IMPLICIT AUGMENTED LAGRANGIAN AND GENERALIZED OPTIMIZATION

ALBERTO DE MARCHI

Institute of Applied Mathematics and Scientific Computing, Department of Aerospace Engineering,
University of the Bundeswehr Munich, 85577 Neubiberg, Germany

Abstract. Generalized nonlinear programming is considered without any convexity assumption, capturing a variety of problems that include nonsmooth objectives, combinatorial structures, and set-membership nonlinear constraints. We extend the augmented Lagrangian framework to this broad problem class, preserving an implicit formulation and introducing auxiliary variables merely as a formal device. This, however, gives rise to a generalized augmented Lagrangian function that lacks regularity. Based on parametric optimization, we develop a tailored stationarity concept to better qualify the iterates, generated as approximate solutions to a sequence of subproblems. Using this variational characterization and the lifted representation, asymptotic properties and convergence guarantees are established for a safeguarded augmented Lagrangian scheme. Numerical examples showcase the modelling versatility gained by dropping convexity assumptions and the practical benefits of the advocated implicit approach.

Keywords. Generalized augmented Lagrangians; Generalized nonlinear programming; Multiplier methods; Nonsmooth nonconvex optimization; Proximal algorithms.

2020 Mathematics Subject Classification. 65K10, 90C26, 65K05.

1. INTRODUCTION

Optimization problems often bring plenty of mathematical structure to work with. In this paper we investigate the setting of finite-dimensional generalized programming without any convexity, namely optimization problems of the form

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad \varphi(x) := f(x) + g(c(x)), \tag{P}$$

where x is the decision variable, f and c are smooth mappings and g is merely lower semicontinuous. Problem data 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. The composition of a nonsmooth term g with a smooth mapping c results in great modeling flexibility and encompasses a broad spectrum of problems [43]; both regularizers and constraints can be encoded, as well as combinations thereof [8, 10, 14]. Nonsmooth composite problems like (P) have been extensively studied in the past decades, but most often with convex (or even polyhedral) nonsmooth term g; see [22, 35, 48]. Prominent special cases of (P) are nonlinear [4] and disjunctive [2] programming, structured optimization [39, 47], and constrained structured

---

E-mail address: alberto.demarchi@unibw.de

optimization [17]. The abstract problem (P) is also referred to as "extended nonlinear program-ming" in [43], promoting better representation of structures found in applications by featuring a composite term, beyond the conventional format of nonlinear programming.

In general, handling the composition $g \circ c$ can be numerically difficult, especially without any convexity assumptions. Rather than focusing on conditions to catalogue cases that may arise, here we intend to address (P) by decoupling g and c, with a generally applicable approach, based on lifting and marginalization. Introducing an auxiliary variable $z \in \mathbb{R}^m$, (P) can be equivalently rewritten in the form

$$\underset{x\in\mathbb{R}^n,\, z\in\mathbb{R}^m}{\text{minimize}} \quad f(x) + g(z) \qquad \text{subject to} \quad c(x) - z = 0, \tag{P_S}$$

which has a simpler, separable objective function but is subject to some (explicit) constraints. It has been observed in [17, Lemma 3.1] that incorporating auxiliary variables in this manner does not affect minimizers and stationary points when compared to the original (P) — some-what remarkably in light of [3]. A fundamental technique for solving *constrained* optimization problems such as (P_S) is the augmented Lagrangian (AL) framework [4, 6, 42], which can also effortlessly handle nonsmoothness [17, 18, 21, 23].

The following blanket assumptions are considered throughout, without further mention. Tech-nical definitions are given in Section 3.1.

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 continuously differentiable;
 (ii) $g \colon \mathbb{R}^m \to \overline{\mathbb{R}}$ is proper, lower semicontinuous, and prox-bounded;
(iii) $\inf \varphi \in \mathbb{R}$.

Assumption 1.1(iii) is relevant to the well-posedness of the problem, as the cost function $\varphi := f + g \circ c$ has nonempty domain and is bounded from below. Prox-boundedness in As-sumption 1.1(ii) guarantees that the proximal mapping of g is well-defined, for some suitable stepsizes. Moreover, we proceed under the working assumption that (only) the following com-putational oracles are available or simple to evaluate:

  • given $x \in \mathbb{R}^n$, cost function value $f(x) \in \mathbb{R}$ and gradient $\nabla f(x) \in \mathbb{R}^n$;
  • given $v \in \mathbb{R}^m$ and $\gamma \in (0, \gamma_g)$, with γ_g being the prox-boundedness threshold of g, proxi-mal point $z \in \text{prox}_{\gamma g}(v)$ and cost function value $g(z) \in \mathbb{R}$ therein;
  • given $x \in \mathbb{R}^n$ and $v \in \mathbb{R}^m$, constraint function value $c(x) \in \mathbb{R}^m$ and transposed Jacobian-vector product $c'(x)^\top v \in \mathbb{R}^n$, where $c' \colon \mathbb{R}^n \to \mathbb{R}^{m \times n}$ denotes the Jacobian of c.

