. Thus, we have

$$\limsup_{k\to\infty}\left(\left(\frac{1}{2} - \mu_k c_1\right)\|z^k\|^2 - (\mu_k c_2 + \|\tilde z^k\|)\|z^k\| - \mu_k c_3\right) \le 0$$

from (2.4). By $\mu_k \downarrow 0$ and boundedness of $\{\tilde z^k\}$, this implies that $\{z^k\}$ is bounded as well. Hence, $\{h(z^k)\}$ is bounded below, which gives $\liminf_{k\to\infty} \mu_k h(z^k) \ge 0$. Now, (2.4) yields the claim. □

2.2 VARIATIONAL ANALYSIS AND GENERALIZED DIFFERENTIATION

TANGENT AND NORMAL CONES

We start by repeating the definition of some cones which are well known in variational analysis. Therefore, we fix some closed set $\Omega \subseteq \mathbb{R}^m$ and $\bar z \in \Omega$. We refer to

$$T_\Omega(\bar z) := \left\{v \in \mathbb{R}^m \,\middle|\, \exists\{t_k\} \subset (0,\infty),\ \exists\{v^k\} \subseteq \mathbb{R}^m\colon\ t_k \downarrow 0,\ v^k \to v,\ \bar z + t_k v^k \in \Omega\ \forall k \in \mathbb{N}\right\}$$

as the *tangent cone* to Ω at $\bar z$, and we point out that it is always a closed cone. Furthermore, we make use of

$$\begin{aligned}
\widehat{N}_\Omega(\bar z) &:= \{v \in \mathbb{R}^m \,|\, \forall z \in \Omega\colon\ \langle v, z - \bar z\rangle \le o(\|z - \bar z\|)\},\\
N_\Omega(\bar z) &:= \{v \in \mathbb{R}^m \,|\, \exists \{z^k\} \subseteq \Omega,\ \exists \{v^k\} \subseteq \mathbb{R}^m\colon\ z^k \to \bar z,\ v^k \to v,\ v^k \in \widehat{N}_\Omega(z^k)\ \forall k \in \mathbb{N}\}
\end{aligned}$$

which are called *regular* (or Fréchet) and *limiting* (or Mordukhovich) *normal cone* to Ω at $\bar z$. Both of these cones are closed, and $\widehat{N}_\Omega(\bar z)$ is, additionally, convex. For a convex set Ω, we have

$$\widehat{N}_\Omega(\bar z) = N_\Omega(\bar z) = \{v \in \mathbb{R}^m \,|\, \forall z \in \Omega\colon\ \langle v, z - \bar z\rangle \le 0\}.$$

We would like to point out the polar relation

$$\widehat{N}_\Omega(\bar z) = T_\Omega(\bar z)^\circ, \tag{2.5}$$

Here, we made use of $A^\circ := \{v \in \mathbb{R}^m \,|\, \forall z \in A\colon\ \langle v, z\rangle \le 0\}$, the *polar cone* of $A \subseteq \mathbb{R}^m$.

SUBDIFFERENTIALS AND STATIONARITY

For a lower semicontinuous function $h\colon \mathbb{R}^m \to \overline{\mathbb{R}}$ and $\bar z \in \operatorname{dom} h$,

$$\begin{aligned}
\widehat{\partial} h(\bar z) &:= \{v \in \mathbb{R}^m \,|\, (v, -1) \in \widehat{N}_{\operatorname{epi} h}(\bar z, h(\bar z))\},\\
\partial h(\bar z) &:= \{v \in \mathbb{R}^m \,|\, (v, -1) \in N_{\operatorname{epi} h}(\bar z, h(\bar z))\},\\
\partial^\infty h(\bar z) &:= \{v \in \mathbb{R}^m \,|\, (v, 0) \in N_{\operatorname{epi} h}(\bar z, h(\bar z))\}
\end{aligned}$$

are referred to as the the *regular* (or Fréchet), *limiting* (or Mordukhovich), and *singular* (or horizon) *subdifferential* of h at $\bar z$. Whenever h is Lipschitz continuous around $\bar z$, then $\partial^\infty h(\bar z) = \{0\}$. Let us mention that, among others, the subdifferential operators $\widehat{\partial}$, ∂, and ∂^∞ are compatible with respect to smooth additions. Indeed, for each continuously differentiable function $h_0\colon \mathbb{R}^m \to \mathbb{R}$, it holds

$$\widehat{\partial}(h_0 + h)(\bar z) = \nabla h_0(\bar z) + \widehat{\partial} h(\bar z), \quad \partial(h_0 + h)(\bar z) = \nabla h_0(\bar z) + \partial h(\bar z), \quad \partial^\infty(h_0 + h)(\bar z) = \partial^\infty h(\bar z).$$

Whenever $h := \delta_\Omega$, where δ_Ω is the (proper and lower semicontinuous) *indicator function* of the nonempty, closed set $\Omega \subseteq \mathbb{R}^m$, vanishing on Ω and being ∞ otherwise, we have $\operatorname{dom} \delta_\Omega = \Omega$, and

$$\widehat{\partial} \delta_\Omega(\bar z) = \widehat{N}_\Omega(\bar z), \quad \partial \delta_\Omega(\bar z) = \partial^\infty \delta_\Omega(\bar z) = N_\Omega(\bar z)$$

for $\bar z \in \Omega$. The proximal mapping of δ_Ω is the projection Π_Ω, so that Π_Ω is locally bounded.

Lemma 2.3. *Let $h\colon \mathbb{R}^m \to \overline{\mathbb{R}}$ be proper, lower semicontinuous, and positively homogeneous of degree 2. Then, for each $\bar z \in \operatorname{dom} h$ and $v \in \partial h(\bar z)$, $h(\bar z) = \frac{1}{2}\langle v, \bar z\rangle$ is valid.*

Proof. Let us note that the assertion is trivially true whenever $\bar z = 0$ holds, so let us assume that $\bar z \neq 0$. First, suppose that $v \in \widehat{\partial} h(\bar z)$. Due to [43, Thm 1.26], this yields

$$\liminf_{z \to \bar z,\, z \neq \bar z} \frac{h(z) - h(\bar z) - \langle v, z - \bar z\rangle}{\|z - \bar z\|} \ge 0.$$

Considering $z := (1 \pm t)\bar z$ for $t \downarrow 0$ in the above limit, and exploiting positive homogeneity of degree 2 of h, we find $(\pm 2h(\bar z) \mp \langle v, \bar z\rangle)/\|\bar z\| \ge 0$ which yields $h(\bar z) = \frac{1}{2}\langle v, \bar z\rangle$. To obtain the lemma's assertion in the more general case where $v \in \partial h(\bar z)$, we combine the above findings with [43, Thm 1.28]. ◻

A point $\bar{z} \in \operatorname{dom} h$ is said to be *M-stationary* whenever $0 \in \partial h(\bar{z})$ is valid, and this constitutes a necessary condition for the local minimality of $\bar{z}$ for h also known as *Fermat's rule*, see [55, Thm 10.1]. It should be noted that $0 \in \widehat{\partial} h(\bar{z})$ serves as a (potentially sharper) necessary optimality condition as well. Given some tolerance $\varepsilon \geq 0$, an approximate M-stationarity concept for the minimization of h refers to $\operatorname{dist}(0, \partial h(\bar{z})) \leq \varepsilon$ which we refer to as *ε-M-stationarity*. By closedness of $\partial h(\bar{z})$, ε-M-stationarity with $\varepsilon = 0$ recovers the notion of M-stationarity.

Below, we introduce a stationarity concept that will be use later to qualify the iterates of our implicit AL algorithm, see [16, Sec. 4.2]. Therefore, let us consider a parametric optimization problem with an objective $p\colon \mathbb{R}^n \to \overline{\mathbb{R}}$ and an oracle $\mathbf{O}\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ given by

$$p(x) := \inf_{z \in \mathbb{R}^m} P(x, z), \qquad \mathbf{O}(x) := \operatorname*{arg\,min}_{z \in \mathbb{R}^m} P(x, z) \tag{2.6}$$

for a proper, lower semicontinuous function $P\colon \mathbb{R}^n \times \mathbb{R}^m \to \overline{\mathbb{R}}$. Recalling that the notion of uniform level-boundedness corresponds to a parametric extension of level-boundedness, see [55, Def. 1.16], we suppose that P is level-bounded in z (second argument) locally uniformly in x (first argument). Then, from [55, Thm 10.13] we have for every $\bar{x} \in \operatorname{dom} p$ the inclusion

$$\partial p(\bar{x}) \subseteq \Upsilon(\bar{x}) := \bigcup_{\bar{z} \in \mathbf{O}(\bar{x})} \{\xi \in \mathbb{R}^n \mid (\xi, 0) \in \partial P(\bar{x}, \bar{z})\}. \tag{2.7}$$

In the setting (2.6), because of the parametric nature of p, the subdifferential mapping $\partial p\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is not a simple object in general, making M-stationarity difficult to check. Therefore, for the minimization of p, one can resort to the concept of Υ-*stationarity*, coined in [16, Def. 4.1].

Definition 2.4 (Υ-stationarity). Let $\varepsilon \geq 0$ be fixed and let $P\colon \mathbb{R}^n \times \mathbb{R}^m \to \overline{\mathbb{R}}$ be chosen as specified above. Define $p\colon \mathbb{R}^n \to \overline{\mathbb{R}}$ and $\Upsilon\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ as in (2.6) and (2.7), respectively. Then, relatively to the minimization of p, a point $\bar{x} \in \operatorname{dom} p$ is called *ε-Υ-stationary* if $\operatorname{dist}(0, \Upsilon(\bar{x})) \leq \varepsilon$. In the case $\varepsilon = 0$, such a point $\bar{x}$ is said to be *Υ-stationary*.

Notice that the inclusion $\bar{x} \in \operatorname{dom} p$ is implicitly required to have the set $\Upsilon(\bar{x})$ nonempty. Furthermore, in the exact case, namely $\varepsilon = 0$, Υ-stationarity of $\bar{x}$ coincides with $0 \in \Upsilon(\bar{x})$, by closedness of $\Upsilon(\bar{x})$ [55, Thm 10.13]. We shall point out that Υ-stationarity provides an intermediate qualification between (the stronger) M-stationarity for p and (the weaker) M-stationarity for P, see [16, Prop. 4.1 and 4.2] for details. Finally, it appears from (2.7) that an ε-Υ-stationary point $\bar{x} \in \mathbb{R}^n$ can be associated to a (possibly nonunique) *certificate* $\bar{z} \in \mathbf{O}(\bar{x})$ that satisfies

$$\operatorname{dist}(0, \Upsilon(\bar{x})) \leq \min_{\xi \in \mathbb{R}^n} \{\|\xi\| \mid (\xi, 0) \in \partial P(\bar{x}, \bar{z})\} \leq \varepsilon.$$

Given such upper bound, the *pair* $(\bar{x}, \bar{z})$ certificates the ε-Υ-stationarity of $\bar{x}$ for p.

Generalized derivatives of set-valued mapping

Let us fix a set-valued mapping $\Gamma\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ and some point $(\bar{x}, \bar{z}) \in \operatorname{gph} \Gamma$. We refer to the set-valued mappings $D\Gamma(\bar{x}, \bar{z})\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ and $D^*\Gamma(\bar{x}, \bar{z})\colon \mathbb{R}^m \rightrightarrows \mathbb{R}^n$, given by

$$\begin{aligned} D\Gamma(\bar{x}, \bar{z})(u) &:= \{v \in \mathbb{R}^m \mid (u, v) \in T_{\operatorname{gph} \Gamma}(\bar{x}, \bar{z})\}, \\ D^*\Gamma(\bar{x}, \bar{z})(y^*) &:= \{x^* \in \mathbb{R}^n \mid (x^*, -y^*) \in N_{\operatorname{gph} \Gamma}(\bar{x}, \bar{z})\}, \end{aligned}$$

as the *graphical derivative* and the *limiting coderivative* of Γ at $(\bar{x}, \bar{z})$. Subsequently, we review some stability properties of set-valued mappings, see e.g. [7, 55]. We say that Γ is *metrically regular* at $(\bar{x}, \bar{z})$ whenever there are neighborhoods $U \subseteq \mathbb{R}^n$ of $\bar{x}$ and $V \subseteq \mathbb{R}^m$ of $\bar{z}$ as well as a constant $\kappa > 0$ such that

$$\forall x \in U,\ \forall z \in V\colon \qquad \operatorname{dist}(x, \Gamma^{-1}(z)) \leq \kappa \operatorname{dist}(z, \Gamma(x)).$$

If just

$$\forall x \in U: \quad \operatorname{dist}(x, \Gamma^{-1}(\bar z)) \le \kappa \operatorname{dist}(\bar z, \Gamma(x))$$

holds, i.e., if $z := \bar z$ can be fixed in the estimate required for metric regularity, then Γ is called *metrically subregular* at $(\bar x, \bar z)$. Furthermore, Γ is said to be *strongly metrically subregular* at $(\bar x, \bar z)$, whenever there exist a neighborhood $U \subseteq \mathbb{R}^n$ of $\bar x$ and a constant $\kappa > 0$ such that

$$\forall x \in U: \quad \|x - \bar x\| \le \kappa \operatorname{dist}(\bar z, \Gamma(x))$$

is valid. Recall that strong metric subregularity of Γ at $(\bar x, \bar z)$ is equivalent to $\ker D\Gamma(\bar x, \bar z) = \{0\}$ by the so-called *Levy–Rockafellar criterion*, see [21, Thm 4E.1] and [39, Prop. 4.1]. Furthermore, Γ is metrically regular at $(\bar x, \bar z)$ if and only if $\ker D^*\Gamma(\bar x, \bar z) = \{0\}$ by the so-called *Mordukhovich criterion*, see [55, Thm 9.40].

Let $F\colon \mathbb{R}^n \to \mathbb{R}^m$ be continuously differentiable and let $\Omega \subseteq \mathbb{R}^m$ be closed. We consider the so-called *feasibility mapping* $\Phi\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ given by

$$\Phi(x) := F(x) - \Omega. \tag{2.8}$$

We fix some point $\bar x \in \mathbb{R}^n$ satisfying $F(\bar x) \in \Omega$, i.e., $(\bar x, 0) \in \operatorname{gph} \Phi$. It is well known that Φ is metrically regular at $(\bar x, 0)$ if and only if

$$N_\Omega(F(\bar x)) \cap \ker F'(\bar x)^\top = \{0\} \tag{2.9}$$

is valid, as we have

$$D^*\Phi(\bar x, 0)(y) = \begin{cases} \{F'(\bar x)^\top y\} & y \in N_\Omega(F(\bar x)), \\ \emptyset & \text{otherwise} \end{cases}$$

from the change-of-coordinates formula in [55, Ex. 6.7] and the representation

$$\operatorname{gph} \Phi = \{(x, w) \in \mathbb{R}^n \times \mathbb{R}^m \mid F(x) - w \in \Omega\}. \tag{2.10}$$

The following lemma, which is a direct consequence of [21, Thm 4B.1], provides a certain openness-type property of the feasibility mapping from (2.8) around points of its graph where it is metrically regular.

Lemma 2.5. *Fix $(\bar x, 0) \in \operatorname{gph} \Phi$ where Φ, the mapping given in (2.8), is metrically regular. Then there exist $s > 0$ and $\varepsilon > 0$ such that*

$$\mathbb{B}_s(0) \subseteq F'(x)\, \mathbb{B}_1(0) - \big(T_\Omega(z) \cap \mathbb{B}_1(0)\big)$$

holds true for all $x \in \mathbb{B}_\varepsilon(\bar x)$ and all $z \in \Omega \cap \mathbb{B}_\varepsilon(F(\bar x))$.

Proof. Since Φ is assumed to be metrically regular at $(\bar x, 0)$, [21, Thm 4B.1] yields the existence of constants $\delta > 0$ and $r > 0$ such that

$$\forall (x, w) \in \operatorname{gph} \Phi \cap \mathbb{B}_\delta(\bar x, 0),\ \forall v \in \mathbb{B}_1(0),\ \exists u \in \mathbb{B}_r(0): \quad v \in D\Phi(x, w)(u).$$

We apply [55, Ex. 6.7] once more to the representation (2.10) in order to find

$$D\Phi(x, w)(u) = F'(x)u - T_\Omega(F(x) - w)$$

for $(x, w) \in \operatorname{gph} \Phi$. Hence, we have

$$\forall (x, w) \in \operatorname{gph} \Phi \cap \mathbb{B}_\delta(\bar x, 0): \quad \mathbb{B}_1(0) \subseteq F'(x)\, \mathbb{B}_r(0) - T_\Omega(F(x) - w).$$

By continuous differentiablity of F, there is a constant $C > 0$ such that $\|F'(x)\| \le C$ holds for all $x \in \mathbb{B}_\delta(\bar x)$. With the aid of $\kappa := \max(r, 1 + Cr)$, this yields

$$\forall (x, w) \in \operatorname{gph} \Phi \cap \mathbb{B}_\delta(\bar x, 0): \quad \mathbb{B}_1(0) \subseteq F'(x)\, \mathbb{B}_\kappa(0) - \big(T_\Omega(F(x) - w) \cap \mathbb{B}_\kappa(0)\big).$$

Let us now choose $\varepsilon \in (0, \delta/2)$ so small such that $\|F(x) - F(\bar{x})\| \leq \delta/2$ holds for all $x \in \mathbb{B}_\varepsilon(\bar{x})$. Then, for arbitrary $x \in \mathbb{B}_\varepsilon(\bar{x})$ and $z \in \Omega \cap \mathbb{B}_\varepsilon(F(\bar{x}))$, we can set $w := F(x) - z$ in order to find $(x, w) \in \operatorname{gph} \Phi \cap \mathbb{B}_\delta(\bar{x}, 0)$, and the above guarantees

$$\mathbb{B}_{1/\kappa}(0) \subseteq F'(x)\,\mathbb{B}_1(0) - \big(T_\Omega(z) \cap \mathbb{B}_1(0)\big).$$

Choosing $s := 1/\kappa$, the assertion follows. □

SUBDERIVATIVES

Let us fix a lower semicontinuous function $h\colon \mathbb{R}^m \to \overline{\mathbb{R}}$. For $\bar{z} \in \operatorname{dom} h$ and $v \in \mathbb{R}^m$, the lower limit

$$\mathrm{d}h(\bar{z})(v) := \liminf_{t \downarrow 0,\, v' \to v} \frac{h(\bar{z} + tv') - h(\bar{z})}{t}$$

is called the *subderivative* of h at $\bar{z}$ in direction v, and the mapping $v \mapsto \mathrm{d}h(\bar{z})(v)$, which, by definition, is lower semicontinuous and positively homogeneous, is referred to as the subderivative of h at $\bar{z}$. We note that $\operatorname{epi} \mathrm{d}h(\bar{z}) = T_{\operatorname{epi} h}(\bar{z}, h(\bar{z}))$, see [55, Thm 8.2(a)]. Furthermore, for $\bar{y} \in \mathbb{R}^m$,

$$\mathrm{d}^2 h(\bar{z}, \bar{y})(v) := \liminf_{t \downarrow 0,\, v' \to v} \frac{h(\bar{z} + tv') - h(\bar{z}) - t\langle \bar{y}, v' \rangle}{\frac{1}{2}t^2} \tag{2.11}$$

is called the *second subderivative* of h at $\bar{z}$ for $\bar{y}$ in direction v. The mapping $v \mapsto \mathrm{d}^2 h(\bar{z}, \bar{y})(v)$, which, by definition, is lower semicontinuous and positively homogeneous of degree 2, is referred to as the second subderivative of h at $\bar{z}$ for $\bar{y}$. The recent study [8] presents an overview of calculus rules addressing these variational tools.