Relying only on these primitives, the numerical method presented in the following is first-order and matrix-free by construction; as such, it involves only simple operations and has low memory footprint.

The remainder of this paper is structured as follows. Related works and our contributions are discussed in Section 2, followed by notation and preliminary material in Section 3. The augmented Lagrangian scheme is detailed and characterized in Section 4, before developing and analyzing a solver for the inner AL problems in Section 5. Finally, illustrative numerical examples are reported in Section 6, with some concluding remarks in Section 7.

2. RELATED WORK AND CONTRIBUTIONS

The template (P) covers many classical problem classes. These include structured two-term problems $f + g$ [39] and three-term problems $f + g + h$, as well as settings with the nonsmooth term composed with a linear operator, with two-term $g + h \circ L$, three-term problems $f + g + h \circ L$ [34], or with linear constraints as for the alternating direction method of multipliers [8]. Most of these problem classes have been extensively studied in the convex setting, while nonconvexity has been mostly considered for structured two-term problems $f + g$ only [20, 47].

Since their inception [30, 40, 41], AL schemes have been deeply and extensively investigated [4, 6, 11, 24, 25], also in infinite dimensions [32]. In the convex setting, it was recognized that the method of multipliers [30, 40] corresponds to solving an associated dual problem via the proximal point algorithm (PPA) [41, 42], with duality theory playing a key role in this parallel. Then, sufficient optimality conditions and local duality led to analogous results in nonlinear programming by Bertsekas [4]. Pivoting on some kind of local duality, in turn enabled by some local convexity, the AL scheme can be traced back to an application of the PPA, thereby establishing convergence.

This pattern was recently followed by Rockafellar [44, 45] to extend these ideas to the much broader setting of (P) with g convex. In contrast with this line of work, which goes back at least to [43, Section 2], we are interested here in investigating AL methods for generalized nonlinear programming without any convexity assumptions. With g potentially nonconvex, the abstract setting (P) not only covers a variety of problems, but it also provides a more versatile and expressive modeling paradigm, with no need for "suitable reformulations". Dropping the convexity of g comes at a price, though, as it appears more difficult to leverage the perspective of PPA. Nonetheless, the shifted-penalty approach underpinning the seminal method of multipliers still applies, and can be exploited as such. Yet difficulties arise as the nonconvexity of g translates into potential set-valuedness of its proximal mapping, which in turn leads to a lack of regularity. Here, we seek a better understanding of the consequences on the analysis and design of numerical methods for (P), concerning theoretical properties, convergence guarantees, and practical aspects.

In conjunction with AL schemes, non-Lipschitzian or merely lower semicontinuous cost functions have been recently considered in [7, 9, 17, 18, 23, 28]. The issue of attentive convergence, defined in Section 3.1 and seemingly necessary to perform limiting analysis, was overlooked in [9] and brought to attention in [17]. Since it is difficult to attest a priori, as a sufficient to guarantee attentive convergence one may additionally assume g continuous relative to its domain. While much weaker than convexity, this assumption is satisfied by many practical problems, as pointed out in [19].

Another issue is that of well-definedness of the AL algorithm. In fact, the AL subproblems could be ill-posed, as there is no a priori guarantee that the AL function is bounded from below, due to the constraints relaxation [6]. As we are concerned with the role of g, in this work, we neglect this possibility and assume that all subproblems are well-posed. Several approaches for moving beyond this assumption may be applicable, although possibly imposing other requirements. Among these are localized minimizations by exploiting non-relaxable constraints [1] or a trust-region included in the subproblems, as well as different AL schemes [23, 25].

Constrained structured optimization [17] entails the minimization of the sum $f(x) + g(x)$ over $x \in \mathbb{R}^n$ subject to set-membership constraints $c(x) \in D$, where c is a smooth mapping and D a

nonempty closed set. As such, it is more general than nonlinear and disjunctive programming [2, 4], but covered by the template (P). Conversely, possibly introducing some auxiliary variables, (P) can always be reformulated as a constrained structured program. Thus, the problem class investigated in [17] is as general as (P) under Assumption 1.1, but our solution strategy here is essentially different.