Lemma 2.6. *Let $h\colon \mathbb{R}^m \to \overline{\mathbb{R}}$ be a lower semicontinuous function, and fix $\bar{z} \in \operatorname{dom} h$ and $\bar{y} \in \mathbb{R}^m$. Then we have $\mathrm{d}^2 h(\bar{z}, \bar{y})(0) \in \{-\infty, 0\}$.*

Proof. Observe that $\mathrm{d}^2 h(\bar{z}, \bar{y})(0) \leq 0$ holds by definition of the second subderivative simply by choosing $v' := 0$ in (2.11). Positive homogeneity of degree 2 of the second subderivative guarantees validity of $\mathrm{d}^2 h(\bar{z}, \bar{y})(0) = \alpha^2 \mathrm{d}^2 h(\bar{z}, \bar{y})(0)$ for each $\alpha > 0$, and this is only possible if $\mathrm{d}^2 h(\bar{z}, \bar{y})(0) \in \{-\infty, 0\}$. □

For brevity of presentation, we do not formally introduce the notions of prox-regularity, subdifferential continuity, and twice epi-differentiability, which will be used in the next lemma. Instead, as the precise meaning of these concepts is not exploited in this paper, we refer the interested reader to [55, Def. 13.27, 13.28, and 13.6(b)] for proper definitions.

Lemma 2.7. *Let $h\colon \mathbb{R}^m \to \overline{\mathbb{R}}$ be a lower semicontinuous function, and fix $\bar{z} \in \operatorname{dom} h$ and $\bar{y} \in \partial h(\bar{z})$. Assume that h is prox-regular, subdifferentially continuous, and twice epi-differentiable at $\bar{z}$ for $\bar{y}$. Then we have*

$$\forall v \in \mathbb{R}^m\colon \quad D(\partial h)(\bar{z}, \bar{y})(v) = \frac{1}{2} \partial \mathrm{d}^2 h(\bar{z}, \bar{y})(v),$$

and

$$\forall v, w \in \mathbb{R}^m\colon \quad w \in D(\partial h)(\bar{z}, \bar{y})(v) \implies \mathrm{d}^2 h(\bar{z}, \bar{y})(v) = \langle w, v \rangle.$$

Proof. Recalling that $\mathrm{d}^2 h(\bar{z}, \bar{y})$ is positively homogeneous of degree 2, the second property follows with the aid of Lemma 2.3 from the first one, which is taken from [55, Thm 13.40]. □

3 FUNDAMENTALS OF COMPOSITE OPTIMIZATION

We now move our attention to (P) and discuss relevant optimality and stationarity notions. Furthermore, we investigate local characterizations using second-order tools, regularity concepts, and error bounds.

3.1 Stationarity concepts and Lagrangian-type functions

Before dealing with optimality conditions, we consider some Lagrangian terminology and notions useful for first-order analysis. Including an auxiliary variable $z \in \mathbb{R}^m$, we can lift (P) as ($\mathrm{P_S}$) involving merely equality constraints but no (nontrivial) compositions. Introducing a Lagrange multiplier $y \in \mathbb{R}^m$ for the constraints, we define a Lagrangian-type function $\mathcal{L}^{\mathrm{S}}\colon \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^m \to \overline{\mathbb{R}}$ associated with ($\mathrm{P_S}$) by means of

$$\mathcal{L}^{\mathrm{S}}(x, z, y) \coloneqq f(x) + g(z) + \langle y, c(x) - z\rangle. \tag{3.1}$$

Focusing on those terms of $\mathcal{L}^{\mathrm{S}}$ depending on x, we call the function $\mathcal{L}\colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ given by

$$\mathcal{L}(x, y) \coloneqq f(x) + \langle y, c(x)\rangle \tag{3.2}$$

the *Lagrangian* function of (P). Then, acting as a precursor of $\mathcal{L}$, we refer to $\mathcal{L}^{\mathrm{S}}$ as the *pre-Lagrangian* function of (P). These objects are tightly related to the so-called M-stationarity conditions of both problems ($\mathrm{P_S}$) and (P), see Definition 3.1 below. In fact, these first-order optimality conditions can be expressed in Lagrangian form as

$$0 \in \partial_x \mathcal{L}^{\mathrm{S}}(\bar{x}, \bar{z}, \bar{y}), \qquad 0 \in \partial_z \mathcal{L}^{\mathrm{S}}(\bar{x}, \bar{z}, \bar{y}), \qquad 0 \in \partial_y \mathcal{L}^{\mathrm{S}}(\bar{x}, \bar{z}, \bar{y})$$

or more explicitly as

$$0 = \nabla_x \mathcal{L}(\bar{x}, \bar{y}), \qquad \bar{y} \in \partial g(\bar{z}), \qquad 0 = c(\bar{x}) - \bar{z}. \tag{3.3}$$

Equivalently, albeit omitting the auxiliary variable $\bar{z} = c(\bar{x})$, these read

$$0 = \nabla_x \mathcal{L}(\bar{x}, \bar{y}), \tag{3.4a}$$

$$\bar{y} \in \partial g(c(\bar{x})). \tag{3.4b}$$

Notice that (3.4b) implicitly requires the feasibility of $\bar{x}$ for (P), namely $c(\bar{x}) \in \operatorname{dom} g$, for otherwise the subdifferential $\partial g(c(\bar{x}))$ is empty.

Interpreting (P) as an *unconstrained* problem, first-order necessary optimality conditions using the notion of M-stationarity pertain a point $\bar{x} \in \mathbb{R}^n$ such that $0 \in \partial\varphi(\bar{x})$. We now aim to rewrite this condition in terms of initial problem data, i.e., first-order (generalized) derivatives of f, c, and g. Exploiting compatibility of the limiting subdifferential with respect to smooth additions, we find $0 \in \nabla f(x) + \partial(g \circ c)(\bar{x})$. It has been recognized, e.g. in [31, Sec. 3.2], that metric subregularity of the set-valued mapping $\Xi\colon \mathbb{R}^n \times \mathbb{R} \rightrightarrows \mathbb{R}^m \times \mathbb{R}$ given by

$$\Xi(x, \alpha) \coloneqq (c(x), \alpha) - \operatorname{epi} g \tag{3.5}$$

is enough to guarantee that a subdifferential chain rule can be used to approximate the limiting subdifferential of $g \circ c$ from above in terms of the subdifferential of g and the derivative of c. More precisely, if Ξ is metrically subregular at $((\bar{x}, g(c(\bar{x}))), (0, 0))$, then $\partial(g \circ c)(\bar{x}) \subseteq c'(\bar{x})^\top \partial g(c(\bar{x}))$. From a Lagrangian perspective, this gives the existence of some Lagrange multiplier $\bar{y} \in \mathbb{R}^m$ such that the stationarity conditions (3.4) hold. This stationarity characterization resembles, at least in spirit, the so called Karush–Kuhn–Tucker (or KKT) conditions in nonlinear programming, see e.g. [10, 11].

These considerations lead to the following definition, which uses in accordance with (3.4) the Lagrangian function $\mathcal{L}\colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ associated with (P), given by (3.2).

Definition 3.1 (M-stationarity). Relative to (P), a point $\bar{x} \in \mathbb{R}^n$ is called *M-stationary* if there exists a multiplier $\bar{y} \in \mathbb{R}^m$ such that (3.4) holds. Let $\Lambda(\bar{x})$ denote the set of multipliers $\bar{y} \in \mathbb{R}^m$ such that the M-stationarity conditions (3.4) are satisfied by $(\bar{x}, \bar{y})$.

As a reminder of the possible gap highlighted above, where metric subregularity of Ξ is invoked to formulate first-order optimality conditions in Lagrangian terms, the notion given in Definition 3.1 could be referred to as *KKT*-stationarity, as in [16]. For simplicity, we stick to the nomenclature of *M*-stationarity.

Following the nomenclature in [12, 56], $\mathcal{L}^{\mathrm{S}}$ would be referred to as the Lagrangian of ($\mathrm{P_S}$), as a standalone problem, and not only as the pre-Lagrangian in view of (P). In fact, the definition of $\mathcal{L}^{\mathrm{S}}$ in (3.1) complies with the classical concept of Lagrangian function for equality-constrained optimization problems, such as ($\mathrm{P_S}$), and reflects the (nonsmooth, extended real-valued) objective $(x,z) \mapsto f(x)+g(z)$ of ($\mathrm{P_S}$) and its equality constraints $c(x) - z = 0$. However, containing (primal) nonsmooth terms, $\mathcal{L}^{\mathrm{S}}$ is not differentiable. The object $\mathcal{L}$ defined in (3.2) corresponds to the *ordinary* Lagrangian function of (P) as described in [54], and this is consistent with several other papers which exploit the variational analysis approach to composite optimization, see e.g. [5, 8, 27, 28, 42, 44, 45] and, particularly, the setting of standard nonlinear programming, see Example 3.4 below.

Above, we derived the M-stationarity conditions of (P) at some feasible point $\bar{x} \in \mathbb{R}^n$ by using the chain rule for the limiting subdifferential which, in general, requires a qualification condition like metric subregularity of Ξ from (3.5) at $((\bar{x}, g(c(\bar{x}))), (0,0))$. Note that (2.9) reduces to

$$\partial^\infty g(c(\bar{x})) \cap \ker c'(\bar{x})^\top = \{0\} \tag{3.6}$$

when applied to Ξ at the given reference point, and the latter is equivalent to the mapping Ξ being metrically regular at $((\bar{x}, g(c(\bar{x}))), (0,0))$, which also extends to a neighborhood of the point of interest. Thus, (3.6) is sufficient for the subregularity requirement stated earlier. Clearly, (3.6) is valid whenever g is locally Lipschitz continuous at $c(\bar{x})$ or if $c'(\bar{x})$ has full row rank. As we know that $0 \in \partial\varphi(\bar{x})$ provides a necessary optimality condition for the local optimality of $\bar{x}$, the M-stationarity conditions from Definition 3.1 do so as well in the presence of a suitable CQ like (3.6) as outlined above.

Augmented Lagrangian We shall introduce *augmented* Lagrangian functions, which not only offer the basic component for AL methods, but also closely relate to first-order optimality concepts. An AL function for (P) can be obtained in two steps: augmenting the pre-Lagrangian $\mathcal{L}^{\mathrm{S}}$ with a penalty term, and then marginalizing over the auxiliary variables.

For some penalty parameter $\mu > 0$, the AL function $\mathcal{L}^{\mathrm{S}}_\mu \colon \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^m \to \overline{\mathbb{R}}$ associated to ($\mathrm{P_S}$) entails the sum of the pre-Lagrangian $\mathcal{L}^{\mathrm{S}}$ and a quadratic penalty for the constraint violation, scaled by μ. This leads to the definition of $\mathcal{L}^{\mathrm{S}}_\mu$ as

$$\mathcal{L}^{\mathrm{S}}_\mu(x,z,y) \coloneqq \mathcal{L}^{\mathrm{S}}(x,z,y) + \frac{1}{2\mu}\|c(x) - z\|^2.$$

Then, since ($\mathrm{P_S}$) involves the minimization over both original and auxiliary variables, whereas the latter ones do not appear in the original problem (P), we consider the marginalization of $\mathcal{L}^{\mathrm{S}}$ over z, which yields the *augmented Lagrangian* function $\mathcal{L}_\mu \colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ associated to (P):

$$\begin{aligned} \mathcal{L}_\mu(x,y) \coloneqq \inf_z \mathcal{L}^{\mathrm{S}}_\mu(x,z,y) &= f(x) + \inf_z \left\{ g(z) + \frac{1}{2\mu}\|c(x) + \mu y - z\|^2 \right\} - \frac{\mu}{2}\|y\|^2 \\ &= f(x) + g^\mu(c(x) + \mu y) - \frac{\mu}{2}\|y\|^2. \end{aligned} \tag{3.7}$$

Notice that the minimization over z is well-defined only for sufficiently small penalty parameters, relative to the prox-boundedness threshold of g, in particular $\mu \in (0, \mu_g)$. Moreover, we highlight that the Moreau envelope $g^\mu \colon \mathbb{R}^m \to \mathbb{R}$ of g is real-valued and strictly continuous [55, Ex. 10.32], but not continuously differentiable in general, as the proximal mapping of g is possibly set-valued.

With the AL tools at hand, one can readily recover the M-stationarity conditions (3.4) for (P). Through the augmented pre-Lagrangian function $\mathcal{L}^{\mathrm{S}}_\mu$ of (P), the first-order optimality conditions in the form of (3.3) can be written, for any $\mu > 0$, as $0 \in \partial_x \mathcal{L}^{\mathrm{S}}_\mu(\bar{x}, c(\bar{x}), \bar{y})$, $0 \in \partial_z \mathcal{L}^{\mathrm{S}}_\mu(\bar{x}, c(\bar{x}), \bar{y})$, and $0 \in \partial_y \mathcal{L}^{\mathrm{S}}_\mu(\bar{x}, c(\bar{x}), \bar{y})$, which hold if and only if $(\bar{x}, \bar{y}) \in \mathbb{R}^n \times \mathbb{R}^m$ satisfies (3.4).

The following lemma, which is inspired by [13, Lem. 3.1], will come in handy later on.

Lemma 3.2. *If $x \in \mathbb{R}^n$ is feasible for (P), then $\mathcal{L}_\mu(x, y) \le \varphi(x)$ for all $\mu > 0$ and $y \in \mathbb{R}^m$.*

Proof. By feasibility of x, we have $c(x) \in \operatorname{dom} g$. Then, for all $\mu > 0$ and $y \in \mathbb{R}^m$,

$$g^\mu(c(x) + \mu y) = \inf_z \left\{ g(z) + \frac{1}{2\mu}\|c(x) + \mu y - z\|^2 \right\} \le g(c(x)) + \frac{1}{2\mu}\|\mu y\|^2 = g(c(x)) + \frac{\mu}{2}\|y\|^2$$

by selecting $z = c(x) \in \operatorname{dom} g$. Hence, $\mathcal{L}_\mu(x, y) = f(x) + g^\mu(c(x) + \mu y) - \frac{\mu}{2}\|y\|^2 \le f(x) + g(c(x)) = \varphi(x)$, concluding the proof. □

In view of the AL subproblems arising in Section 4 below, the subsequent remark considers the notion of Υ-stationarity, discussed already in Section 2.2, to the AL function $\mathcal{L}_\mu$.

Remark 3.3. Motivated by the minimization of the AL function $\mathcal{L}_\mu(\cdot, \hat{y})$ where $\hat{y} \in \mathbb{R}^m$ is fixed, we are interested in pairs $(\bar{x}, \bar{z}) \in \mathbb{R}^n \times \mathbb{R}^m$ which certificate ε-Υ-stationarity of $\bar{x}$ for $\mathcal{L}_\mu(\cdot, \hat{y})$, for some given $\varepsilon \ge 0$. A simple calculation reveals that

$$\Upsilon(\bar{x}) = \bigcup_{\bar{z} \in \operatorname{prox}_{\mu g}(c(\bar{x}) + \mu \hat{y})} \left\{ \nabla_x \mathcal{L}(\bar{x}, \bar{y}) \,\middle|\, \bar{y} := \hat{y} + \frac{c(\bar{x}) - \bar{z}}{\mu} \in \partial g(\bar{z}) \right\}$$

holds in this situation. Clearly, $\bar{z} \in \operatorname{prox}_{\mu g}(c(\bar{x}) + \mu \hat{y})$ always gives $\hat{y} + (c(\bar{x}) - \bar{z})/\mu \in \partial g(\bar{z})$ by Fermat's rule (the converse is true for convex g), so ε-Υ-stationarity boils down to the existence of some $\bar{z} \in \operatorname{prox}_{\mu g}(c(\bar{x}) + \mu \hat{y})$ such that $\|\nabla_x \mathcal{L}(\bar{x}, \bar{y})\| \le \varepsilon$ where $\bar{y} := \hat{y} + (c(\bar{x}) - \bar{z})/\mu$. Note that this implicitly demands $\mu \in (0, \mu_g)$. Obviously, for arbitrary $\mu > 0$ and any pair $(\bar{x}, \bar{z})$ certificating Υ-stationarity (where $\varepsilon := 0$) of $\bar{x}$ for $\mathcal{L}_\mu(\cdot, \hat{y})$ in the above sense such that $\bar{z} = c(\bar{x})$ holds, $\bar{x}$ is also M-stationary. The converse is true whenever g is a convex function, and, in this case, the proximal mapping is single-valued.

Some of the concepts addressed in this section are visualized in the following example in terms of standard nonlinear programming.

Example 3.4. Nonlinear programming can be cast in the form (P) via many reformulations. Let us consider the setting

$$\text{(NLP)} \qquad \operatorname*{minimize}_x \quad f(x) \qquad \text{subject to} \quad c(x) \in C$$

with $g \equiv \delta_C$ being the indicator of a nonempty, closed, convex set $C := [c_l, c_u]$. Allowing entries of c_l and c_u to take infinite values, namely $c_l \in (\mathbb{R} \cup \{-\infty\})^m$ and $c_u \in (\mathbb{R} \cup \{\infty\})^m$, the model includes equalities, inequalities, and bounds in a compact form, and the constraint set C is convex polyhedral. The pre-Lagrangian for (NLP) with auxiliary variable $z \in \mathbb{R}^m$ and multiplier $y \in \mathbb{R}^m$ reads

$$\mathcal{L}^{\mathrm{S}}(x, z, y) = f(x) + \delta_C(z) + \langle y, c(x) - z \rangle.$$

The M-stationarity conditions of (NLP) can be expressed as

$$\nabla f(\bar{x}) + c'(\bar{x})^\top \bar{y} = 0, \qquad \bar{y} \in N_C(c(\bar{x})),$$

where the inclusion coincides with the classical complementarity conditions and imposes the feasibility condition $c(\bar{x}) \in C$ as well. The Lagrangian $\mathcal{L}$ for (NLP) is $\mathcal{L}(x, y) = f(x) + \langle y, c(x)\rangle$ and the AL $\mathcal{L}_\mu$ is given by

$$\mathcal{L}_\mu(x, y) = f(x) + \frac{1}{2\mu} \operatorname{dist}^2_C(c(x) + \mu y) - \frac{\mu}{2}\|y\|^2,$$

recovering all classical quantities. As C is convex in (NLP), the squared distance term in the AL function is continuously differentiable, see e.g. [2, Cor. 12.30].