The core idea of this work traces back to the author's doctoral dissertation [14, Chapter 1], whose further developments took a different spirit and eventually became [17]. Therein auxiliary variables are introduced as in (P_S), making the constraints *explicit* and then directly applying a traditional AL scheme. To the contrary, here we maintain the *implicit* approach advocated for in [14], offering a rigorous convergence analysis. In particular, we use the lifted reformulation (P_S) only as a theoretical device to develop our strategy for solving (P) and investigate its properties via sharpened variational tools. Marginalizing the AL function of (P_S) over the auxiliary variable z yields the generalized (or, implicit) AL function for (P) [14, 18, 21, 45]. However, by the nonconvexity of g, the generalized AL function fails to be continuously differentiable (or prox-friendly) in general. For overcoming this lack of regularity,

- we devise the concept of Υ-stationarity, which allows us to characterize precisely the sought (approximate) solutions to the AL subproblems,

thereby better capturing the intended philosophy. Rooted in parametric optimization, Υ-stationarity is strong enough for the outer AL scheme but still computable and easy to handle numerically. Indeed, we show this concept provides an intermediate qualification between M-stationarity for the implicit and explicit formulations. With the concept of Υ-stationarity at hand, we focus on solving the AL subproblems, namely minimizing the AL function, which constitutes the major computational task and a key step in AL algorithms. We explore the possibility to directly solve the AL subproblems, without auxiliary variables. In particular,

- we design and analyze a subsolver that exploits the proximal oracle of g for minimizing the generalized AL function and, by generating approximate Υ-stationary points, complies with the requirements of the AL scheme.

The rationale behind the subsolver is to include auxiiary variables only as algorithmic by-products, and not as decision variables. These are then adopted within an iterative scheme to generate updates that yield sufficient improvement and to terminate with a suitable certificate of (approximate) stationarity. Overall, the advantages of weaker assumptions and stronger stationarity concepts would not be attractive without convergence results. Therefore,

- we investigate the asymptotic properties of a safeguarded implicit AL scheme for (P) and establish (global) convergence guarantees on par with those in classical nonlinear programming.

Moreover, numerical examples illustrate how following an implicit approach can result in enhanced practical performance, without hindering modeling flexibility. Not only does sticking to a compact problem formulation avoid increasing the problem size, but it also leads to better use of the available computational oracles and tends to avoid spurious (local) minima and stationary points.

3. NOTATION AND FUNDAMENTALS

In this section, we comment on notation, preliminary definitions and useful results.

3.1. **Preliminaries.** With $\mathbb{R}$ and $\overline{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$ we denote the real and extended-real line, respectively. The vectors in $\mathbb{R}^n$ with all elements equal to 0 or 1 are denoted as 0_n and 1_n; whenever n is clear from context we simply write 0 and 1, respectively. With $\mathbb{B}_r(x)$ we indicate the closed ball centered at x with radius r. The *effective domain* of an extended-real-valued function $h \colon \mathbb{R}^n \to \overline{\mathbb{R}}$ is denoted by $\operatorname{dom} h := \{x \in \mathbb{R}^n \,|\, h(x) < \infty\}$. We say that h is *proper* if $\operatorname{dom} h \neq \emptyset$ and *lower semicontinuous* (lsc) if $h(\bar{x}) \leq \liminf_{x \to \bar{x}} h(x)$ for all $\bar{x} \in \mathbb{R}^n$. For some constant $\tau \in \mathbb{R}$, $\operatorname{lev}_{\leq \tau} h := \{x \in \mathbb{R}^n \,|\, h(x) \leq \tau\}$ denotes the τ-*sublevel set* associated with h. Function h is (lower) *level-bounded* if for every $\tau \in \mathbb{R}$ its τ-sublevel set is bounded. Then, we introduce the following notion of *uniform level boundedness*, which plays an important role in parametric minimization [46, Definition 1.16]. A function $f \colon \mathbb{R}^n \times \mathbb{R}^m \to \overline{\mathbb{R}}$ with values $f(x, u)$ is *level-bounded in x locally uniformly in u* if for each $\alpha \in \mathbb{R}$ and $\bar{u} \in \mathbb{R}^m$ there exists $\varepsilon > 0$ such that the set $\{(x, u) \,|\, f(x, u) \leq \alpha, \|u - \bar{u}\| \leq \varepsilon\}$ is bounded in $\mathbb{R}^n \times \mathbb{R}^m$. We use the notation $F \colon \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ to indicate a point-to-set function $F \colon \mathbb{R}^n \to \mathscr{P}(\mathbb{R}^m)$, where $\mathscr{P}(\mathbb{R}^m)$ is the power set of $\mathbb{R}^m$ (the set of all subsets of $\mathbb{R}^m$).

Given a proper and lsc function $h \colon \mathbb{R}^n \to \overline{\mathbb{R}}$ and a point $\bar{x}$ with $h(\bar{x})$ finite, we may avoid to assume h continuous and instead appeal to *h-attentive* convergence of a sequence $\{x^k\}$, denote as $x^k \xrightarrow{h} \bar{x}$ and given by $x^k \to \bar{x}$ with $h(x^k) \to h(\bar{x})$. Following [46, Definition 8.3], we denote by $\hat{\partial} h \colon \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ the *regular subdifferential* of h, where

$$v \in \hat{\partial} h(\bar{x}) \quad :\Leftrightarrow \quad \liminf_{\substack{x \to \bar{x} \\ x \neq \bar{x}}} \frac{h(x) - h(\bar{x}) - \langle v, x - \bar{x} \rangle}{\|x - \bar{x}\|} \geq 0.$$

The (limiting) *subdifferential* of h is $\partial h \colon \mathbb{R}^n \rightrightarrows \mathbb{R}^n$, where $v \in \partial h(\bar{x})$ if and only if $\bar{x} \in \operatorname{dom} h$ and there exist sequences $\{x^k\}$ and $\{v^k\}$ such that $x^k \xrightarrow{h} \bar{x}$ and $v^k \in \hat{\partial} h(x^k)$ with $v^k \to v$. The subdifferential of h at $\bar{x}$ satisfies $\partial(h + h_0)(\bar{x}) = \partial h(\bar{x}) + \nabla h_0(\bar{x})$ for any $h_0 \colon \mathbb{R}^n \to \overline{\mathbb{R}}$ continuously differentiable around $\bar{x}$ [46, Exercise 8.8].

Given a parameter value $\gamma > 0$, the *proximal* mapping $\operatorname{prox}_{\gamma h}$ is defined by

$$\operatorname{prox}_{\gamma h}(x) := \arg\min_z \left\{ h(z) + \frac{1}{2\gamma} \|z - x\|^2 \right\}. \tag{3.1}$$

We say that h is *prox-bounded* if it is proper and $h + \|\cdot\|^2/(2\gamma)$ is bounded below on $\mathbb{R}^n$ for some $\gamma > 0$ [46, Definition 1.23]. The supremum of all such γ is the threshold γ_h of prox-boundedness for h. In particular, if h is bounded below by an affine function, then $\gamma_h = \infty$. When h is lsc, for any $\gamma \in (0, \gamma_h)$ the proximal mapping $\operatorname{prox}_{\gamma h}$ is locally bounded, nonempty- and compact-valued [46, Theorem 1.25]. The value function of the minimization problem defining the proximal mapping is the *Moreau envelope* with parameter $\gamma > 0$, denoted $h^\gamma \colon \mathbb{R}^n \to \mathbb{R}$, namely

$$h^\gamma(x) := \inf_z \left\{ h(z) + \frac{1}{2\gamma} \|z - x\|^2 \right\}.$$

Some tools of variational analysis will be exploited in order to describe the geometry of some nonempty, closed, but not necessarily convex, set $D \subset \mathbb{R}^m$. The *projection* mapping $\Pi_D \colon \mathbb{R}^m \rightrightarrows \mathbb{R}^m$ and the *distance* function $\operatorname{dist}_D \colon \mathbb{R}^m \to \mathbb{R}$ are defined by

$$\Pi_D(v) := \arg\min_{z \in D} \|z - v\| \qquad \text{and} \qquad \operatorname{dist}_D(v) \equiv \operatorname{dist}(v, D) := \inf_{z \in D} \|z - v\|.$$

The former is a set-valued mapping whenever D is nonconvex, whereas the latter is always single-valued. The *indicator* of D is denoted $\delta_D \colon \mathbb{R}^m \to \overline{\mathbb{R}}$ and defined by $\delta_D(v) = 0$ if $v \in D$,

and $\delta_D(v) = \infty$ otherwise. If D is nonempty and closed, then δ_D is proper and lsc. The proximal mapping of δ_D is the projection Π_D; thus, Π_D is locally bounded. Given $z \in D$, the *limiting normal cone* to D at z is the closed cone

$$N_D^{\lim}(z) := \limsup_{v \to z} \text{cone}\,(v - \Pi_D(v)).$$

The limiting normal cone is robust (or stable) in the sense that $N_D^{\lim}(z) = \limsup_{v \to z} N_D^{\lim}(v)$ holds for all $z \in D$. For any proper and lsc function $h \colon \mathbb{R}^n \to \overline{\mathbb{R}}$ and point $\bar{x} \in \text{dom}\,h$, we have

$$\partial h(\bar{x}) = \left\{ v \in \mathbb{R}^n \,|\, (v, -1) \in N_{\text{epi}\,h}^{\lim}(\bar{x}, h(\bar{x})) \right\}, \tag{3.2}$$

where $\text{epi}\,h := \{(x, \alpha) \in \mathbb{R}^n \times \mathbb{R} \,|\, h(x) \le \alpha\}$ denotes the *epigraph* of h [46, Theorem 8.9].