Remark 3.5. Yet another way to the definition of a Lagrangian-type function in composite optimization with *convex* function g has been promoted by Rockafellar in his recent papers [53, 54] where he introduces the so-called *generalized* Lagrangian of (P) by marginalization of the pre-Lagrangian $\mathcal{L}^S$ given in (3.1). The marginalization step enters here because ($\mathrm{P_S}$) involves the minimization over both original and auxiliary variables, whereas the latter ones do not appear in (P). Marginalization of $\mathcal{L}^S$ over z results in the generalized Lagrangian $\ell\colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R} \cup \{-\infty\}$ given by

$$\ell(x, y) := \inf_z \mathcal{L}^S(x, z, y) = f(x) + \langle y, c(x)\rangle + \inf_z\{g(z) - \langle y, z\rangle\} = \mathcal{L}(x, y) - g^*(y),$$

where $\mathcal{L}$ is the Lagrangian defined in (3.2). Clearly, it is $\nabla_x \mathcal{L} = \nabla_x \ell$. One could hope that the generalized Lagrangian ℓ encapsulates all information needed to state the M-stationarity conditions (3.4) for (P), exemplary as $0 \in \partial_x \ell(\bar{x}, \bar{y})$, $0 \in \partial_y(-\ell)(\bar{x}, \bar{y})$. The negative sign appearing for the multipliers relates to the (generalized) saddle-point properties of the (generalized) Lagrangian. Expanding terms, this gives

$$0 = \nabla_x \mathcal{L}(\bar{x}, \bar{y}), \tag{3.8a}$$

$$c(\bar{x}) \in \partial g^*(\bar{y}). \tag{3.8b}$$

If g is convex, (3.8b) is equivalent to (3.4b), see [55, Prop. 11.3], so that M-stationarity can be fully characterized via the derivatives of the generalized Lagrangian. Whenever g is a nonconvex function, however, this reasoning is no longer possible. Under additional assumptions on g (and g^*), one may apply the convex hull property, see e.g. [19, formula (2.7)], and a marginal function rule, see e.g. [7, Thm 5.1] or [55, Thm 10.13], in order to find

$$\partial g^*(y) \subseteq -\operatorname{conv} \partial(-g^*)(y) \subseteq -\operatorname{conv}\{-z \in \mathbb{R}^m \mid y \in \partial g(z)\} = \operatorname{conv}(\partial g)^{-1}(y).$$

Hence, whenever $(\partial g)^{-1}(y)$ is convex, (3.8b) yields (3.4b) if the aforementioned calculus rules apply. Consequently, under additional assumptions, (3.8) implies the M-stationarity conditions (3.4) even for nonconvex g. However, the converse implication is likely to fail even in very simple situations, as illustrated in the subsequently stated Example 3.6.

Example 3.6. We investigate the model problem

$$\underset{x}{\text{minimize}} \quad f(x) + \|c(x)\|_0, \tag{3.9}$$

where g plays the role of the ℓ_0-quasi-norm $\|\cdot\|_0\colon \mathbb{R}^m \to \mathbb{R}$, which simply counts the nonzero entries of the argument vector. Clearly, $\|\cdot\|_0$ is a merely lower semicontinuous function and is not convex. For some point $\bar{x} \in \mathbb{R}^n$, we will exploit the index sets

$$I^0(\bar{x}) := \{i \in \{1, \ldots, m\} \mid c_i(\bar{x}) = 0\}, \quad I^\pm(\bar{x}) := \{1, \ldots, m\} \setminus I^0(\bar{x}).$$

One can easily check that

$$\partial\|\cdot\|_0(c(\bar{x})) = \{y \in \mathbb{R}^m \mid \forall i \in I^\pm(\bar{x})\colon\ y_i = 0\}$$

holds true. Hence, $\bar{x}$ is M-stationary for (3.9) if and only if there is some $\bar{y} \in \mathbb{R}^m$ such that (3.4a) is valid and, for all $i \in I^\pm(\bar{x})$, it is $\bar{y}_i = 0$. A simple calculation reveals that $\|\cdot\|_0^* = \delta_{\{0\}}$, which is why condition (3.8b) reduces to $\bar{y} = 0$. This is a much stronger requirement on the multiplier than the one demanded by M-stationarity.

3.2 Second-order sufficient optimality conditions

In this subsection, we briefly review the second-order sufficient optimality condition for (P) which has been derived in [8, Sec. 6].

Let us fix a feasible point $\bar{x} \in \mathbb{R}^n$ of (P), and define the *critical cone* of (P) at $\bar{x}$ by means of

$$C(\bar{x}) := \{u \in \mathbb{R}^n \mid f'(\bar{x})u + \mathrm{d}g(c(\bar{x}))(c'(\bar{x})u) \leq 0\}.$$

Furthermore, for each $u \in C(\bar{x})$, we make use of the *directional multiplier set* $\Lambda(\bar{x}, u)$ given by

$$\Lambda(\bar{x}, u) := \{y \in \mathbb{R}^m \mid \nabla_x \mathcal{L}(\bar{x}, y) = 0,\ \mathrm{d}g(c(\bar{x}))(c'(\bar{x})u) = \langle y, c'(\bar{x})u\rangle,\ \mathrm{d}^2 g(c(\bar{x}), y)(c'(\bar{x})u) > -\infty\}.$$

Let us mention that this definition can be stated equivalently in terms of the so-called directional proximal subdifferential of g, see [8, Sec. 3.2] for details.

Definition 3.7 (Second-order sufficient condition). For a feasible point $\bar{x} \in \mathbb{R}^n$ of (P), we say that the *Second-Order Sufficient Condition* (SOSC for brevity) is valid, whenever

$$\forall u \in C(\bar{x}) \setminus \{0\},\ \exists y \in \Lambda(\bar{x}, u): \quad \nabla^2_{xx}\mathcal{L}(\bar{x}, y)[u, u] + \mathrm{d}^2 g(c(\bar{x}), y)(c'(\bar{x})u) > 0.$$

Let us fix a feasible point $\bar{x} \in \mathbb{R}^n$ of (P) where SOSC is valid. To avoid trivial situations, we assume that $C(\bar{x})$ contains a non-vanishing direction $u \in \mathbb{R}^n$. Clearly, SOSC requires that $\Lambda(\bar{x}, u)$ is nonempty, and since we have $\Lambda(\bar{x}, u) \subseteq \Lambda(\bar{x})$ from [5, Prop. 2.9] and [8, Cor. 3.14], $\Lambda(\bar{x})$ is nonempty as well, i.e., $\bar{x}$ is M-stationary. Thus, checking validity of SOSC is only reasonable at M-stationary points.

The following result is a consequence of [8, Thm 6.1] and the associated discussions.

Proposition 3.8. *Let $\bar{x} \in \mathbb{R}^n$ be a feasible point of* (P) *where SOSC is valid. Then there exist constants $\varepsilon > 0$ and $\beta > 0$ such that the second-order growth condition*

$$\forall x \in \mathbb{B}_\varepsilon(\bar{x}): \quad \varphi(x) - \varphi(\bar{x}) \geq \frac{\beta}{2}\|x - \bar{x}\|^2 \tag{3.10}$$

is valid. Particularly, $\bar{x}$ is a strict local minimizer of (P).

Since the composite optimization problem (P) can be recast as the constrained problem

$$\operatorname*{minimize}_{x,\tau} \quad f(x) + \tau \quad \text{subject to} \quad (c(x), \tau) \in \operatorname{epi} g,$$

Proposition 3.8 is also a consequence of [5, Thm 3.3] when taking into account the inequality

$$\mathrm{d}^2\, \delta_{\operatorname{epi} g}((c(\bar{x}), g(c(\bar{x}))), (y, -1))(c'(\bar{x})u, v) \geq \mathrm{d}^2 g(c(\bar{x}), y)(c'(\bar{x})u)$$

which holds for each $u \in \mathbb{R}^n$ and $v \in \mathbb{R}$, see [8, Prop. 3.13].

The following corollary provides a sufficient condition for SOSC which will be of interest later on.

Corollary 3.9. *Let $\bar{x} \in \mathbb{R}^n$ be an M-stationary point for* (P)*, and fix $\bar{y} \in \Lambda(\bar{x})$. Furthermore, let the condition*

$$\forall u \in C(\bar{x}) \setminus \{0\}: \quad \nabla^2_{xx}\mathcal{L}(\bar{x}, \bar{y})[u, u] + \mathrm{d}^2 g(c(\bar{x}), \bar{y})(c'(\bar{x})u) > 0 \tag{3.11}$$

hold. Then SOSC and, thus, the second-order growth condition (3.10) *are valid.*

Proof. We will show that $\bar{y} \in \Lambda(\bar{x}, u)$ holds for each $u \in C(\bar{x}) \setminus \{0\}$. Then it is clear that (3.11) implies validity of SOSC, and the final assertion is just a consequence of Proposition 3.8.

Fix $u \in C(\bar{x}) \setminus \{0\}$. Then (3.11) obviously implies $\mathrm{d}^2 g(c(\bar{x}), \bar{y})(c'(\bar{x})u) > -\infty$, and [8, formula (5)] immediately yields $\mathrm{d}g(c(\bar{x}))(c'(\bar{x})u) \geq \langle \bar{y}, c'(\bar{x})u\rangle$. By definition of the critical cone and $\nabla_x \mathcal{L}(\bar{x}, \bar{y}) = 0$, we also have

$$\mathrm{d}g(c(\bar{x}))(c'(\bar{x})u) \leq -f'(\bar{x})u = \langle \bar{y}, c'(\bar{x})u\rangle.$$

Thus, $\bar{y} \in \Lambda(\bar{x}, u)$ follows. ◻

3.3 Error bounds

Here, we aim to establish a connection between the second-order sufficient conditions from Definition 3.7 and an error bound property. Relating to stability properties and involving the distance to the primal-dual solution set, error bounds are an essential ingredient for deriving rates of local convergence for numerical methods addressing (P). In order to quantify the violation of the M-stationarity conditions from Definition 3.1, it is (almost) natural to define the residual mapping

$$\Theta(x, z, y) := \|\nabla_x \mathcal{L}(x, y)\| + \|c(x) - z\| + \operatorname{dist}(y, \partial g(z)). \tag{3.12}$$

Clearly, the M-stationarity conditions (3.4) for some $(\bar{x}, \bar{y}) \in \mathbb{R}^n \times \mathbb{R}^m$ are equivalent to $\Theta(\bar{x}, c(\bar{x}), \bar{y}) = 0$. We shall see now that, under certain assumptions, Θ allows us to quantify not only the violation of (3.4), but also the distance to the (primal-dual) solution set.

The proof of the following proposition, which provides the foundations of our analysis in this section, relates the error bound property of our interest with the strong metric subregularity of a certain set-valued mapping. Moreover, the latter characterization is quantifiable via a condition which can be stated in terms of initial problem data, thanks to the Levy–Rockafellar criterion.

Proposition 3.10. *Let $\bar{x} \in \mathbb{R}^n$ be an M-stationary point of (P) and pick $\bar{y} \in \Lambda(\bar{x})$. Assume that the qualification condition*

$$0 = \nabla^2_{xx} \mathcal{L}(\bar{x}, \bar{y})u + c'(\bar{x})^\top \eta,\ \eta \in D(\partial g)(c(\bar{x}), \bar{y})(c'(\bar{x})u) \implies u = 0,\ \eta = 0 \tag{3.13}$$

is valid. Then there are a constant $\varrho_\mathrm{u} > 0$ and a neighborhood U of $(\bar{x}, c(\bar{x}), \bar{y})$ such that, for each $(x, z, y) \in U \cap (\mathbb{R}^n \times \operatorname{dom} g \times \mathbb{R}^m)$, we have the upper estimate

$$\|x - \bar{x}\| + \|z - c(\bar{x})\| + \|y - \bar{y}\| \le \varrho_\mathrm{u}\, \Theta(x, z, y). \tag{3.14}$$

Proof. We define a set-valued mapping $G\colon \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^m \rightrightarrows \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^m$ by means of

$$G(x, z, y) := (\nabla_x \mathcal{L}(x, y), c(x) - z, y) - \{0\} \times \{0\} \times \partial g(z).$$

By continuous differentiability of the single-valued part of G, one can easily check, e.g., by means of the change-of-coordinates formula for tangents from [55, Ex. 6.7], that

$$\begin{aligned} &DG((\bar{x}, c(\bar{x}), \bar{y}), (0, 0, 0))(u, v, \eta) \\ &\qquad = (\nabla^2_{xx} \mathcal{L}(\bar{x}, \bar{y})u + c'(\bar{x})^\top \eta, c'(\bar{x})u - v, \eta) - \{0\} \times \{0\} \times D(\partial g)(\bar{y}, c(\bar{x}))(v) \end{aligned}$$

holds. Thus, (3.13) is equivalent to $\ker DG((\bar{x}, c(\bar{x}), \bar{y}), (0, 0, 0)) = \{(0, 0, 0)\}$. By the Levy–Rockafellar criterion, G is strongly metrically subregular at $(\bar{x}, c(\bar{x}), \bar{y})$, and the latter is equivalent to the desired error bound condition. □

Note that the proof of Proposition 3.10 actually shows that (3.13) is *equivalent* to the local validity of the error bound property (3.14).

Corollary 3.11. *Let $\bar{x} \in \mathbb{R}^n$ be an M-stationary point for (P), and fix $\bar{y} \in \Lambda(\bar{x})$. If the second-order condition (3.11) is valid, and if we have*

$$\forall u \in \mathbb{R}^n,\ \forall \eta \in D(\partial g)(c(\bar{x}), \bar{y})(c'(\bar{x})u)\colon \quad \langle \eta, c'(\bar{x})u \rangle \ge \mathrm{d}^2 g(c(\bar{x}), \bar{y})(c'(\bar{x})u) \tag{3.15}$$

and

$$D(\partial g)(c(\bar{x}), \bar{y})(0) \cap \ker c'(\bar{x})^\top = \{0\}, \tag{3.16}$$

then there are a constant $\varrho_\mathrm{u} > 0$ and a neighborhood U of $(\bar{x}, c(\bar{x}), \bar{y})$ such that the upper estimate (3.14) holds for each triplet $(x, z, y) \in U \cap (\mathbb{R}^n \times \operatorname{dom} g \times \mathbb{R}^m)$.

Proof. We just show that the qualification condition (3.13) is valid. Then the assertion follows from Proposition 3.10.

Thus, pick $u \in \mathbb{R}^n$ and $\eta \in \mathbb{R}^m$ with $0 = \nabla^2_{xx}\mathcal{L}(\bar{x}, \bar{y})u + c'(\bar{x})^\top \eta$ and $\eta \in D(\partial g)(c(\bar{x}), \bar{y})(c'(\bar{x})u)$. Taking the scalar product of the equation with u gives $\nabla^2_{xx}\mathcal{L}(\bar{x}, \bar{y})[u, u] + \langle \eta, c'(\bar{x})u\rangle = 0$, so that (3.15) gives $\nabla^2_{xx}\mathcal{L}(\bar{x}, \bar{y})[u, u] + \mathrm{d}^2 g(c(\bar{x}), \bar{y})(c'(\bar{x})u) \leq 0$. If $u \notin C(\bar{x})$, we have

$$\mathrm{d}g(c(\bar{x}))(c'(\bar{x})u) > -f'(\bar{x})u = \langle \bar{y}, c'(\bar{x})u\rangle \tag{3.17}$$

from $\nabla_x \mathcal{L}(\bar{x}, \bar{y}) = 0$, which gives $\mathrm{d}^2 g(c(\bar{x}), \bar{y})(c'(\bar{x})u) = \infty$, see [8, formula (5)], and, thus, a contradiction. Hence, we have $u \in C(\bar{x})$, and (3.11) gives $u = 0$. Thus, $\eta \in D(\partial g)(c(\bar{x}), \bar{y})(0) \cap \ker c'(\bar{x})^\top$, and (3.16) yields $\eta = 0$. Consequently, (3.13) is valid, and the assertion follows. □

Remark 3.12. Let us note that (3.15) is valid whenever g is prox-regular, subdifferentially continuous, and twice epi-differentiable at $c(\bar{x})$ for $\bar{y}$, see Lemma 2.7. Due to [55, Ex. 13.30], each proper, lower semicontinuous, convex function is prox-regular and subdifferentially continuous on its domain. Exemplary, whenever g is a convex piecewise linear-quadratic function or the indicator function of the second-order cone, then it is twice epi-differentiable as well, see [55, Prop. 13.9] and [26, Thm 3.1].

Observe that validity of (3.13) is equivalent to validity of (3.16) and

$$0 = \nabla^2_{xx}\mathcal{L}(\bar{x}, \bar{y})u + c'(\bar{x})^\top \eta,\ \eta \in D(\partial g)(c(\bar{x}), \bar{y})(c'(\bar{x})u) \implies u = 0. \tag{3.18}$$

The proof of Corollary 3.11 shows that validity of (3.11) and (3.15) implies that (3.18) holds. Let us elaborate on (3.16). Similar considerations in a much more specific setting can be found in [42, Sec. 8] and [45, Sec. 4] where g is assumed to be a convex function of special type. Notice that the results obtained in [42, 45] are, expectedly, slightly stronger.