3.2. Stationarity Concepts.

We now define some basic concepts and discuss optimality and stationarity conditions for (P). As the cost function φ is possibly extended-real-valued, the concept of feasibility for a point refers to the domain of this cost function. A point $x^* \in \mathbb{R}^n$ is called *feasible* for (P) if $x^* \in \text{dom}\,\varphi$, that is, if $c(x^*) \in \text{dom}\,g$. With (P), we seek points where the minimum of φ is attained, which are necessarily feasible. Notions of optimality can have global or local scope. A feasible point $x^* \in \mathbb{R}^n$ is said *locally optimal*, or called a *local minimizer*, for (P) if there exists $\delta > 0$ such that $\varphi(x^*) \le \varphi(x)$ holds for all feasible $x \in \mathbb{B}_\delta(x^*)$. Additionally, if the inequality holds for all feasible $x \in \mathbb{R}^n$, then x^* is said *(globally) optimal* for (P).

Focusing on the development of an "affordable" algorithm, our method is designed to seek *candidate* minimizers, guided by necessary optimality conditions. Working without any convexity assumptions, a plausible stationarity concept is that based on limiting subdifferentials; cf. [36, Section 3] and [37, Theorem 5.48]. With respect to the minimization of a proper lsc function $h \colon \mathbb{R}^n \to \overline{\mathbb{R}}$, we say that $x^* \in \text{dom}\,h$ is *M-stationary* if $0 \in \partial h(x^*)$, which constitutes a necessary condition for the optimality of x^* [46, Theorem 10.1]. When addressing (P), the composite structure gives rise to first-order optimality conditions expressed in Lagrangian terms [44, 45], as it intuitively follows from (P$_\text{S}$). Although closely related to M-stationarity, as discussed in [18, Section 4.1], here we prefer the nomenclature *KKT*-stationarity to emphasize the Lagrangian structure.

Definition 3.1 (KKT-stationarity). Relative to (P), a point $x^* \in \mathbb{R}^n$ is called *KKT-stationary* if there exists a multiplier $y^* \in \mathbb{R}^m$ such that

$$0 = \nabla f(x^*) + c'(x^*)^\top y^*, \tag{3.3a}$$

$$y^* \in \partial g(c(x^*)). \tag{3.3b}$$

Notice that (3.3b) implicitly requires the feasibility of x^*, namely $c(x^*) \in \text{dom}\,g$, for otherwise the subdifferential $\partial g(c(x^*))$ is empty. Moreover, note that Definition 3.1 coincides with the celebrated Karush-Kuhn-Tucker (KKT) conditions in nonlinear programming if g in (P) is the indicator of a nonempty closed convex set [4, 17, 45].

Although potentially ambiguous in the nonconvex setting, one could refer to the Lagrange multiplier y as to the *dual* variable and to (x, y) as to a *primal-dual* pair. In fact, a sort of local duality emerged while investigating convergence of AL methods in nonconvex settings, as revealed by Bertsekas [4] and recently by Rockafellar in a broader context [44, 45]—but this seems premature for (P) under Assumption 1.1.

Given some tolerance $\varepsilon \geq 0$, an approximate M-stationarity concept for the minimization of h refers to $\mathrm{dist}(0, \partial h(x^*)) \leq \varepsilon$. We may also consider relaxed conditions for an approximate counterpart of Definition 3.1.

Definition 3.2 (Approximate KKT-stationarity). Let $\varepsilon \geq 0$ be given. Relative to (P), a point $x^* \in \mathbb{R}^n$ is called *ε-KKT-stationary* if there exist $z^* \in \mathrm{dom}\, g$ and a multiplier $y^* \in \mathbb{R}^m$ such that

$$\|\nabla f(x^*) + c'(x^*)^\top y^*\| \leq \varepsilon, \tag{3.4a}$$

$$\mathrm{dist}(y^*, \partial g(z^*)) \leq \varepsilon, \tag{3.4b}$$

$$\|c(x^*) - z^*\| \leq \varepsilon. \tag{3.4c}$$

It should be noticed that ε-KKT-stationarity with $\varepsilon = 0$ recovers the notion of KKT-stationarity, analogously to M-stationarity. This fact follows from $\partial g(z)$ being closed at any point $z \in \mathrm{dom}\, g$ [46, Theorem 8.6], which implies that, for any $v \in \mathbb{R}^m$, $\mathrm{dist}(v, \partial g(z)) = 0$ is equivalent to the inclusion $v \in \partial g(z)$.

The approximation allowed by Definition 3.2 entails not only stationarity (3.4a) and complementarity (3.4b), but affects also feasibility: owing to (3.4c) it may be that $c(x^*) \notin \mathrm{dom}\, g$. Encompassing this inexactness is particularly important in practice as it might be difficult to deal with the composition $g \circ c$ as soon as c is a nonlinear mapping. On the contrary, when c is the identity mapping we apply a narrower definition compared to Definition 3.2, as we deem feasibility of an approximate M-stationary point a significant yet attainable property in this case. Therefore, relative to (P) with $c := \mathrm{Id}$, a point $x^* \in \mathbb{R}^n$ is called *ε-M-stationary* if $\mathrm{dist}(-\nabla f(x^*), \partial g(x^*)) \leq \varepsilon$. Along with feasibility, namely $x^* \in \mathrm{dom}\, g$, this demands a tighter approximation than (3.4). As it suffices to invoke the proximal mapping of g to reconcile the "constraint" $x \in \mathrm{dom}\, g$, many feasible methods exist for such two-term problems $f + g$; see e.g. [10, 20, 39, 47] and references therein.