Lemma 3.13. *Let $\bar{x} \in \mathbb{R}^n$ be an M-stationary point for* (P)*, and fix $\bar{y} \in \Lambda(\bar{x})$. We investigate the mapping $H\colon \mathbb{R}^m \rightrightarrows \mathbb{R}^n \times \mathbb{R}^m$ given by*

$$\forall y \in \mathbb{R}^m\colon \quad H(y) := (\nabla_x \mathcal{L}(\bar{x}, y), y) - \{0\} \times \partial g(c(\bar{x})). \tag{3.19}$$

Then (3.16) *implies that H is strongly metrically subregular at $(\bar{y}, (0, 0))$, and the converse holds true whenever $(\partial g)^{-1}$ is metrically subregular at $(\bar{y}, c(\bar{x}))$.*

Proof. Patterning the proof of Proposition 3.10, we find

$$DH(\bar{y}, (0, 0))(\eta) = (c'(\bar{x})^\top \eta, \eta) - \{0\} \times T_{\partial g(c(\bar{x}))}(\bar{y}).$$

Furthermore, we have $T_{\partial g(c(\bar{x}))}(\bar{y}) \subseteq D(\partial g)(c(\bar{x}), \bar{y})(0)$, and the converse holds true whenever $(\partial g)^{-1}$ is metrically subregular at $(\bar{y}, c(\bar{x}))$, see [7, Thm 3.2]. Thus, (3.16) implies

$$\ker DH(\bar{y}, (0, 0)) = \{0\},$$

and the converse holds true under the additional subregularity of $(\partial g)^{-1}$. Hence, the assertion follows from the Levy–Rockafellar criterion. □

Corollary 3.14. *Let $\bar{x} \in \mathbb{R}^n$ be an M-stationary point for* (P)*, and fix $\bar{y} \in \Lambda(\bar{x})$. Then the following assertions hold.*

- *(a) If* (3.16) *holds, then there is a neighborhood $V \subseteq \mathbb{R}^m$ of $\bar{y}$ such that $\Lambda(\bar{x}) \cap V = \{\bar{y}\}$. Particularly, if $\Lambda(\bar{x})$ is convex (which happens if $\partial g(c(\bar{x}))$ is convex), then $\Lambda(\bar{x}) = \{\bar{y}\}$.*
- *(b) If $\Lambda(\bar{x}) = \{\bar{y}\}$, if the mapping H from* (3.19) *is metrically subregular at $(\bar{y}, (0, 0))$, and if $(\partial g)^{-1}$ is metrically subregular at $(\bar{y}, c(\bar{x}))$ (both subregularity assumptions are satisfied if ∂g is a polyhedral mapping as this also gives polyhedrality of H), then* (3.16) *holds.*

Proof. Due to Lemma 3.13, the assumptions in statement (a) guarantee that H from (3.19) is strongly metrically subregular at $(\bar{y},(0,0))$. Hence, we find a neighborhood $V \subseteq \mathbb{R}^m$ of $\bar{y}$ and a constant $\kappa > 0$ such that

$$\forall y \in V: \quad \|y - \bar{y}\| \leq \kappa\big(\|\nabla_x \mathcal{L}(\bar{x}, y)\| + \operatorname{dist}(y, \partial g(c(\bar{x})))\big). \tag{3.20}$$

Particularly, this estimate shows that whenever $y \in V$ is different from $\bar{y}$, then $y \notin \Lambda(\bar{x})$. The additional statement in assertion (a) readily follows.

For assertion (b), notice first that $H^{-1}(0,0) = \Lambda(\bar{x})$ is valid. Then metric subregularity of H at $(\bar{y},(0,0))$ together with $\Lambda(\bar{x}) = \{\bar{y}\}$ show that (3.20) is valid for some neighborhood $V \subseteq \mathbb{R}^m$ of $\bar{y}$ and some constant $\kappa > 0$. Hence, H is strongly metrically subregular at $(\bar{y},(0,0))$. Finally, metric subregularity of $(\partial g)^{-1}$ at $(\bar{y}, c(\bar{x}))$ and Lemma 3.13 can be used to infer validity of (3.16). ◻

In the following remark, we comment on (3.18).

Remark 3.15. Let $\bar{x} \in \mathbb{R}^n$ be an M-stationary point for (P), and fix $\bar{y} \in \Lambda(\bar{x})$. Suppose that the second-order condition (3.11) is valid, and that $0 \in \operatorname{dom} \mathrm{d}^2 g(c(\bar{x}), \bar{y})$. We first note that this means that $\bar{u} := 0$ is the uniquely determined global minimizer of

$$\underset{u}{\text{minimize}} \quad \frac{1}{2}\nabla^2_{xx}\mathcal{L}(\bar{x},\bar{y})[u,u] + \frac{1}{2}\mathrm{d}^2 g(c(\bar{x}),\bar{y})(c'(\bar{x})u). \tag{3.21}$$

In order to see this, one has to observe two facts. First, for each $u \notin C(\bar{x})$, we have (3.17) which gives $\mathrm{d}^2 g(c(\bar{x}),\bar{y})(c'(\bar{x})u) = \infty$ as mentioned earlier. Secondly, $\mathrm{d}^2 g(c(\bar{x}),\bar{y})(0) = 0$ follows from Lemma 2.6.

Under the assumptions of Lemma 2.7, the limiting subdifferential of the scaled second subderivative $v \mapsto \frac{1}{2}\mathrm{d}^2 g(c(\bar{x}),\bar{y})(v)$ can be computed in terms of the graphical derivative of ∂g at $(c(\bar{x}),\bar{y})$. Hence, under some suitable assumptions, the chain rule from [43, Cor. 4.6] can be applied in order to derive first-order necessary optimality conditions for problem (3.21), and these conditions take the following shape:

$$0 = \nabla^2_{xx}\mathcal{L}(\bar{x},\bar{y})u + c'(\bar{x})^\top \eta, \quad \eta \in D(\partial g)(c(\bar{x}),\bar{y})(c'(\bar{x})u).$$

That is why (3.18) demands, roughly speaking, that $\bar{u} := 0$ is the uniquely determined stationary point of (3.21). This is clearly different from postulating that this point is the uniquely determined global minimizer of this problem, i.e., (3.11). Equivalence of these conditions can only be guaranteed under some additional convexity of the second subderivative.

Following [44, Def. 3.1], validity of (3.18) demands that the multiplier $\bar{y}$ is so-called *noncritical*. The concept of critical multipliers dates back to [32, 34] where it has been introduced for standard nonlinear programs. In [34], it has been pointed out that the presence of critical multipliers slows down the convergence of Newton-type methods when applied for the solution of stationarity systems, and this observation can be extended to certain composite optimization problems as shown in [44, 45]. Let us mention that [44, Thm 4.1] and [45, Thm 5.6] justify that (3.18) is equivalent to a local primal-dual (upper) error bound, based on a not necessarily unique Lagrange multiplier, whenever g is a function whose epigraph is a convex polyhedral set or the indicator function of a so-called C^2-cone reducible set, and in the latter case, further assumptions are required. For brevity, we abstain here from investigating further refinements of the error bound (3.14) to situations where the Lagrange multiplier is not uniquely determined, but indicate that this is an interesting topic for future research.

The above puts our comments from Remark 3.12 into some new light. On the one hand, our arguments highlight that in situations where the second subderivative of g is not convex, (3.11) might be too weak to yield the error bound of our interest. On the other hand, in the absence of any additional assumptions, (3.13) may not be sufficient for (3.11) as uniqueness of stationary points says nothing about the existence of a global minimizer for (3.21). However, the condition

$$\forall u \in \mathbb{R}^n \setminus \{0\},\ \forall \eta \in D(\partial g)(c(\bar{x}),\bar{y})(c'(\bar{x})u): \quad \nabla^2_{xx}\mathcal{L}(\bar{x},\bar{y})[u,u] + \langle \eta, c'(\bar{x})u\rangle > 0 \tag{3.22}$$

is clearly sufficient for (3.18) and, together with (3.16), gives (3.13).

The following example shows that (3.18) does not necessarily imply (3.11). Furthermore, Example 5.1 below visualizes that (3.11) does not necessarily imply (3.18). These two conditions are, thus, independent in general as indicated in Remark 3.15.

Example 3.16. We consider (P) for the functions $f, c, g\colon \mathbb{R} \to \mathbb{R}$ given by $f(x) := \frac{1}{2}x^2$, $c(x) := x$, $g(z) := -z^2$, and choose $\bar{x}$ to be the origin in $\mathbb{R}$. Note that $\bar{x}$ is M-stationary with $\Lambda(\bar{x}) = \{0\}$. Thus, we consider the uniquely determined multiplier $\bar{y} := 0$. Obviously, $\bar{x}$ is a strict local maximizer of (P) which is why (3.11) fails to hold at $\bar{x}$ for $\bar{y}$, see Corollary 3.9. Clearly, we have $\nabla^2_{xx}\mathcal{L}(\bar{x}, \bar{y}) = 1$, and twice continuous differentiability of g gives $D(\partial g)(c(\bar{x}), \bar{y})(c'(\bar{x})u) = \{-2u\}$ for each $u \in \mathbb{R}$. Hence, one can easily check that (3.18) is valid.

Remark 3.17 (Geometric constraints). In the special case where $g := \delta_D$ holds for some closed set $D \subseteq \mathbb{R}^m$, the qualification conditions (3.13), (3.16), (3.18), and (3.22) involve the graphical derivative of the (limiting) normal cone mapping associated with D. For several different choices of D, including convex cones (like the semidefinite or the second-order cone) or convex sets given via smooth convex inequality constraints, explicit ready-to-use formulas for this variational object are available, see e.g. [24, 58]. For diverse nonconvex sets D of special structure, like sparsity sets of type $\{x \in \mathbb{R}^n \mid \|x\|_0 \leq \kappa\}$, $k \in \{1, \ldots, n-1\}$, the explicit computation of this tool is possible as well.

As we have seen above, the *upper* error bound in (3.14) only holds in the presence of comparatively strong assumptions. Unfortunately, due to our definition of Θ in (3.12) which comprises the distance to the subdifferential of g, the converse *lower* error bound seems to demand even more prohibitive assumptions, as the following Remark 3.18 illustrates. Nonetheless, we will circumnavigate this potentially crucial observation later on in Section 4 by the design of our algorithm.

Remark 3.18. Let $\bar{x} \in \mathbb{R}^n$ be an M-stationary point of (P) and pick $\bar{y} \in \Lambda(\bar{x})$. Then, relying on the definition of Θ in (3.12), it appears indispensable for estimating the distance to the subdifferential to assume its *inner calmness*, see [4, Def. 2.2] and [3, Sec. 2] for a discussion of this property. In particular, ∂g shall be inner calm at $(c(\bar{x}), \bar{y})$, which entails the existence of $\kappa > 0$ and a neighborhood $V \subseteq \mathbb{R}^m$ of $c(\bar{x})$ such that

$$\forall z \in V\colon \qquad \operatorname{dist}(\bar{y}, \partial g(z)) \leq \kappa\|z - c(\bar{x})\|.$$

With this property at hand and exploiting (3.4), for each triplet $(x, z, y) \in \mathbb{R}^n \times \operatorname{dom} g \times \mathbb{R}^m$ such that $z \in V$, the triangle inequality yields

$$\begin{aligned}\Theta(x, z, y) &= \|\nabla_x\mathcal{L}(x, y)\| + \|c(x) - z\| + \operatorname{dist}(y, \partial g(z)) \\ &\leq \|\nabla_x\mathcal{L}(x, y) - \nabla_x\mathcal{L}(\bar{x}, \bar{y})\| + \|c(x) - c(\bar{x})\| + \|c(\bar{x}) - z\| + \|y - \bar{y}\| + \operatorname{dist}(\bar{y}, \partial g(z)) \\ &\leq \|\nabla_x\mathcal{L}(x, y) - \nabla_x\mathcal{L}(\bar{x}, \bar{y})\| + \|c(x) - c(\bar{x})\| + (\kappa + 1)\|z - c(\bar{x})\| + \|y - \bar{y}\|.\end{aligned}$$

Then, noting that $\nabla_x\mathcal{L}$ and c are locally Lipschitz continuous by Assumption 1.1(i), there are a constant $\varrho_1 > 0$ and a neighborhood U of $(\bar{x}, c(\bar{x}), \bar{y})$ such that, for each $(x, z, y) \in U \cap (\mathbb{R}^n \times \operatorname{dom} g \times \mathbb{R}^m)$, we obtain the lower estimate

$$\varrho_1\,\Theta(x, z, y) \leq \|x - \bar{x}\| + \|z - c(\bar{x})\| + \|y - \bar{y}\|,$$

which patterns the upper counterpart in (3.14).

However, inner calmness of the subdifferential is an impractical assumption, even for convex g, and it would restrict our considerations mainly to points where g is smooth in practice. Exemplary, consider the absolute value function $g := |\cdot|$. Since ∂g is single-valued on $\mathbb{R} \setminus \{0\}$, one can easily check that inner calmness of ∂g at $(0, \bar{y})$ fails for every $\bar{y} \in [-1, 1] = \partial g(0)$.

In the following Section 4, we will *not* rely on any additional property of the subdifferential, but leverage instead the algorithmic scheme to derive a lower error bound *along the iterates*, see Lemma 4.10 below.

4 AUGMENTED LAGRANGIAN SCHEME AND CONVERGENCE

This section is devoted to describing a numerical scheme for solving (P) and to investigating its convergence properties under suitable assumptions. In particular, we consider the implicit AL scheme from [16, Alg. 4.1], so called because it makes use of the AL function $\mathcal{L}_\mu\colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ associated to (P), as defined in (3.7). It deviates in this respect from [17, Alg. 1], which builds upon ($\mathrm{P_S}$) and treats the auxiliary variable explicitly, see [6] for a discussion on implicit variables (and concealed benefits thereof) in optimization.

4.1 IMPLICIT APPROACH

The numerical method considered for addressing (P) is stated in Algorithm 4.1. Fitting into the AL framework [10, 11, 15], the main step of the iterative procedure involves minimizing the AL function with respect to the primal variable. At the k-th iteration of the AL scheme, with some given penalty parameter $\mu_k > 0$ and multiplier estimate $\hat{y}^k \in \mathbb{R}^m$, a subproblem involving the minimization of $\mathcal{L}_{\mu_k}(x, \hat{y}^k)$ over $x \in \mathbb{R}^n$ has to be solved approximately. However, the AL function may lack regularity since, for g nonconvex, the Moreau envelope g^μ is in general *not* continuously differentiable. Therefore, the concept of (approximate) Υ-stationarity introduced in Section 2.2 plays a role in characterizing adequate solutions of this AL subproblem, see Remark 3.3 as well. Let us note that, in practice, such points can be computed with the aid of a nonmonotone descent method, see [16, Sec. 5] for details. Then, following classical update rules [11], the multiplier estimate $\hat{y}$ and penalty parameter μ are adjusted, along with the subproblem's tolerance ε.

Compared to the classical AL approach for the solution of nonlinear programs, see [10, 15], this variant uses a safeguarded update rule for the Lagrange multipliers and has stronger global convergence properties, as demonstrated in [38]. The safeguarded multiplier estimate $\hat{y}^k$ is drawn from a bounded set $Y \subseteq \mathbb{R}^m$ at Step 2. In practice, it is advisable to choose the safeguarded estimate $\hat{y}^k$ as the projection of the multiplier y^{k-1} onto Y. We refer to [11, Sec. 4.1] for a detailed discussion.

The monotonicity test at Step 3 is adopted to monitor primal infeasibility along the iterates and update the penalty parameter accordingly. Aimed at driving V_k to zero, the penalty parameter μ_k is reduced in case of insufficient improvement.

Algorithm 4.1: Safeguarded implicit AL method for (P)

Data: $\mu_0 \in (0, \mu_g)$, $\theta \in (0,1)$, $\kappa \in (0,1)$, $Y \subseteq \mathbb{R}^m$ nonempty bounded
1 **for** $k = 0, 1, 2 \ldots$ **do**
2 Select $\hat{y}^k \in Y$ and $\varepsilon_k \geq 0$
3 Compute a pair (x^k, z^k) certificating ε_k-Υ-stationary of x^k for $\mathcal{L}_{\mu_k}(\cdot, \hat{y}^k)$:

$$\|\nabla_x \mathcal{L}^{\mathrm{S}}_{\mu_k}(x^k, z^k, \hat{y}^k)\| \leq \varepsilon_k, \qquad z^k \in \operatorname{prox}_{\mu_k g}(c(x^k) + \mu_k \hat{y}^k)$$

Set $y^k \leftarrow \hat{y}^k + \mu_k^{-1}[c(x^k) - z^k]$ and $V_k \leftarrow \|c(x^k) - z^k\|$
if $k = 0$ **or** $V_k \leq \theta V_{k-1}$ **then** $\mu_{k+1} \leftarrow \mu_k$, **else** $\mu_{k+1} \leftarrow \kappa \mu_k$

Before investigating the convergence properties of Algorithm 4.1, we provide some characterizations of the iterates $\{(x^k, z^k, y^k)\}$. These are direct consequences of z^k being a certificate of ε_k-Υ-stationarity for x^k, by Step 3, and of the dual update at Step 3, see Remark 3.3 as well.

Proposition 4.1. *Let $\{(x^k, z^k, y^k)\}$ be a sequence generated by Algorithm 4.1. Then, for each $k \in \mathbb{N}$, $z^k \in \operatorname{prox}_{\mu_k g}(c(x^k) + \mu_k \hat{y}^k) \subseteq \operatorname{dom} g$, $\|\nabla f(x^k) + c'(x^k)^\top y^k\| \leq \varepsilon_k$, and $y^k \in \partial g(z^k)$.*

Throughout the convergence analysis, based on, and extending, that of [16], it is assumed that Algorithm 4.1 is well-defined, thus requiring that each subproblem at Step 3 admits an approximately stationary point. Moreover, the existence of some accumulation point $\bar{x}$ for a sequence $\{x^k\}$ generated by Algorithm 4.1 requires, in general, coercivity or (level) boundedness arguments.

While asymptotic M-stationarity as given in [16, Def. 3.3] demands the existence of a sequence of, in a certain sense, approximately M-stationary points for (P) converging to the point of interest, no quantitative bound on this *approximativity* is required. However, for Algorithm 4.1 to terminate, there is a need for an approximate version of the M-stationarity concept of Definition 3.1. We refer to the notion delineated in [16, Def. 3.2], which comes along with an explicit bound quantifying violation of M-stationarity, while aligning with the asymptotic stipulation.

4.2 GLOBAL CONVERGENCE

In this section, we are concerned with global convergence properties related to Algorithm 4.1, i.e., we are going to study properties of accumulation points of sequences it generates, regardless of how it is initialized. For that purpose, we will assume that Algorithm 4.1 is well-defined and produces an infinite sequence of iterates.

Our first results pertain to a *global* optimization perspective on the subproblems at Step 3 of Algorithm 4.1, compare [11, Ch. 5] and [38, Sec. 4]. Solving each subproblem up to approximate global optimality, not necessarily with vanishing inner tolerance, one finds in the limit a global minimizer of the infeasibility measure, see [13, Lem. 4.2] for a related result.

Lemma 4.2. *Let $\{(x^k, z^k, y^k)\}$ be a sequence generated by Algorithm 4.1 with $\{\varepsilon_k\}$ bounded. Assume that* $\operatorname{dom} g$ *is closed and that*

$$\forall k \in \mathbb{N},\ \forall x \in \mathbb{R}^n: \quad \mathcal{L}_{\mu_k}(x^k, \hat{y}^k) \leq \mathcal{L}_{\mu_k}(x, \hat{y}^k) + \varepsilon_k. \tag{4.1}$$

Fix an accumulation point $\bar{x} \in \mathbb{R}^n$ of $\{x^k\}$. Then $\bar{x}$ is a global minimizer of $\operatorname{dist}(c(\cdot), \operatorname{dom} g)$. *In particular, $\bar{x}$ is feasible if the feasible set of (P) is nonempty.*

Proof. Let us consider two cases. If $\{\mu_k\}$ remains bounded away from zero, then Step 3 demands that $\|c(x^k) - z^k\| \to 0$. As we have $\{z^k\} \subseteq \operatorname{dom} g$ from Proposition 4.1, it follows that

$$0 \leq \operatorname{dist}(c(x^k), \operatorname{dom} g) \leq \|c(x^k) - z^k\| \to 0.$$

Owing to $\operatorname{dom} g$ being closed, we obtain $c(\bar{x}) \in \operatorname{dom} g$, proving that $\bar{x}$ is feasible for (P).