We add to our theoretical toolbox also an asymptotic counterpart of KKT-stationarity. The following closely patterns the *approximate* KKT conditions of [6, Definition 3.1], and so does the meaning of primal iterates x^k and Lagrange multipliers y^k.

Definition 3.3 (AKKT-stationarity). Relative to (P), a feasible point $x^* \in \mathbb{R}^n$ is called *asymptotically KKT-stationary* if there exist sequences $\{x^k\}_{k \in \mathbb{N}} \subseteq \mathbb{R}^n$ and $\{y^k\}_{k \in \mathbb{N}}, \{z^k\}_{k \in \mathbb{N}} \subseteq \mathbb{R}^m$ such that $z^k \xrightarrow{g} c(x^*)$ and

$$\nabla f(x^k) + c'(x^k)^\top y^k \to 0, \tag{3.5a}$$

$$y^k \in \partial g(z^k), \tag{3.5b}$$

$$c(x)^k - z^k \to 0. \tag{3.5c}$$

Notice that Definition 3.3 does not require the sequence $\{y^k\}_{k \in \mathbb{N}}$ to converge. A measure of (primal) infeasibility is suggested by (3.5c), where z^k is introduced to account in (3.5b) for the fact that the condition $c(x^k) \in \mathrm{dom}\, g$ can be violated along the iterates, though it should hold asymptotically. Similarly, with (3.5a) one can monitor the inexactness in the stationarity condition, or *dual* infeasibility. Finally, the attentive convergence $z^k \xrightarrow{g} c(x^*)$ plays a role when taking the limit in (3.5), since we aim to recover the limiting subdifferential for KKT-stationarity.

A local minimizer for (P) is KKT-stationary under validity of a suitable qualification condition, which, by non-Lipschitzness of g, will depend on function g as well, see [26]. However, we

can show that each local minimizer of (P) is always AKKT-stationary, regardless of additional regularity properties. Related results are [27, Theorem 2.1], [33, Theorem 6.2], [36, Section 5.1], and [17, Proposition 2.5].

Proposition 3.1. *Let $x^* \in \mathbb{R}^n$ be a local minimizer for* (P). *Then, x^* is AKKT-stationary.*

In order to guarantee that local minimizers for (P) are not only AKKT- but also KKT-stationary, the presence of a qualification condition is necessary. Analogously to [33, Definition 6.8] and [17, Definition 2.6], the following regularity notion generalizes the (comparatively weak) constraint qualification from [36, Section 3.2] to the non-Lipschitzian setting of (P).

Definition 3.4 (AKKT-regularity). Let $\bar{x} \in \mathbb{R}^n$ be a feasible point for (P) and define the set-valued mapping $\mathscr{M} : \mathbb{R}^n \times \mathbb{R}^m \rightrightarrows \mathbb{R}^n$ by $\mathscr{M}(x,z) := c'(x)^\top \partial g(z)$. Then, $\bar{x}$ is called *asymptotically KKT-regular* for (P) if

$$\limsup_{\substack{x \to \bar{x} \\ z \to c(\bar{x})}} \mathscr{M}(x,z) \subset \mathscr{M}(\bar{x}, c(\bar{x})).$$

Sufficient conditions for the validity of the more general qualification condition from Definition 3.4 can be captured following the arguments of [33]. A direct consequence of Proposition 3.1 is the next result, analogous in spirit to [33, Proposition 6.9] and [17, Corollary 2.7].

Corollary 3.1. *Let $x^* \in \mathbb{R}^n$ be an AKKT-regular AKKT-stationary point for* (P). *Then, x^* is KKT-stationary for* (P). *In particular, each AKKT-regular local minimizer for* (P) *is KKT-stationary.*

4. AUGMENTED LAGRANGIAN METHOD

Algorithms exhibiting a nested structure naturally arise in the AL framework, which builds upon a sequence of subproblems. We develop our implicit approach and derive the generalized AL function in Section 4.1. Then, in Section 4.2 we discuss stationarity notions relevant for the AL subproblem, in order to precisely qualify a subproblem "solution", and thus the output expected from a subsolver. The following Section 4.3 contains a detailed statement of our algorithmic framework, whose convergence analysis is presented in Section 4.4. Suitable termination criteria are discussed in Section 4.5. The subsequent Section 5 is dedicated to the numerical solution of the AL subproblems.