Consider now the case $\mu_k \downarrow 0$. Let $\{x^k\}_{k \in K}$ be a subsequence such that $x^k \to_K \bar{x}$. Then (4.1) guarantees

$$\forall x \in \mathbb{R}^n: \quad f(x^k) + g^{\mu_k}(c(x^k) + \mu_k \hat{y}^k) \leq f(x) + g^{\mu_k}(c(x) + \mu_k \hat{y}^k) + \varepsilon_k.$$

Multiplying by μ_k, taking the lower limit as $k \to_K \infty$, and using boundedness of $\{f(x^k)\}_{k \in K}$ and $\{\varepsilon_k\}$ yield

$$\liminf_{k \to_K \infty} \mu_k g^{\mu_k}(c(x^k) + \mu_k \hat{y}^k) \leq \liminf_{k \to_K \infty} \mu_k g^{\mu_k}(c(x) + \mu_k \hat{y}^k)$$

for each $x \in \mathbb{R}^n$. Together with Lemma 2.1, this gives

$$\begin{aligned} \frac{1}{2} \operatorname{dist}^2(c(\bar{x}), \operatorname{dom} g) &= \liminf_{k \to_K \infty} \inf_z \left\{ \mu_k g(z) + \frac{1}{2} \|z - (c(x^k) + \mu_k \hat{y}^k)\|^2 \right\} \\ &= \liminf_{k \to_K \infty} \mu_k g^{\mu_k}(c(x^k) + \mu_k \hat{y}^k) \leq \liminf_{k \to_K \infty} \mu_k g^{\mu_k}(c(x) + \mu_k \hat{y}^k) \\ &= \liminf_{k \to_K \infty} \inf_z \left\{ \mu_k g(z) + \frac{1}{2} \|z - (c(x) + \mu_k \hat{y}^k)\|^2 \right\} = \frac{1}{2} \operatorname{dist}^2(c(x), \operatorname{dom} g) \end{aligned}$$

for all $x \in \mathbb{R}^n$, where we used boundedness of $\{\hat{y}^k\}$ and $\mu_k \downarrow 0$. This shows that $\bar{x}$ globally minimizes $\operatorname{dist}(c(\cdot), \operatorname{dom} g)$. Then, by closedness of $\operatorname{dom} g$, if (P) is feasible, so is $\bar{x}$, concluding the proof. ◻

Notice that, in the proof of Lemma 4.2, the requirement of closed $\operatorname{dom} g$ does not affect the case with $\mu_k \downarrow 0$, and that we did not exploit the actual qualitative requirements regarding the subproblem solver stated in Step 3, apart from the necessity of having a sequence $\{z^k\} \subset \operatorname{dom} g$ at hand. Indeed, approximate global optimality of x^k for the subproblem in the sense of (4.1) is all we needed. Conversely, the upcoming two theorems rely on $z^k \in \operatorname{prox}_{\mu_k g}(c(x^k) + \mu_k \hat{y}^k)$, additionally, but the specification $\|\nabla_x \mathcal{L}^{\mathrm{S}}_{\mu_k}(x^k, z^k, \hat{y}^k)\| \le \varepsilon_k$ is still not required.

If, in addition to the assumptions of Lemma 4.2, the sequence $\{\varepsilon_k\}$ is chosen so that $\varepsilon_k \to 0$, then (primal) accumulation points of the sequence generated by Algorithm 4.1 correspond to global minimizers of (P), see [13, Thm 4.12] for a related result.

Theorem 4.3. *Let $\{(x^k, z^k, y^k)\}$ be a sequence generated by Algorithm 4.1 with $\varepsilon_k \to 0$. Let (4.1) hold, and assume that the feasible set of (P) is nonempty while $\operatorname{dom} g$ is closed. Then $\limsup_{k\to\infty}(f(x^k) + g(z^k)) \le \varphi(x)$ holds for all feasible $x \in \mathbb{R}^n$. Moreover, every accumulation point $\bar{x} \in \mathbb{R}^n$ of $\{x^k\}$ is globally optimal for (P). For any index set $K \subseteq \mathbb{N}$ such that $x^k \to_K \bar{x}$, it is also $z^k \to_K c(\bar{x})$ and $f(x^k) + g(z^k) \to_K \varphi(\bar{x})$.*

Proof. Let $x \in \mathbb{R}^n$ be a fixed feasible point of (P). Then, due to (4.1), Lemma 3.2, and Proposition 4.1, we have

$$\begin{aligned} f(x^k) + g(z^k) - \frac{\mu_k}{2}\|\hat{y}^k\|^2 &\le f(x^k) + g(z^k) + \frac{1}{2\mu_k}\|c(x^k) + \mu_k\hat{y}^k - z^k\|^2 - \frac{\mu_k}{2}\|\hat{y}^k\|^2 \\ &= \mathcal{L}_{\mu_k}(x^k, \hat{y}^k) \le \mathcal{L}_{\mu_k}(x, \hat{y}^k) + \varepsilon_k \le \varphi(x) + \varepsilon_k < \infty \end{aligned}$$

for all $k \in \mathbb{N}$. If $\mu_k \downarrow 0$, then $\mu_k\|\hat{y}^k\|^2 \to 0$ by boundedness of $\{\hat{y}^k\}$. In this case, $\varepsilon_k \to 0$ and $z^k \in \operatorname{dom} g$ imply that $\limsup_{k\to\infty}(f(x^k) + g(z^k)) \le \varphi(x)$.

Let us now focus on the case where $\{\mu_k\}$ is bounded away from zero. This is possible only if $\|c(x^k) - z^k\| \to 0$ by Step 3. Similar as above, we find

$$f(x^k) + g(z^k) + \frac{1}{2\mu_k}\|c(x^k) - z^k\|^2 + \langle \hat{y}^k, c(x^k) - z^k\rangle = \mathcal{L}_{\mu_k}(x^k, \hat{y}^k) \le \varphi(x) + \varepsilon_k.$$

Since $\{\hat{y}^k\}$ is bounded, $\|c(x^k) - z^k\| \to 0$, and $\varepsilon_k \to 0$, we can take the upper limit in the above estimate to find $\limsup_{k\to\infty}(f(x^k) + g(z^k)) \le \varphi(x)$. Finally, let $\bar{x}$ be an accumulation point of $\{x^k\}$ and $K \subseteq \mathbb{N}$ an index set such that $x^k \to_K \bar{x}$. Then, as (P) admits feasible points, $\bar{x}$ is feasible by Lemma 4.2.

Let us show that $\|c(x^k) - z^k\| \to_K 0$ holds. If $\{\mu_k\}$ remains bounded away from zero, this is obvious by Step 3. In the case where $\mu_k \downarrow 0$, we can exploit feasibility of $\bar{x}$, boundedness of $\{\hat{y}^k\}$, and Lemma 2.1 to find

$$\mu_k g(z^k) + \frac{1}{2}\|z^k - c(x^k) - \mu_k\hat{y}^k\|^2 \to_K 0.$$

Boundedness of $\{\hat{y}^k\}$ and $\mu_k \downarrow 0$ allow us to apply Lemma 2.2 which yields $\mu_k g(z^k) \to_K 0$ as well as $\|z^k - c(x^k) - \mu_k\hat{y}^k\| \to_K 0$, and the latter gives $\|z^k - c(x^k)\| \to_K 0$.

Due to $\|c(x^k) - z^k\| \to_K 0$, continuity of c gives $z^k \to_K c(\bar{x})$. Now, lower semicontinuity of g yields

$$\begin{aligned} \varphi(\bar{x}) = f(\bar{x}) + g(c(\bar{x})) &\le \liminf_{k\to_K\infty} f(x^k) + \liminf_{k\to_K\infty} g(z^k) \\ &\le \liminf_{k\to_K\infty}(f(x^k) + g(z^k)) \le \limsup_{k\to_K\infty}(f(x^k) + g(z^k)) \le \varphi(x), \end{aligned}$$

where the last inequality is due to the upper bound obtained previously in the proof. As x is an arbitrary feasible point of (P), we have shown that $\bar{x}$ is globally optimal for (P). Finally, with the particular choice $x = \bar{x}$, the previous inequalities give that $f(x^k) + g(z^k) \to_K \varphi(\bar{x})$, concluding the proof. □

Under an additional assumption on the multiplier estimate $\hat{y}^k$ in Step 2, a stronger result can be proved that concerns the behavior of the iterates for infeasible problems. By resetting the multiplier estimate when signs of infeasibility are detected, the algorithm tends to minimize the objective function subject to minimal constraint violation, see e.g. [11, Thm 5.3] for a related result.

Theorem 4.4. *Let $\{(x^k, z^k, y^k)\}$ be a sequence generated by Algorithm 4.1 with $\varepsilon_k \to 0$. Let (4.1) hold, suppose that $\operatorname{dom} g$ is closed, and that, for all $k \in \mathbb{N}$, $\hat{y}^{k+1} = 0$ if $y^k \notin Y$. Let $\bar{x}$ be an accumulation point of $\{x^k\}$. Then $\bar{x}$ is a global minimizer of $\operatorname{dist}(c(\cdot), \operatorname{dom} g)$ and, for all $(x, z) \in \mathbb{R}^n \times \operatorname{dom} g$ such that $\|c(x) - z\| = \operatorname{dist}(c(\bar{x}), \operatorname{dom} g)$, it holds $\limsup_{k\to\infty}(f(x^k) + g(z^k)) \leq f(x) + g(z)$.*

Proof. If $\operatorname{dist}(c(\bar{x}), \operatorname{dom} g) = 0$, namely $c(\bar{x}) \in \operatorname{dom} g$ by closedness of $\operatorname{dom} g$, then $\bar{x}$ is feasible and the claim follows from Theorem 4.3. So, let us assume that $\operatorname{dist}(c(\bar{x}), \operatorname{dom} g) > 0$. Together with Step 3 and Proposition 4.1, this implies that $\mu_k \downarrow 0$. Since $\bar{x}$ is a global minimizer of $\operatorname{dist}(c(\cdot), \operatorname{dom} g)$ by Lemma 4.2, then

$$\forall k \in \mathbb{N}: \quad \|c(x^k) - z^k\| \geq \operatorname{dist}(c(x^k), \operatorname{dom} g) \geq \operatorname{dist}(c(\bar{x}), \operatorname{dom} g). \tag{4.2}$$

Thus, by the dual update rule at Step 3, boundedness of Y, and $\mu_k \downarrow 0$, for all $k \in \mathbb{N}$ large enough it is $y^k \notin Y$. Therefore, by the estimate choice stated in the premises, it is $\hat{y}^k = 0$ for all large enough $k \in \mathbb{N}$. Then, for all $x \in \mathbb{R}^n$, we have that

$$\begin{aligned} f(x^k) + g(z^k) + \frac{1}{2\mu_k}\|c(x^k) - z^k\|^2 = f(x^k) + g^{\mu_k}(c(x^k)) = \mathcal{L}_{\mu_k}(x^k, 0) \\ \leq \mathcal{L}_{\mu_k}(x, 0) + \varepsilon_k = f(x) + g^{\mu_k}(c(x)) + \varepsilon_k \end{aligned}$$

for all $k \in \mathbb{N}$ large enough. This holds, in particular, for $x \in \mathbb{R}^n$ a global minimizer of $\operatorname{dist}(c(\cdot), \operatorname{dom} g)$, namely such that $\operatorname{dist}(c(x), \operatorname{dom} g) = \operatorname{dist}(c(\bar{x}), \operatorname{dom} g)$. Then there is some $z \in \operatorname{dom} g$ such that $\|c(x) - z\| = \operatorname{dist}(c(\bar{x}), \operatorname{dom} g)$, and (4.2) gives $\|c(x^k) - z^k\| \geq \|c(x) - z\|$. Hence, we find

$$\begin{aligned} f(x^k) + g(z^k) + \frac{1}{2\mu_k}\|c(x^k) - z^k\|^2 \leq f(x) + g^{\mu_k}(c(x)) + \varepsilon_k \\ \leq f(x) + g(z) + \frac{1}{2\mu_k}\|c(x) - z\|^2 + \varepsilon_k \leq f(x) + g(z) + \frac{1}{2\mu_k}\|c(x^k) - z^k\|^2 + \varepsilon_k \end{aligned}$$

for all $k \in \mathbb{N}$ large enough. Subtracting the squared norm term on both sides and taking the upper limit yields the claim since $\varepsilon_k \to 0$. □

In Lemma 4.2 and Theorems 4.3 and 4.4, it has been assumed that the AL subproblem can be solved up to approximate global optimality. This, however, might be a delicate issue whenever f or g are nonconvex or c is difficult enough. In practice, affordable solvers only have local scope and return stationary points as candidate local minimizers. Nevertheless, (primal) accumulation points of a sequence generated by Algorithm 4.1 can be shown to be at least *asymptotically* M-, or AM-, stationary points for (P), see [16, Def. 3.3, Thm 4.1] for a detailed discussion. Notice that the (approximate) global optimality is not relaxed to *local* optimality, but to mere (Υ-)stationarity, while the subsequential g-attentive convergence of certain iterates is required. We refer to [17, Ex. 3.4] for an illustration of the importance of attentive convergence. It should be noted that a mild asymptotic regularity condition is enough to guarantee that AM-stationary point of (P) are indeed M-stationary, see [16, Cor. 3.1] or [17, Cor. 2.7] for related results. Thus, Algorithm 4.1 is likely to compute M-stationary points of (P).

A result analogous to Lemma 4.2 is available in this affordable setting, too, see [16, Prop. 4.5], whereas [16, Prop. 4.3] provides some sufficient conditions for the feasibility of accumulation points.

4.3 Local convergence

In this section, we investigate the behavior of Algorithm 4.1 in the vicinity of stationary points of (P) under various assumptions. Particularly, we are interested in the existence of strict local minimizers of the AL subproblems in a neighborhood of a strict local minimizer to (P) and convergence rates associated with Algorithm 4.1 in such situations.

4.3.1 *Existence of local minimizers*

Let us consider the existence of local minimizers of the AL function $\mathcal{L}_\mu(\cdot, \hat{y})$ for $\mu > 0$ sufficiently small and an approximate multiplier $\hat{y} \in Y$, where $Y \subseteq \mathbb{R}^m$ is a bounded set, see Algorithm 4.1. Note that, by construction, any such local minimizer would be an Υ-stationary point of $\mathcal{L}_\mu(\cdot, \hat{y})$, and Step 3 would be meaningful, see Remark 3.3 as well.

We fix some strict local minimizer $\bar{x} \in \mathbb{R}^n$ of (P) and proceed as suggested in [13, Sec. 7]. For sufficiently small $r > 0$, consider the localized AL subproblem

$$\text{(4.3)} \qquad \operatorname*{minimize}_{x\in\mathbb{R}^n} \quad \mathcal{L}_{\mu_k}(x, \hat{y}^k) \qquad \text{subject to} \quad x \in \mathbb{B}_r(\bar{x}),$$

where $\hat{y}^k \in Y$ is the chosen multiplier estimate and $\mu_k \in (0, \mu_g)$. Clearly, by continuity of the Moreau envelope, see e.g. [55, Thm 1.25], (4.3) possesses a global minimizer $x^k \in \mathbb{R}^n$. Under suitable assumptions it is possible to show $\|x^k - \bar{x}\| < r$ for sufficiently small μ_k, and the localization in (4.3) becomes superfluous. In fact, if $\mu_k \downarrow 0$, we are in position to verify $x^k \to \bar{x}$ as desired.

As we will see in the subsequent lemma, strict local minimality of some feasible point $\bar{x} \in \mathbb{R}^n$ of (P) serves as a sufficient condition for a sequence of asymptotically feasible points to converge to $\bar{x}$, see [13, Cor. 6.2] for a related result under stronger assumptions.

Lemma 4.5. *Let $\bar{x} \in \mathbb{R}^n$ be a strict local minimizer of (P). Then there exists $r > 0$ such that, whenever $\{x^k\} \subseteq \mathbb{B}_r(\bar{x})$ and $\{z^k\} \subseteq \operatorname{dom} g$ are sequences with $\|c(x^k) - z^k\| \to 0$ and $\limsup_{k\to\infty}(f(x^k) + g(z^k)) \le \varphi(\bar{x})$, then $x^k \to \bar{x}$ and $z^k \to c(\bar{x})$.*

Proof. The stated assumptions guarantee the existence of $r > 0$ such that

$$\text{(4.4)} \qquad \forall x \in \mathbb{B}_r(\bar{x}) \setminus \{\bar{x}\}: \quad \varphi(x) > \varphi(\bar{x}).$$

Let us pick arbitrary sequences $\{x^k\}$ and $\{z^k\}$ satisfying the requirements. Suppose now that $x^k \nrightarrow \bar{x}$. Hence, as $\{x^k\}$ belongs to the compact set $\mathbb{B}_r(\bar{x})$, it possesses an accumulation point $\tilde{x} \in \mathbb{B}_r(\bar{x})$ such that $\tilde{x} \neq \bar{x}$. Let $K \subseteq \mathbb{N}$ be an index set such that $x^k \to_K \tilde{x}$. Continuity of c yields $c(x^k) \to_K c(\tilde{x})$, and $\|c(x^k) - z^k\| \to 0$ gives $z^k \to_K c(\tilde{x})$. Furthermore, continuity of f and lower semicontinuity of g can be used to infer

$$\varphi(\bar{x}) \ge \limsup_{k\to\infty}(f(x^k) + g(z^k)) \ge \liminf_{k\to_K\infty}(f(x^k) + g(z^k)) \ge \varphi(\tilde{x}),$$

but this contradicts (4.4). Thus, it must be $x^k \to \bar{x}$. Then, continuity of c yields $c(x^k) \to c(\bar{x})$, and $z^k \to c(\bar{x})$ follows from $\|c(x^k) - z^k\| \to 0$. □

Observe that SOSC from Definition 3.7 provides a sufficient condition for strict local minimality which can be checked in terms of initial problem data, see Proposition 3.8. It is remarkable that, in analogous ways, one can show that [13, Cor. 6.2] (stated in an infinite-dimensional setting) remains true whenever the considered point of interest therein is supposed to be a strict local minimizer. One does not need to assume validity of a second-order sufficient condition for that purpose.

The following is an analog of [13, Lem. 7.1].