4.1. **Implicit Augmented Lagrangian.** Introducing an auxiliary variable $z \in \mathbb{R}^m$, (P) can be equivalently rewritten in the form (P_S), which has a structured objective function and some (explicit) equality constraints. Following [7, 17, 18, 28], for some penalty parameter $\mu > 0$, let us define the AL function $\mathscr{L}_\mu^S : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^m \to \overline{\mathbb{R}}$ associated to (P_S) by

$$\mathscr{L}_\mu^S(x,z,y) := f(x) + g(z) + \langle y, c(x) - z \rangle + \frac{1}{2\mu} \|c(x) - z\|^2$$

$$= f(x) + g(z) + \frac{1}{2\mu} \|c(x) + \mu y - z\|^2 - \frac{\mu}{2} \|y\|^2. \qquad (4.1)$$

Within AL schemes [4, 6, 11], it is required to minimize (at least, up to approximate stationarity) the AL function. Entailing the sum of a smooth and a prox-friendly function, the minimization of $\mathscr{L}_\mu^S$ can be tackled by exploiting its structure and the computational oracles via, e.g., proximal-gradient methods [10, 39, 47]. Owing to the convenient structure of $\mathscr{L}_\mu^S$, the

method developed in [17] relies on this "explicit" approach. Here, instead of performing a joint minimization of $\mathscr{L}_\mu^S(\cdot,\cdot,y)$ over both (x,z), we exploit the structure arising from the original problem (P), formally eliminating the auxiliary variable z, on the vein of the proximal AL approach [14, 15, 21, 29]. Given some $\mu > 0$ and $y \in \mathbb{R}^m$, we define the oracle $\mathbf{O}_\mu(\cdot,y)$ as the mapping associated to the pointwise minimization of $\mathscr{L}_\mu^S(\cdot,\cdot,y)$ with respect to z. Owing to the definitions in (3.1) and (4.1), the set-valued mapping $\mathbf{O}_\mu(\cdot,y) \colon \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ is given by

$$\mathbf{O}_\mu(x,y) := \underset{z\in\mathbb{R}^m}{\arg\min}\,\mathscr{L}_\mu^S(x,z,y) = \underset{z\in\mathbb{R}^m}{\arg\min}\left\{g(z)+\frac{1}{2\mu}\|c(x)+\mu y - z\|^2\right\}$$
$$= \mathrm{prox}_{\mu g}\,(c(x)+\mu y). \tag{4.2}$$

Injecting back into $\mathscr{L}_\mu^S(x,\cdot,y)$ any arbitrary element of $\mathbf{O}_\mu(x,y)$, namely evaluating the AL on the set corresponding to the minimization over z, we obtain the (single-valued) AL function $\mathscr{L}_\mu \colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ associated to (P):

$$\mathscr{L}_\mu(x,y) := \inf_{z\in\mathbb{R}^m}\mathscr{L}_\mu^S(x,z,y) = \mathscr{L}_\mu^S(x,z,y)\Big|_{z\in\mathbf{O}_\mu(x,y)}$$
$$= f(x) + g^\mu(c(x)+\mu y) - \frac{\mu}{2}\|y\|^2. \tag{4.3}$$

We highlight that the Moreau envelope $g^\mu \colon \mathbb{R}^m \to \mathbb{R}$ of g is real-valued and strictly continuous [38], [46, Example 10.32], but not continuously differentiable in general, as the proximal mapping of g is possibly set-valued. Regarded as the AL for (P), $\mathscr{L}_\mu$ was called *generalized* AL function in [45] (for g convex). Compared to $\mathscr{L}_\mu^S$ with the *explicit* auxiliary variable z for the lifted formulation (P$_S$), we may also refer to $\mathscr{L}_\mu$ as to the *implicit* AL for (P).

The algorithm we are going to present in Section 4.3 fits in the AL framework and is a natural adaptation thereof [4, 6, 17, 45]. The iterative scheme generates a sequence of iterates x^k and Lagrange multipliers y^k according to

$$x^k \leftsquigarrow \underset{x\in\mathbb{R}^n}{\arg\min}\,\mathscr{L}_{\mu_k}(x,\hat{y}^k), \qquad y^k \leftsquigarrow \hat{y}^k + \frac{1}{\mu_k}\partial_y\mathscr{L}_{\mu_k}(x^k,\hat{y}^k), \tag{4.4}$$

with respect to suitable sequences of parameter values $\mu_k > 0$ and multiplier estimates $\hat{y}^k \in \mathbb{R}^m$. Here we intend to give the reader only a prospect about the algorithm before looking into the details; by using the symbol "$\leftsquigarrow$" in (4.4) we emphasize that the notation should not be taken too seriously. Most especially, for the primal update in (4.4) we do not require a *global*, nor an *exact* local minimizer of $\mathscr{L}_{\mu_k}(\cdot,\hat{y}^k)$, but only an approximate stationary point thereof (in the sense described in Section 4.2). However, it should be noticed that, by considering the implicit AL function, each iterate x^k is associated to some $z^k \in \mathbf{O}_{\mu_k}(x^k,\hat{y}^k)$, by (4.3). In other words, the lifted AL subproblem