Lemma 4.6. *Let $\bar{x} \in \mathbb{R}^n$ be a strict local minimizer of* (P)*. Furthermore, let $Y \subseteq \mathbb{R}^m$ be bounded. Then there is a radius $r > 0$ such that whenever $\{\hat{y}^k\} \subseteq Y$, $\mu_k \downarrow 0$, $\varepsilon_k \to 0$, and, for all $k \in \mathbb{N}$, x^k is an ε_k-minimizer of* (4.3) *in the sense that*

$$\forall x \in \mathbb{B}_r(\bar{x})\colon \quad \mathcal{L}_{\mu_k}(x^k, \hat{y}^k) \leq \mathcal{L}_{\mu_k}(x, \hat{y}^k) + \varepsilon_k, \tag{4.5}$$

then $x^k \to \bar{x}$.

Proof. Let $r > 0$ be as in Lemma 4.5. For large enough $k \in \mathbb{N}$, $\mu_k \in (0, \mu_g)$ holds, and we can fix $z^k \in \operatorname{prox}_{\mu_k g}(c(x^k) + \mu_k \hat{y}^k)$. For any such $k \in \mathbb{N}$, (4.5) and Lemma 3.2 yield

$$\begin{aligned} f(x^k) + g(z^k) &+ \frac{1}{2\mu_k}\|c(x^k) + \mu_k \hat{y}^k - z^k\|^2 - \frac{\mu_k}{2}\|\hat{y}^k\|^2 \\ &= f(x^k) + g^{\mu_k}\left(c(x^k) + \mu_k \hat{y}^k\right) - \frac{\mu_k}{2}\|\hat{y}^k\|^2 \\ &= \mathcal{L}_{\mu_k}(x^k, \hat{y}^k) \leq \mathcal{L}_{\mu_k}(\bar{x}, \hat{y}^k) + \varepsilon_k \leq \varphi(\bar{x}) + \varepsilon_k < \infty. \end{aligned}$$

Multiplying by μ_k, by the boundedness of $\{\hat{y}^k\}$ and $\{f(x^k)\}$, $\mu_k \downarrow 0$, and $\varepsilon_k \to 0$, we obtain

$$\limsup_{k\to\infty} \left(\mu_k g(z^k) + \frac{1}{2}\|z^k - c(x^k) - \mu_k \hat{y}^k\|^2\right) \leq 0.$$

We apply Lemma 2.2 to find $\|z^k - c(x^k) - \mu_k \hat{y}^k\| \to 0$ and, thus, $\|c(x^k) - z^k\| \to 0$. Moreover, the above estimate also guarantees $\limsup_{k\to\infty}(f(x^k) + g(z^k)) \leq \varphi(\bar{x})$, again by boundedness of $\{\hat{y}^k\}$ and $\mu_k \downarrow 0$. Hence, Lemma 4.5 is applicable and yields the desired convergence. □

As a consequence of Lemma 4.6, we find the following result which parallels [13, Thm 7.2].

Theorem 4.7. *Let $\bar{x} \in \mathbb{R}^n$ be a strict local minimizer of* (P)*. Furthermore, let $Y \subseteq \mathbb{R}^m$ be bounded. Then there is a radius $r > 0$ such that, for every $\hat{y} \in Y$ and $\mu \in (0, \mu_g)$, the function $\mathcal{L}_\mu(\cdot, \hat{y})$ admits a local minimizer $x(\mu, \hat{y})$ which lies in $\mathbb{B}_r(\bar{x})$. Moreover, $x(\mu, \hat{y}) \to \bar{x}$ uniformly on Y as $\mu \downarrow 0$.*

Proof. Let $r > 0$ be as in Lemma 4.6. For $\mu \in (0, \mu_g)$, the Moreau envelope g^μ is a continuous function, and this extends to $\mathcal{L}_\mu(\cdot, \hat{y})$ for arbitrary $\hat{y} \in Y$. Hence, this function possesses a global minimizer over $\mathbb{B}_r(\bar{x})$ which we denote by $x(\mu, \hat{y})$. As $\mu \downarrow 0$, we find $x(\mu, \hat{y}) \to \bar{x}$ from Lemma 4.6, and this convergence is uniform for $\hat{y} \in Y$. □

Let us note that Lemma 4.6 and Theorem 4.7 merely assume strict local minimality of the reference point. As mentioned before, this holds true whenever SOSC in valid at the point of interest. Note that SOSC does not demand uniqueness of the underlying Lagrange multiplier.

4.3.2 RATES OF CONVERGENCE

Our rates-of-convergence analysis of Algorithm 4.1 is based on a primal-dual pair $(\bar{x}, \bar{y}) \in \mathbb{R}^n \times \mathbb{R}^m$ which solves the M-stationarity system (3.4) associated with (P) such that an (upper) error bound condition is valid.

For brevity of notation, we will partially make use of the following assumptions.

Assumption 4.8 (Rates of convergence).

(i) Let $\bar{x} \in \mathbb{R}^n$ be an M-stationary point of (P), and let $\bar{y} \in \Lambda(\bar{x})$ be chosen such that there are a constant $\varrho_{\mathrm{u}} > 0$ and a neighborhood U of $(\bar{x}, c(\bar{x}), \bar{y})$ such that the upper error bound condition (3.14) holds for each triplet $(x, z, y) \in U \cap (\mathbb{R}^n \times \operatorname{dom} g \times \mathbb{R}^m)$.

(ii) Let $\{(x^k, z^k, y^k)\}$ be a sequence generated by Algorithm 4.1 with $\varepsilon_k \to 0$.

(iii) The primal-dual sequence $\{(x^k, y^k)\}$ converges to $(\bar{x}, \bar{y})$.

(iv) For each $k \in \mathbb{N}$ large enough, $\hat{y}^k = y^{k-1}$ is valid.

Note that we already know, by Proposition 3.8 and Theorem 4.7, that the AL admits approximate local minimizers and stationary points in a neighborhood of some M-stationary point $\bar{x} \in \mathbb{R}^n$ which satisfies SOSC. We shall now see that, under the error bound conditions from Section 3.3 involving a fixed multiplier $\bar{y} \in \Lambda(\bar{x})$, if the algorithm chooses these local minimizers (or any other points sufficiently close to $\bar{x}$), then we automatically obtain the convergence $(x^k, z^k, y^k) \to (\bar{x}, c(\bar{x}), \bar{y})$. In this case, the sequence $\{y^k\}$ is necessarily bounded, so it is reasonable to assume that the safeguarded multipliers are eventually chosen as $\hat{y}^k = y^{k-1}$.

Let us recall that even validity of the second-order condition (3.11), which is more restrictive than SOSC, may not be sufficient for the error bound condition, see Remark 3.15. However, Section 3.3 provides a number of sufficient conditions which still can be checked in terms of initial problem data, so we will abstain here from postulating any more specific assumptions on the upper error bound. Furthermore, we do not stipulate any lower error bound conditions, deviating from all other related papers, where the lower estimate was never problematic, see Remark 3.18.

The following result, which is motivated by [57, Prop. 4.29], can be considered as (retrospective) justification for Assumption 4.8(iii)–(iv) in the presence of Assumption 4.8(i)–(ii). Besides the error bound condition, a CQ is needed. As we require an M-stationary point of (P), this is not too restrictive.

Proposition 4.9. *Let Assumption 4.8(i)–(ii) hold and suppose that (at least) one of the following conditions is valid:*

(a) $c'(\bar{x})$ possesses full row rank m;

(b) condition (3.6) is valid, $\operatorname{dom} g$ *is closed, and g is continuous relative to its domain.*

Then there exists a radius $r > 0$ such that, if $x^k \in \mathbb{B}_r(\bar{x})$ for all sufficiently large $k \in \mathbb{N}$, then we have the convergences $\Theta(x^k, z^k, y^k) \to 0$ and $(x^k, z^k, y^k) \to (\bar{x}, c(\bar{x}), \bar{y})$ as $k \to \infty$.

Proof. Let $r > 0$ be small enough so that (3.14) holds for all $(x, z, y) \in \mathbb{R}^n \times \operatorname{dom} g \times \mathbb{R}^m$ with $x \in \mathbb{B}_r(\bar{x})$. Assume now that $x^k \in \mathbb{B}_r(\bar{x})$ for all $k \in \mathbb{N}$ sufficiently large. The proof is divided into multiple steps.

We first show that $c(x^k) - z^k \to 0$ as $k \to \infty$. Consider two cases. If $\{\mu_k\}$ remains bounded away from zero, this assertion readily follows from the penalty updating scheme at Step 3. Instead, if $\mu_k \downarrow 0$, then we can argue from Proposition 4.1 that

$$c'(x^k)^\top \left[c(x^k) + \mu_k \hat{y}^k - z^k\right] \to 0 \tag{4.6}$$

as $\mu_k \downarrow 0$, by boundedness of $\{\nabla f(x^k)\}$, $\{\hat{y}^k\}$, and $\{\varepsilon_k\}$. Let us now show $c(x^k) + \mu_k \hat{y}^k - z^k \to 0$, which readily yields $c(x^k) - z^k \to 0$ since $\{\hat{y}^k\}$ is bounded and $\mu_k \downarrow 0$. In case (a) where $c'(\bar{x})$ has full row rank, the matrices $c'(x)c'(x)^\top$ are uniformly invertible on $\mathbb{B}_r(\bar{x})$, potentially after shrinking r, and (4.6) gives $c(x^k) + \mu_k \hat{y}^k - z^k \to 0$. Next, for case (b), assume that (3.6) holds while $\operatorname{dom} g$ is closed and g is continuous on its domain. Note that, for each $k \in \mathbb{N}$, we even have $\mu_k^{-1}(c(x^k) + \mu_k \hat{y}^k - z^k) \in \widehat{\partial} g(z^k)$ by definition of the prox-operator and compatibility of the regular subdifferential with smooth additions, and this also gives $(c(x^k) + \mu_k \hat{y}^k - z^k, -\mu_k) \in \widehat{N}_{\operatorname{epi} g}(z^k, g(z^k))$. Recall that (3.6) is equivalent to the metric regularity of Ξ from (3.5) at $((\bar{x}, g(c(\bar{x}))), (0, 0))$. Lemma 2.5 now yields the existence of $s > 0$ such that, for all sufficiently large $k \in \mathbb{N}$, we have

$$\mathbb{B}_s(0,0) \subseteq \begin{bmatrix} c'(x^k) & 0 \\ 0 & 1 \end{bmatrix} \mathbb{B}_1(0,0) - \left(T_{\operatorname{epi} g}(z^k, g(z^k)) \cap \mathbb{B}_1(0,0)\right) \tag{4.7}$$

as g is continuous on $\operatorname{dom} g$. In order to see this, we need to make sure that $(z^k, g(z^k))$ is sufficiently close to $(c(\bar{x}), g(c(\bar{x})))$ for large enough $k \in \mathbb{N}$, and due to the continuity of g, this boils down to showing that z^k is sufficiently close to $c(\bar{x})$ for large enough $k \in \mathbb{N}$.

Along the tail of the sequence (without relabeling), we have that $\{x^k\}$ is close to $\bar{x}$. For every k, the optimality of z^k in the proximal minimization subproblem reads

$$\forall z \in \mathbb{R}^m: \qquad \mu_k g(z^k) + \frac{1}{2}\|c(x^k) + \mu_k \hat{y}^k - z^k\|^2 \leq \mu_k g(z) + \frac{1}{2}\|c(x^k) + \mu_k \hat{y}^k - z\|^2.$$

Taking the specific choice $z \coloneqq c(\bar{x}) \in \operatorname{dom} g$ and dividing both sides by $\mu_k > 0$ results in

$$g(z^k) + \frac{1}{2\mu_k}\|c(x^k) + \mu_k \hat{y}^k - z^k\|^2 \leq g(c(\bar{x})) + \frac{1}{2\mu_k}\|c(x^k) + \mu_k \hat{y}^k - c(\bar{x})\|^2 < \infty.$$

By invoking the triangle, Cauchy–Schwarz, and Young's inequalities, this implies that

$$\begin{aligned}
\|z^k - c(\bar{x})\|^2 &= \|z^k - [c(x^k) + \mu_k \hat{y}^k] - c(\bar{x}) + [c(x^k) + \mu_k \hat{y}^k]\|^2 \\
&\leq \|z^k - [c(x^k) + \mu_k \hat{y}^k]\|^2 + \|c(\bar{x}) - [c(x^k) + \mu_k \hat{y}^k]\|^2 \\
&\qquad + 2\|z^k - [c(x^k) + \mu_k \hat{y}^k]\|\|c(\bar{x}) - [c(x^k) + \mu_k \hat{y}^k]\| \\
&\leq 2\|z^k - [c(x^k) + \mu_k \hat{y}^k]\|^2 + 2\|c(\bar{x}) - [c(x^k) + \mu_k \hat{y}^k]\|^2 \\
&\leq 4\left[\mu_k g(c(\bar{x})) - \mu_k g(z^k) + \|c(x^k) + \mu_k \hat{y}^k - c(\bar{x})\|^2\right].
\end{aligned}$$

Rearranging gives

$$\mu_k g(z^k) + \frac{1}{4}\|z^k - c(\bar{x})\|^2 \leq \mu_k g(c(\bar{x})) + \|c(x^k) + \mu_k \hat{y}^k - c(\bar{x})\|^2.$$

Since $c(\bar{x}) \in \operatorname{dom} g$, the term $\mu_k g(c(\bar{x}))$ vanishes as $\mu_k \downarrow 0$, and so does $\mu_k \hat{y}^k$. Therefore, possibly shrinking the neighborhood considered around $\bar{x}$, the right-hand side remains bounded by some arbitrarily small $C > 0$ for all large $k \in \mathbb{N}$, by continuous differentiability of c, i.e.,

$$\mu_k g(z^k) + \frac{1}{4}\|z^k - c(\bar{x})\|^2 \leq C$$

holds for all $k \in \mathbb{N}$ large enough. By virtue of the prox-boundedness of g, [18, Lem. 4.1] yields boundedness of $\{z^k\}$ and, thus, of $\{g(z^k)\}$ by continuity of g on its domain which is assumed to be closed (Heine's theorem yields uniform continuity of g on closed, bounded subsets of $\operatorname{dom} g$). As $\mu_k \downarrow 0$ and $C > 0$ can be made arbitrarily small if only $r > 0$ is chosen small enough, it follows that $\{z^k\}$ is arbitrarily close to $c(\bar{x})$ for all large enough $k \in \mathbb{N}$.

Pick $w \in \mathbb{B}_s(0)$ arbitrary. Then, for each sufficiently large $k \in \mathbb{N}$, we can rely on (4.7) to find $(u^k, \alpha_k) \in \mathbb{B}_1(0,0)$ and $(v^k, \beta_k) \in T_{\operatorname{epi} g}(z^k, g(z^k)) \cap \mathbb{B}_1(0,0)$ such that $(w, 0) = (c'(x^k)u^k - v^k, \alpha_k - \beta_k)$. From $(v^k, \beta_k) \in T_{\operatorname{epi} g}(z^k, g(z^k))$, we find $\langle (v^k, \beta_k), (c(x^k) + \mu_k \hat{y}^k - z^k, -\mu_k)\rangle \leq 0$ due to (2.5). Thus

$$\begin{aligned}
\langle w, c(x^k) + \mu_k \hat{y}^k - z^k\rangle &= \langle c'(x^k)u^k - v^k, c(x^k) + \mu_k \hat{y}^k - z^k\rangle \\
&= \langle u^k, c'(x^k)^\top (c(x^k) + \mu_k \hat{y}^k - z^k)\rangle - \mu_k \beta_k - \langle (v^k, \beta_k), (c(x^k) + \mu_k \hat{y}^k - z^k, -\mu_k)\rangle \\
&\geq \langle u^k, c'(x^k)^\top (c(x^k) + \mu_k \hat{y}^k - z^k)\rangle - \mu_k \beta_k \to 0,
\end{aligned}$$

where we used boundedness of $\{u^k\}$ and $\{\beta_k\}$ as well as $\mu_k \downarrow 0$ and (4.6). Testing this expression with $w \coloneqq \pm s(c(x^k) + \mu_k \hat{y}^k - z^k)/\|c(x^k) + \mu_k \hat{y}^k - z^k\|$ gives $c(x^k) + \mu_k \hat{y}^k - z^k \to 0$.

We now demonstrate that $\Theta(x^k, z^k, y^k) \to 0$ as $k \to \infty$. Observe that

$$\nabla_x \mathcal{L}(x^k, y^k) = \nabla_x \mathcal{L}^{\mathrm{S}}(x^k, z^k, y^k) = \nabla_x \mathcal{L}^{\mathrm{S}}_{\mu_k}(x^k, z^k, \hat{y}^k)$$

holds for all $k \in \mathbb{N}$ by construction of the dual update rule in Step 3. Then the first summand in $\Theta(x^k, z^k, y^k)$ satisfies

$$\|\nabla_x \mathcal{L}(x^k, y^k)\| = \|\nabla_x \mathcal{L}^{\mathrm{S}}_{\mu_k}(x^k, z^k, \hat{y}^k)\| \leq \varepsilon_k,$$

which converges to zero by Assumption 4.8(ii). Hence, as $\|c(x^k) - z^k\| \to 0$ was obtained previously, the second term in (3.12) vanishes, too. For the third and last term, it remains to show that $\operatorname{dist}(y^k, \partial g(z^k)) \to 0$. This, however, readily follows from Proposition 4.1.

Finally, recall that $x^k \in \mathbb{B}_r(\bar{x})$ for all $k \in \mathbb{N}$ and that $\Theta(x^k, z^k, y^k) \to 0$. Hence, the convergence $(x^k, z^k, y^k) \to (\bar{x}, c(\bar{x}), \bar{y})$ is an immediate consequence of (3.14). □

Subsequently, we will prove convergence rates for the sequence $\{(x^k, z^k, y^k)\}$ in the presence of Assumption 4.8. Since the distance of (x^k, z^k, y^k) to $(\bar{x}, c(\bar{x}), \bar{y})$ admits an estimate relative to the residual terms $\Theta_k := \Theta(x^k, z^k, y^k)$ by (3.14), we will largely base our analysis on the sequence $\{\Theta_k\}$, and the results on the sequence $\{(x^k, z^k, y^k)\}$ will follow directly. However, this correspondence heavily relies on a *two-sided* error bound, see the proof of Theorem 4.12 below. In stark contrast to Remark 3.18, the following Lemma 4.10 shows that, along a sequence generated by Algorithm 4.1, a lower error bound holds. This exploits the fact that, as a consequence of Proposition 4.1, the distance-to-subdifferential in Θ does not play a role for the error bound *at the iterates*. Therefore, complementing the upper estimate of Assumption 4.8(i), a two-sided error bound becomes *algorithmically* available, enabling the derivation of convergence rates.

Lemma 4.10. *Let $\bar{x} \in \mathbb{R}^n$ be an M-stationary point of (P) and $\bar{y} \in \Lambda(\bar{x})$ be arbitrary. Suppose Assumption 4.8(ii)–(iv) hold. Then there are a constant $\varrho_l > 0$ and a neighborhood U of $(\bar{x}, c(\bar{x}), \bar{y})$ such that, for each triplet $(x^k, z^k, y^k) \in U \cap (\mathbb{R}^n \times \operatorname{dom} g \times \mathbb{R}^m)$, we have*

$$\varrho_l \, \Theta(x^k, z^k, y^k) \leq \|x^k - \bar{x}\| + \|z^k - c(\bar{x})\| + \|y^k - \bar{y}\|. \tag{4.8}$$

Proof. For each triplet $(x^k, z^k, y^k) \in \mathbb{R}^n \times \operatorname{dom} g \times \mathbb{R}^m$, we can exploit (3.4) and the triangle inequality to obtain

$$\begin{aligned} \Theta(x^k, z^k, y^k) &= \|\nabla_x \mathcal{L}(x^k, y^k)\| + \|c(x^k) - z^k\| \\ &\leq \|\nabla_x \mathcal{L}(x^k, y^k) - \nabla_x \mathcal{L}(\bar{x}, \bar{y})\| + \|c(x^k) - c(\bar{x})\| + \|z^k - c(\bar{x})\|, \end{aligned}$$

where the equality is due to Assumption 4.8(ii) and Proposition 4.1, which imply $\operatorname{dist}(y^k, \partial g(z^k)) = 0$ for all $k \in \mathbb{N}$. Then, noting that $\nabla_x \mathcal{L}$ and c are locally Lipschitz continuous, the claim follows. □

Our next result, preparatory for Theorem 4.12 below, has been inspired by [57, Lem. 4.30].

Lemma 4.11. *Let Assumption 4.8 hold and set $\Theta_k := \Theta(x^k, z^k, y^k)$ for each $k \in \mathbb{N}$. Then*

$$(1 - \varrho_{\mathrm{u}} \mu_k) \Theta_k \leq \varepsilon_k + \varrho_{\mathrm{u}} \mu_k \Theta_{k-1}$$

for all $k \in \mathbb{N}$ large enough, where $\varrho_{\mathrm{u}} > 0$ is the constant from (3.14).

Proof. Due to Proposition 4.1, $y^k \in \partial g(z^k)$ holds for all $k \in \mathbb{N}$. Then, by (3.12) and Step 3, we have

$$\Theta_k \leq \varepsilon_k + \|c(x^k) - z^k\| = \varepsilon_k + \mu_k \|y^k - y^{k-1}\| \leq \varepsilon_k + \mu_k \|y^k - \bar{y}\| + \mu_k \|y^{k-1} - \bar{y}\|,$$

where the equality is due to the update rule at Step 3 and Assumption 4.8(iv). By Assumption 4.8(i), since $x^k \to \bar{x}$ and, due to Proposition 4.9, $z^k \to c(\bar{x})$, we find $\|y^k - \bar{y}\| \leq \varrho_{\mathrm{u}} \Theta_k$ for all $k \in \mathbb{N}$ large enough. Hence, $\Theta_k \leq \varepsilon_k + \varrho_{\mathrm{u}} \mu_k (\Theta_k + \Theta_{k-1})$ holds for all $k \in \mathbb{N}$ large enough, and reordering gives the assertion. □

With the above lemma and the two-sided error bound enabled by Assumption 4.8(i) and Lemma 4.10, one can deduce convergence rates for the sequence $\{(x^k, z^k, y^k)\}$, see [57, Thm 4.31] as well. Notice that the condition $\varepsilon_k \in o(\Theta_{k-1})$ can be easily guaranteed in practice. For instance, one could compute the next iterate (x^k, z^k, y^k) with a precision $\varepsilon_k \leq \nu_k \Theta_{k-1}$ where $\{\nu_k\}$ is a given null sequence. It should be mentioned also that the value Θ_k from (3.12) becomes algorithmically available thanks to Proposition 4.1, and can readily be obtained by the dual update rule at Step 3.

Theorem 4.12. *Let Assumption 4.8 hold and assume that $\varepsilon_k \in o(\Theta_{k-1})$, where $\Theta_k := \Theta(x^k, z^k, y^k)$ for each $k \in \mathbb{N}$. Then the following assertions hold.*

(a) For every $q \in (0, 1)$, there exists $\bar{\mu}(q)$ such that, if $\mu_k \leq \bar{\mu}(q)$ for sufficiently large $k \in \mathbb{N}$, then $(x^k, z^k, y^k) \to (\bar{x}, c(\bar{x}), \bar{y})$ Q-linearly with rate q.

(b) If $\mu_k \downarrow 0$, then $(x^k, z^k, y^k) \to (\bar{x}, c(\bar{x}), \bar{y})$ Q-superlinearly.

Proof. Let $k \in \mathbb{N}$ be large enough so that $\hat{y}^k = y^{k-1}$. By Lemma 4.11, if μ_k is small enough so that $1 - \varrho_{\mathrm{u}}\mu_k > 0$, then

$$\frac{\Theta_k}{\Theta_{k-1}} \leq \frac{\varrho_{\mathrm{u}}\mu_k}{1 - \varrho_{\mathrm{u}}\mu_k} + o(1).$$

The desired rates for $\{(x^k, z^k, y^k)\}$ are an easy consequence of the upper and lower estimates in (3.14) and Lemma 4.10, as these give

$$\frac{\|x^k - \bar{x}\| + \|z^k - c(\bar{x})\| + \|y^k - \bar{y}\|}{\|x^{k-1} - \bar{x}\| + \|z^{k-1} - c(\bar{x})\| + \|y^{k-1} - \bar{y}\|} \leq \frac{\varrho_{\mathrm{u}}}{\varrho_{\mathrm{l}}} \frac{\varrho_{\mathrm{u}}\mu_k}{1 - \varrho_{\mathrm{u}}\mu_k} + o(1)$$

for all $k \in \mathbb{N}$ large enough. □

The following result, analogous to [57, Cor. 4.32], establishes the boundedness of $\{\mu_k\}$ away from zero in the case of exact subproblem solutions, thus preventing the fast local convergence of Theorem 4.12(b).

Corollary 4.13. *Let Assumption 4.8 hold and assume that the subproblems occurring at Step 3 of Algorithm 4.1 are solved exactly, i.e., that $\varepsilon_k = 0$ for all $k \in \mathbb{N}$. Then $\{\mu_k\}$ remains bounded away from zero.*

Proof. For each $k \in \mathbb{N}$, we make use of $V_k := \|c(x^k) - z^k\|$ and $\Theta_k := \Theta(x^k, z^k, y^k)$. Let $k \in \mathbb{N}$ be large enough so that $\hat{y}^k = y^{k-1}$. Arguing as in the proof of Lemma 4.11, we have for all $k \in \mathbb{N}$ that $\Theta_k \leq V_k$ since $\varepsilon_k = 0$. Furthermore, using the triangle inequality, the convergences, $x^k \to \bar{x}$, $z^k \to c(\bar{x})$, and Assumption 4.8(iv), we obtain $V_k = \mu_k\|y^k - y^{k-1}\| \leq \varrho_{\mathrm{u}}\mu_k(\Theta_k + \Theta_{k-1})$ from (3.14). Combining these inequalities yields

$$\frac{V_k}{V_{k-1}} \leq \varrho_{\mathrm{u}}\mu_k \left(\frac{\Theta_k}{\Theta_{k-1}} + 1\right).$$

Finally, assuming that $\mu_k \downarrow 0$, we deduce from the proof of Theorem 4.12 that $\Theta_k/\Theta_{k-1} \to 0$, and then $V_k/V_{k-1} \to 0$ follows. Hence, $V_k/V_{k-1} \leq \theta$ for all $k \in \mathbb{N}$ sufficiently large, where $\theta \in (0, 1)$ is a fixed parameter of Algorithm 4.1, so that Step 3 gives a contradiction, thus proving the assertion. □

In summary, local fast convergence of Algorithm 4.1, even for nonconvex functions g, can be obtained in the presence of suitable second-order conditions (one ensuring the existence of minimizers of the subproblems and another one to guarantee validity of an upper error bound) and a first-order CQ which, in principle, gives us the full convergence of the primal-dual sequence.

In comparison with the noteworthy results from [22, 28], these assumptions may seem quite strong. However, let us mention that in the settings discussed in these papers, the (convex) function g under consideration is chosen in such a way that the aforementioned two second-order conditions can already be merged into one, see Remark 3.12. Furthermore, it is likely that the additional postulation of a first-order CQ could be avoided in these papers, too, since g (or at least its derivative) is convex and/or polyhedral *enough* while, for convex functions, the proximal operator is well-behaved. It remains a question for future research whether, for example, a generalized polyhedral structure of g (where its domain and epigraph are unions of finitely many convex polyhedral sets) makes the additional assumption of a first-order CQ superfluous.

5 SOME EXEMPLARY SETTINGS

In light of our theoretical findings for the general problem (P), this section examines two notable illustrative settings: sparsity-promoting and complementarity-constrained optimization.

5.1 SPARSITY-PROMOTING OPTIMIZATION

Here, we take a closer look at the sparsity-promoting optimization problem (3.9) which has been already discussed in Example 3.6.

Let us fix some point $\bar{x} \in \mathbb{R}^n$. For $\bar{y} \in \partial\|\cdot\|_0(c(\bar{x}))$, we make use of

$$I^{00}(\bar{x}, \bar{y}) \coloneqq \{i \in I^0(\bar{x}) \mid \bar{y}_i = 0\}, \quad I^{0\pm}(\bar{x}, \bar{y}) \coloneqq \{i \in I^0(\bar{x}) \mid \bar{y}_i \neq 0\}$$

where $I^0(\bar{x})$ has been defined in Example 3.6. With the definition of $I^\pm(\bar{x})$ therein, one obtains

$$T_{\operatorname{gph} \partial\|\cdot\|_0}(c(\bar{x}), \bar{y}) = \left\{ (v, \eta) \in \mathbb{R}^m \times \mathbb{R}^m \,\middle|\, \begin{array}{ll} \forall i \in I^\pm(\bar{x}): & \eta_i = 0 \\ \forall i \in I^{0\pm}(\bar{x}, \bar{y}): & v_i = 0 \\ \forall i \in I^{00}(\bar{x}, \bar{y}): & v_i \eta_i = 0 \end{array} \right\}. \tag{5.1}$$

This can be used to see that (3.16) reduces to the linear independence of the family $(\nabla c_i(\bar{x}))_{i \in I^0(\bar{x})}$.

For $u \in \mathbb{R}^n \setminus \{0\}$ and $\eta \in D(\partial\|\cdot\|_0)(c(\bar{x}), \bar{y})(c'(\bar{x})u)$, we easily find

$$\langle \eta, c'(\bar{x})u \rangle = \sum_{i \in I^\pm(\bar{x})} \underbrace{\eta_i}_{=0} c_i'(\bar{x})u + \sum_{i \in I^0(\bar{x})} \underbrace{\eta_i c_i'(\bar{x})u}_{=0} = 0,$$

so that (3.22) reduces to

$$\forall u \in \{u' \in \mathbb{R}^n \mid \forall i \in I^{0\pm}(\bar{x}, \bar{y}) \colon\ c_i'(\bar{x})u' = 0\} \setminus \{0\} \colon \quad \nabla^2_{xx}\mathcal{L}(\bar{x}, \bar{y})[u, u] > 0, \tag{5.2}$$

and this corresponds to a classical second-order sufficient condition for the nonlinear program

$$\underset{x}{\text{minimize}} \quad f(x) \quad \text{subject to} \quad c_i(x) = 0 \quad i \in I^{0\pm}(\bar{x}, \bar{y}).$$

From [8, Ex. 6.3], we find

$$C(\bar{x}) = \{u \in \mathbb{R}^n \mid f'(\bar{x})u \leq 0,\ \forall i \in I^0(\bar{x}) \colon\ c_i'(\bar{x})u = 0\}.$$

However, for $u \in \mathbb{R}^n$ satisfying $c_i'(\bar{x})u = 0$ for all $i \in I^0(\bar{x})$, we already have

$$f'(\bar{x})u = -\langle \bar{y}, c'(\bar{x})u \rangle = -\sum_{i \in I^\pm(\bar{x})} \underbrace{\bar{y}_i}_{=0} c_i'(\bar{x})u - \sum_{i \in I^0(\bar{x})} \bar{y}_i \underbrace{c_i'(\bar{x})u}_{=0} = 0,$$

i.e., $u \in C(\bar{x})$ due to $\bar{y} \in \partial\|\cdot\|(c(\bar{x}))$, and this particularly holds for $\bar{y} \in \Lambda(\bar{x})$ which exists whenever $\bar{x}$ is M-stationary. In the latter case, we thus obtain the simplified representation

$$C(\bar{x}) = \{u \in \mathbb{R}^n \mid \forall i \in I^0(\bar{x}) \colon\ c_i'(\bar{x})u = 0\}. \tag{5.3}$$

Noting that $\bar{y}_i = 0$ holds for each $i \in I^{\pm}(\bar{x})$, [8, Ex. 6.3] shows that SOSC is implied by

$$\forall u \in C(\bar{x}) \setminus \{0\},\ \exists y \in \Lambda(\bar{x}): \quad \nabla^2_{xx}\mathcal{L}(\bar{x}, y)[u,u] > 0,$$

while

$$\exists \bar{y} \in \Lambda(\bar{x}),\ \forall u \in C(\bar{x}) \setminus \{0\}: \quad \nabla^2_{xx}\mathcal{L}(\bar{x}, \bar{y})[u,u] > 0$$

is sufficient for (3.11), and these correspond to certain second-order sufficient optimality conditions for the optimization problem

$$\underset{x}{\text{minimize}} \quad f(x) \quad \text{subject to} \quad c_i(x) = 0 \quad i \in I^0(\bar{x}).$$

Clearly, due to (5.3), both conditions are implied by (5.2). The following example shows that (5.2) can, indeed, be stronger than (3.11).

Example 5.1. We consider (3.9) for the functions $f\colon \mathbb{R}^2 \to \mathbb{R}$ and $c\colon \mathbb{R}^2 \to \mathbb{R}^2$ given by

$$f(x) := \frac{1}{2}(x_1 - x_2)^2 + x_1 - x_2, \quad c(x) := \begin{pmatrix} x_1 - x_2 \\ x_1 + x_2 \end{pmatrix},$$

and choose $\bar{x}$ to be the origin in $\mathbb{R}^2$. Note that $I^0(\bar{x}) = \{1,2\}$ and $\Lambda(\bar{x}) = \{(-1,0)\}$, i.e., $\bar{y} := (-1,0)$ is the uniquely determined multiplier in this situation. As the critical cone $C(\bar{x})$ reduces to the origin, (3.11) is trivially satisfied.

Observe that we have

$$\nabla^2_{xx}\mathcal{L}(\bar{x}, \bar{y}) = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix},$$

and for $\bar{u} := (1,1)$ and $\bar{\eta} := (0,0)$, we find $\bar{\eta} \in D(\partial\|\cdot\|_0)(c(\bar{x}), \bar{y})(c'(\bar{x})\bar{u})$ from (5.1). Furthermore, $\nabla^2_{xx}\mathcal{L}(\bar{x}, \bar{y})\bar{u} + c'(\bar{x})^\top\bar{\eta} = 0$ is valid. Hence, (3.18) does not hold, and this also shows that the stronger condition (3.22) fails—the latter being equivalent to (5.2) in the present setting.

5.2 COMPLEMENTARITY-CONSTRAINED OPTIMIZATION

Let $m := 2p$ for some $p \in \mathbb{N}$ and consider the special situation $g := \delta_{C_{\text{cc}}}$ where $C_{\text{cc}} \subseteq \mathbb{R}^{2p}$ is given by

$$C_{\text{cc}} := \{z \in \mathbb{R}^{2p} \mid \forall i \in \{1,\ldots,p\}\colon\ 0 \le z_i \perp z_{p+i} \ge 0\},$$

i.e., C_{cc} is the standard complementarity set. Problem (P), thus, reduces to

$$\text{(MPCC)} \qquad \underset{x}{\text{minimize}} \quad f(x) \quad \text{subject to} \quad c(x) \in C_{\text{cc}},$$

a *mathematical problem with complementarity constraints* (MPCC), see the classical monographs [40, 46]. Note that standard inequality and equality constraints can be added without any difficulty and are omitted here for brevity of presentation.

For a feasible point $\bar{x} \in \mathbb{R}^n$ of (MPCC), we make use of the index sets

$$\begin{aligned} I^{+0}(\bar{x}) &:= \{i \in \{1,\ldots,p\} \mid c_i(\bar{x}) > 0,\ c_{p+1}(\bar{x}) = 0\}, \\ I^{0+}(\bar{x}) &:= \{i \in \{1,\ldots,p\} \mid c_i(\bar{x}) = 0,\ c_{p+i}(\bar{x}) > 0\}, \\ I^{00}(\bar{x}) &:= \{i \in \{1,\ldots,p\} \mid c_i(\bar{x}) = 0,\ c_{p+i}(\bar{x}) = 0\}, \end{aligned}$$

which provide a disjoint partition of $\{1,\ldots,p\}$. As we have

$$\partial g(c(\bar{x})) = \partial^\infty g(c(\bar{x})) = N_{C_{\text{cc}}}(c(\bar{x})) = \left\{ y \in \mathbb{R}^{2p} \,\middle|\, \begin{array}{ll} \forall i \in I^{+0}(\bar{x}): & y_i = 0 \\ \forall i \in I^{0+}(\bar{x}): & y_{p+i} = 0 \\ \forall i \in I^{00}(\bar{x}): & (y_i \le 0 \ \wedge\ y_{p+i} \le 0) \ \vee\ y_i y_{p+i} = 0 \end{array} \right\},$$

we can specify the precise meaning of the CQ (3.6). Note that, as $\delta_{C_{\mathrm{cc}}}$ is continuous on its closed domain C_{cc}, (3.6) can be used in Proposition 4.9. We also note that

$$\widehat{N}_{C_{\mathrm{cc}}}(c(\bar{x})) = \left\{ y \in \mathbb{R}^{2p} \,\middle|\, \begin{array}{ll} \forall i \in I^{+0}(\bar{x}): & y_i = 0 \\ \forall i \in I^{0+}(\bar{x}): & y_{p+i} = 0 \\ \forall i \in I^{00}(\bar{x}): & y_i \le 0,\ y_{p+i} \le 0 \end{array} \right\}.$$

Let us now assume that $\bar{x}$ is an M-stationary point of (MPCC). Some calculations show that the associated critical cone is given by

$$C(\bar{x}) = \left\{ u \in \mathbb{R}^n \,\middle|\, \begin{array}{ll} & f'(\bar{x})u \le 0 \\ \forall i \in I^{+0}(\bar{x}): & c'_{p+i}(\bar{x})u = 0 \\ \forall i \in I^{0+}(\bar{x}): & c'_i(\bar{x})u = 0 \\ \forall i \in I^{00}(\bar{x}): & 0 \le c'_i(\bar{x})u \perp c'_{p+i}(\bar{x})u \ge 0 \end{array} \right\},$$

see [8, Sec. 5.1]. If $\Lambda(\bar{x}) \cap \widehat{N}_{C_{\mathrm{cc}}}(c(\bar{x}))$ is nonempty, i.e., if $\bar{x}$ is so-called strongly stationary, a simplified representation of the critical cone is available which does not involve the $\nabla f(\bar{x})$ anymore but depends on a multiplier $y \in \Lambda(\bar{x}) \cap \widehat{N}_{C_{\mathrm{cc}}}(c(\bar{x}))$ and is given by

$$C(\bar{x}) = \left\{ u \in \mathbb{R}^n \,\middle|\, \begin{array}{ll} & f'(\bar{x})u \le 0 \\ \forall i \in I^{+0}(\bar{x}): & c'_{p+i}(\bar{x})u = 0 \\ \forall i \in I^{0+}(\bar{x}): & c'_i(\bar{x})u = 0 \\ \forall i \in I^{00}_{--}(\bar{x}, y): & c'_i(\bar{x})u = 0,\ c'_{p+i}(\bar{x}) = 0 \\ \forall i \in I^{00}_{-0}(\bar{x}, y): & c'_i(\bar{x})u = 0,\ c'_{p+i}(\bar{x})u \ge 0 \\ \forall i \in I^{00}_{0-}(\bar{x}, y): & c'_i(\bar{x})u \ge 0,\ c'_{p+i}(\bar{x})u = 0 \\ \forall i \in I^{00}_{*}(\bar{x}, y): & 0 \le c'_i(\bar{x})u \perp c'_{p+i}(\bar{x})u \ge 0 \end{array} \right\},$$

see [41, Lem. 4.1]. Here, we used

$$\begin{aligned} I^{00}_{--}(\bar{x}, y) &:= \{ i \in I^{00}(\bar{x}) \mid y_i < 0,\ y_{p+i} < 0 \}, & I^{00}_{-0}(\bar{x}, y) &:= \{ i \in I^{00}(\bar{x}) \mid y_i < 0,\ y_{p+i} = 0 \}, \\ I^{00}_{0-}(\bar{x}, y) &:= \{ i \in I^{00}(\bar{x}) \mid y_i = 0,\ y_{p+i} < 0 \}, & I^{00}_{*}(\bar{x}, y) &:= \{ i \in I^{00}(\bar{x}) \mid y_i = 0,\ y_{p+i} = 0 \}, \end{aligned}$$

which provide a disjoint partition of $I^{00}(\bar{x})$.

Following the arguments provided at the end of [8, Sec. 3.1], we find

$$\Lambda(\bar{x}, u) = \left\{ y \in \Lambda(\bar{x}) \,\middle|\, \begin{array}{ll} \forall i \in I^{00}_{+0}(\bar{x}, u): & y_i = 0 \\ \forall i \in I^{00}_{0+}(\bar{x}, u): & y_{p+i} = 0 \\ \forall i \in I^{00}_{00}(\bar{x}, u): & y_i \le 0,\ y_{p+1} \le 0 \end{array} \right\}.$$

for each $u \in C(\bar{x})$, where we made use of a disjoint partition of $I^{00}(\bar{x})$ given by

$$\begin{aligned} I^{00}_{+0}(\bar{x}, u) &:= \{ i \in I^{00}(\bar{x}) \mid c'_i(\bar{x})u > 0,\ c'_{p+i}(\bar{x})u = 0 \}, \\ I^{00}_{0+}(\bar{x}, u) &:= \{ i \in I^{00}(\bar{x}) \mid c'_i(\bar{x})u = 0,\ c'_{p+i}(\bar{x})u > 0 \}, \\ I^{00}_{00}(\bar{x}, u) &:= \{ i \in I^{00}(\bar{x}) \mid c'_i(\bar{x})u = 0,\ c'_{p+i}(\bar{x})u = 0 \}. \end{aligned}$$

Thus, due to [8, Thm 5.4], SOSC can be stated in the form

$$\forall u \in C(\bar{x}) \setminus \{0\},\ \exists y \in \Lambda(\bar{x}, u): \quad \nabla^2_{xx} \mathcal{L}(\bar{x}, y)[u, u] > 0.$$

As shown in the proof of Corollary 3.9, any multiplier $\bar{y} \in \Lambda(\bar{x})$ suitable to appear the second-order condition (3.11) necessarily belongs to $\bigcap_{u \in C(\bar{x})\setminus\{0\}} \Lambda(\bar{x}, u) \subset \Lambda(\bar{x})$, and for any such multiplier $\bar{y}$, $\mathrm{d}^2\, \delta_{C_{\mathrm{cc}}}(c(\bar{x}), \bar{y})(c'(\bar{x})u)$ vanishes, see [8, Lem. 3.2, Prop. 3.6]. Hence, (3.11) takes the form

$$\exists \bar{y} \in \bigcap_{u \in C(\bar{x})\setminus\{0\}} \Lambda(\bar{x}, u),\ \forall u \in C(\bar{x}) \setminus \{0\} : \quad \nabla^2_{xx}\mathcal{L}(\bar{x}, \bar{y})[u, u] > 0.$$

It follows from [23, Lem. 3.2] that this is a less restrictive assumption than the standard second-order sufficient condition for (MPCC) which takes the form

$$\exists \bar{y} \in \Lambda(\bar{x}) \cap \widehat{N}_{C_{\mathrm{cc}}}(c(\bar{x})),\ \forall u \in C(\bar{x}) \setminus \{0\} : \quad \nabla^2_{xx}\mathcal{L}(\bar{x}, \bar{y})[u, u] > 0$$

and is based on a strongly stationary point. A detailed study on the relationship between SOSC and (3.11) as well as other MPCC-tailored second-order optimality conditions is beyond the scope of this paper, see e.g. [23, 25] for an overview.

The graphical derivative of the limiting normal cone mappings associated with C_{cc} has been computed recently in [9, Sec. 4.4.1], and the obtained formulas can be used to specify the CQs (3.13), (3.16), (3.18), and (3.22) in the recent setting.

6 CONCLUDING REMARKS

The results in this paper could be extended to cover the extra feature in (P) of a geometric *convex* constraint $x \in X$, which was not included here for reasons of exposition. It remains unclear, instead, how to address such additional constraint with nonconvex X, if not reformulating into (P) and accepting $x \in X$ as a soft constraint, see [16, Rem. 5.1].

Another challenging question is whether it is possible to dispose the additional constraint qualification in the nonconvex polyhedral case (i.e., $\operatorname{epi} g$ being the union of finitely many convex polyhedra) in the analysis of Section 4. Such a result would yield convergence rates merely via some second-order sufficient conditions and the (upper) error bound, generalizing [22]. In specific situations, this should be possible even in the nonpolyhedral setting, as [28] has shown for the case of convex linear-quadratic g. As already pointed out in Section 1, however, such a generalization does not seem to be available in nonpolyhedral settings.

Future research may also focus on the relationship between the proximal point algorithm and the augmented Lagrangian method in the fully nonconvex setting, in the vein of [50, 54], and investigate saddle-point properties of the augmented Lagrangian function in primal-dual terms as in [57].

ACKNOWLEDGMENTS

The authors thank the anonymous reviewer whose valuable comments and suggestions helped to strengthen some results, particularly Lemma 4.5, and to improve the presentation of this paper.

REFERENCES

[1] B. Bank, J. Guddat, D. Klatte, B. Kummer, and K. Tammer, *Non-Linear Parametric Optimization*, Birkhäuser, Basel, 1983, doi:10.1007/978-3-0348-6328-5.

[2] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, Springer, New York, 2011, doi:10.1007/978-1-4419-9467-7.

[3] M. Benko, On inner calmness*, generalized calculus, and derivatives of the normal-cone mapping, *Journal of Nonsmooth Analysis and Optimization* 2 (2021), 5881, doi:10.46298/jnsao-2021-5881.

[4] M. Benko, H. Gfrerer, and J. V. Outrata, Calculus for directional limiting normal cones and subdifferentials, *Set-Valued and Variational Analysis* 27 (2019), 713–745, doi:10.1007/s11228-018-0492-5.

[5] M. Benko, H. Gfrerer, J. J. Ye, J. Zhang, and J. Zhou, Second-order optimality conditions for general nonconvex optimization problems and variational analysis of disjunctive systems, *SIAM Journal on Optimization* 33 (2023), 2625–2653, doi:10.1137/22m1484742.

[6] M. Benko and P. Mehlitz, On implicit variables in optimization theory, *Journal of Nonsmooth Analysis and Optimization* 2 (2021), 7215, doi:10.46298/jnsao-2021-7215.

[7] M. Benko and P. Mehlitz, Calmness and calculus: two basic patterns, *Set-Valued and Variational Analysis* 30 (2022), 81–117, doi:10.1007/s11228-021-00589-x.

[8] M. Benko and P. Mehlitz, Why second-order sufficient conditions are, in a way, easy — or — revisiting calculus for second subderivatives, *Journal of Convex Analysis* 30 (2023), 541–589, https://www.heldermann.de/JCA/JCA30/JCA302/jca30031.htm.

[9] M. Benko and P. Mehlitz, On the directional asymptotic approach in optimization theory, *arXiv* (2024), arXiv:2402.16530v1.

[10] D. P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods*, Athena Scientific, 1996.

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

[12] J. Bolte, S. Sabach, and M. Teboulle, Nonconvex Lagrangian-based optimization: monitoring schemes and global convergence, *Mathematics of Operations Research* 43 (2018), 1210–1232, doi:10.1287/moor.2017.0900.

[13] E. Börgens, C. Kanzow, and D. Steck, Local and global analysis of multiplier methods for constrained optimization in Banach spaces, *SIAM Journal on Control and Optimization* 57 (2019), 3694–3722, doi:10.1137/19m1240186.

[14] 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 (2017), 168–193, doi:10.1137/15m1052834.

[15] A. R. Conn, N. I. M. Gould, and P. L. Toint, A globally convergent augmented Lagrangian algorithm for optimization with general constraints and simple bounds, *SIAM Journal on Numerical Analysis* 28 (1991), 545–572, doi:10.1137/0728030.

[16] A. De Marchi, Implicit augmented Lagrangian and generalized optimization, *Journal of Applied and Numerical Optimization* 6 (2024), 291–320, doi:10.23952/jano.6.2024.2.08.

[17] A. De Marchi, X. Jia, C. Kanzow, and P. Mehlitz, Constrained composite optimization and augmented Lagrangian methods, *Mathematical Programming* 201 (2023), 863–896, doi:10.1007/s10107-022-01922-4.

[18] A. De Marchi and A. Themelis, Proximal gradient algorithms under local Lipschitz gradient continuity, *Journal of Optimization Theory and Applications* 194 (2022), 771–794, doi:10.1007/s10957-022-02048-5.

[19] S. Dempe, J. Dutta, and B. S. Mordukhovich, New necessary optimality conditions in optimistic bilevel programming, *Optimization* 56 (2007), 577–604, doi:10.1080/02331930701617551.

[20] N. K. Dhingra, S. Z. Khong, and M. R. Jovanović, The proximal augmented Lagrangian method for nonsmooth composite optimization, *IEEE Transactions on Automatic Control* 64 (2019), 2861–2868, doi:10.1109/tac.2018.2867589.

[21] A. L. Dontchev and R. T. Rockafellar, *Implicit Functions and Solution Mappings*, Springer, Heidelberg, 2014, doi:10.1007/978-0-387-87821-8.

[22] D. Fernández and M. Solodov, Local convergence of exact and inexact augmented Lagrangian methods under the second-order sufficient optimality condition, *SIAM Journal on Optimization* 22 (2012), 384–407, doi:10.1137/10081085x.

[23] H. Gfrerer, Optimality conditions for disjunctive programs based on generalized differentiation with application to mathematical programs with equilibrium constraints, *SIAM Journal on Optimization* 24 (2014), 898–931, doi:10.1137/130914449.

[24] H. Gfrerer and J. V. Outrata, On computation of generalized derivatives of the normal-cone mapping and their applications, *Mathematics of Operations Research* 41 (2016), 1535–1556, doi:10.1287/moor.2016.0789.

[25] L. Guo, G. H. Lin, and J. J. Ye, Second-order optimality conditions for mathematical programs with equilibrium constraints, *Journal of Optimization Theory and Applications* 158 (2013), 33–64, doi:10.1007/s10957-012-0228-x.

[26] N. T. V. Hang, B. S. Mordukhovich, and M. E. Sarabi, Second-order variational analysis in second-order cone programming, *Mathematical Programming* 180 (2020), 75–116, doi:10.1007/s10107-018-1345-6.

[27] N. T. V. Hang, B. S. Mordukhovich, and M. E. Sarabi, Augmented Lagrangian method for second-order cone programs under second-order sufficiency, *Journal of Global Optimization* 82 (2022), 51–81, doi:10.1007/s10898-021-01068-1.

[28] N. T. V. Hang and M. E. Sarabi, Local convergence analysis of augmented Lagrangian methods for piecewise linear-quadratic composite optimization problems, *SIAM Journal on Optimization* 31 (2021), 2665–2694, doi:10.1137/20m1375188.

[29] N. T. V. Hang and M. E. Sarabi, Convergence of augmented Lagrangian methods for composite optimization problems, *arXiv* (2023), arXiv:2307.15627v2.

[30] M. R. Hestenes, Multiplier and gradient methods, *Journal of Optimization Theory and Applications* 4 (1969), 303–320, doi:10.1007/bf00927673.

[31] A. D. Ioffe and J. V. Outrata, On metric and calmness qualification conditions in subdifferential calculus, *Set-Valued Analysis* 16 (2008), 199–227, doi:10.1007/s11228-008-0076-x.

[32] A. F. Izmailov, On the analytical and numerical stability of critical Lagrange multiplier, *Computational Mathematics and Mathematical Physics* 45 (2005), 930–946, https://www.mathnet.ru/php/archive.phtml?wshow=paper&jrnid=zvmmf&paperid=636&option_lang=eng.

[33] A. F. Izmailov, A. S. Kurennoy, and M. V. Solodov, Local convergence of the method of multipliers for variational and optimization problems under the noncriticality assumption, *Computational Optimization and Applications* 60 (2015), 111–140, doi:10.1007/s10589-014-9658-8.

[34] A. F. Izmailov and M. V. Solodov, Stabilized SQP revisited, *Mathematical Programming* 133 (2012), 93–120, doi:10.1007/s10107-010-0413-3.

[35] X. Jia, C. Kanzow, P. Mehlitz, and G. Wachsmuth, An augmented Lagrangian method for optimization problems with structured geometric constraints, *Mathematical Programming* 199 (2023), 1365–1415, doi:10.1007/s10107-022-01870-z.

[36] C. Kanzow and D. Steck, On error bounds and multiplier methods for variational problems in Banach spaces, *SIAM Journal on Control and Optimization* 56 (2018), 1716–1738, doi:10.1137/17m1146518.

[37] C. Kanzow and D. Steck, Improved local convergence results for augmented Lagrangian methods in C^2-reducible constrained optimization, *Mathematical Programming* 177 (2019), 425–438, doi:10.1007/s10107-018-1261-9.

[38] C. Kanzow, D. Steck, and D. Wachsmuth, An augmented Lagrangian method for optimization problems in Banach spaces, *SIAM Journal on Control and Optimization* 56 (2018), 272–291, doi:10.1137/16m1107103.

[39] A. B. Levy, Implicit multifunction theorems for the sensitivity analysis of variational conditions, *Mathematical Programming* 74 (1996), 333–350, doi:10.1007/bf02592203.

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

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

[42] A. Mohammadi, B. S. Mordukhovich, and M. E. Sarabi, Variational analysis of composite models with applications to continuous optimization, *Mathematics of Operations Research* 47 (2022), 397–426, doi:10.1287/moor.2020.1074.

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

[44] B. S. Mordukhovich and M. E. Sarabi, Critical multipliers in variational systems via second-order generalized differentiation, *Mathematical Programming* 169 (2018), 605–648, doi:10.1007/s10107-017-1155-2.

[45] B. S. Mordukhovich and M. E. Sarabi, Criticality of Lagrange multipliers in variational systems, *SIAM Journal on Optimization* 29 (2019), 1524–1557, doi:10.1137/18m1206862.

[46] 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.

[47] M. J. D. Powell, A method for nonlinear constraints in minimization problems, in *Optimization*, Academic Press, 1969, 283–298.

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

[49] R. T. Rockafellar, A dual approach to solving nonlinear programming problems by unconstrained optimization, *Mathematical Programming* 5 (1973), 354–373, doi:10.1007/bf01580138.

[50] R. T. Rockafellar, Augmented Lagrangians and applications of the proximal point algorithm in convex programming, *Mathematics of Operations Research* 1 (1976), 97–116, doi:10.1287/moor.1.2.97.

[51] R. T. Rockafellar, Lagrange multipliers and optimality, *SIAM Review* 35 (1993), 183–238, doi:10.1137/1035044.

[52] R. T. Rockafellar, Extended nonlinear programming, in *Nonlinear Optimization and Related Topics*, G. Di Pillo and F. Giannessi (eds.), volume 36 of Applied Optimization, Kluwer, 2000, 381–399, doi:10.1007/978-1-4757-3226-9_20.

[53] R. T. Rockafellar, Augmented Lagrangians and hidden convexity in sufficient conditions for local optimality, *Mathematical Programming* 198 (2023), 159–194, doi:10.1007/s10107-022-01768-w.

[54] R. T. Rockafellar, Convergence of augmented Lagrangian methods in extensions beyond nonlinear programming, *Mathematical Programming* 199 (2023), 375–420, doi:10.1007/s10107-022-01832-5.

[55] R. T. Rockafellar and R. J. B. Wets, *Variational Analysis*, Springer, 1998, doi:10.1007/978-3-642-02431-3.

[56] S. Sabach and M. Teboulle, Lagrangian methods for composite optimization, in *Processing, Analyzing and Learning of Images, Shapes, and Forms: Part 2*, R. Kimmel and X. C. Tai (eds.), volume 20 of Handbook of Numerical Analysis, Elsevier, 2019, 401–436, doi:10.1016/bs.hna.2019.04.002.

[57] D. Steck, *Lagrange Multiplier Methods for Constrained Optimization and Variational Problems in Banach Spaces*, PhD thesis, University of Würzburg, 2018, https://nbn-resolving.org/urn:nbn:de:bvb:20-opus-174444.

[58] G. Wachsmuth, Strong stationarity for optimization problems with complementarity constraints in absence of polyhedricity, *Set-Valued and Variational Analysis* 25 (2017), 133–175, doi:10.1007/s11228-016-0370-y.

[59] D. W. Walkup and R. J. B. Wets, A Lipschitz characterization of convex polyhedra, *Proceedings of the American Mathematical Society* 23 (1969), 167–173, doi:10.2307/2037511.