$$\underset{x\in\mathbb{R}^n, z\in\mathbb{R}^m}{\text{minimize}}\ \mathscr{L}_{\mu_k}^S(x,z,\hat{y}^k)$$

is solved exactly and to global optimality with respect to z, and up to approximate stationarity relatively to x. Regarding the multiplier, the update rule in (4.4) reduces to the traditional one when (P) specializes to nonlinear programming [4, 6], but if g is nonconvex it may be necessary to select specific elements from the relevant set-valued subdifferentials. This potential choice is also very much related to the stationarity concept introduced in Section 4.2.

At the k-th iteration of the AL scheme, a subproblem in the form

$$\underset{x\in\mathbb{R}^n}{\text{minimize}}\ \mathscr{L}_{\mu_k}(x,\hat{y}^k),\tag{4.5}$$

with some given $\mu_k > 0$ and $\hat{y}^k \in \mathbb{R}^m$, is to be approximately solved. Being g potentially nonconvex, the AL may lack regularity. In the following Section 4.2 we introduce a stationary concept weaker than M-stationarity for (4.5) but sharp enough to guarantee that the iterates generated by the AL scheme detailed in Section 4.3 exhibit opportune asymptotic behavior and properties, as demonstrated later in Section 4.4.

4.2. **Stationarity for Parametric Optimization.** In view of the marginalization step to derive $\mathscr{L}_\mu$ in (4.3), we now study the AL subproblem (4.5) with the lenses of parametric optimization [46, Sections 1.F, 10.C]. Let us consider an objective $p\colon \mathbb{R}^n \to \mathbb{R}$ and an oracle $\mathbf{O}\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ given by

$$p(x) := \inf_{z\in\mathbb{R}^m} P(x,z) \qquad\text{and}\qquad \mathbf{O}(x) := \underset{z\in\mathbb{R}^m}{\arg\min}\, P(x,z)\tag{4.6}$$

for a proper, lsc function $P\colon \mathbb{R}^n \times \mathbb{R}^m \to \overline{\mathbb{R}}$. We are interested in minimizing p, but first in devising suitable stationarity concepts to qualify a "solution". This setting is apparently mirroring subproblem (4.5), with the explicit and implicit AL functions $\mathscr{L}^{\mathrm{S}}$ and $\mathscr{L}$ playing the role of P and p, respectively. Moreover, it highlights our intention to keep the solution process implicit, effectively hiding the auxiliary variable z, by utilizing the oracle $\mathbf{O}$.

Approximate M-stationarity of a point $\bar{x} \in \mathrm{dom}\, p$ for some given $\varepsilon \geq 0$ amounts to requiring that $\mathrm{dist}(0, \partial p(\bar{x})) \leq \varepsilon$. However, because of the parametric nature of p, the subdifferential ∂p is not a simple object in general. Let us consider some results on subdifferentiation in parametric optimization. Recalling from Section 3.1 that the notion of uniform level-boundedness corresponds to a parametric extension of level-boundedness [46, Definition 1.16], we suppose that $P(x,z)$ is level-bounded in z locally uniformly in x. Then, from [46, Theorem 10.13] we have for every $\bar{x} \in \mathrm{dom}\, p$ the inclusion

$$\partial p(\bar{x}) \subseteq \Upsilon(\bar{x}), \qquad\text{where}\qquad \Upsilon(\bar{x}) := \bigcup_{\bar{z}\in\mathbf{O}(\bar{x})} \{v\,|\,(v,0) \in \partial P(\bar{x},\bar{z})\},\tag{4.7}$$

which implies the inequality

$$\mathrm{dist}(0, \partial p(\bar{x})) \geq \mathrm{dist}(0, \Upsilon(\bar{x})).\tag{4.8}$$

This offers a lower bound on the M-stationarity measure, and it does not provide in general an upper bound suitable for a termination condition. Regardless, in the parametric optimization setting of (4.6), we may resort to the following stationarity concept, here denominated Υ-*stationarity*.

Definition 4.1 (Υ-stationarity). Let $\varepsilon \geq 0$ be fixed and consider a proper, lsc function $P\colon \mathbb{R}^n \times \mathbb{R}^m \to \overline{\mathbb{R}}$ with $P(x,z)$ level-bounded in z locally uniformly in x. Define $p\colon \mathbb{R}^n \to \overline{\mathbb{R}}$ and $\Upsilon\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ as in (4.6) and (4.7), respectively. Then, relatively to p, a point $x^* \in \mathrm{dom}\, p$ is called ε-Υ-*stationary* if $\mathrm{dist}(0, \Upsilon(x^*)) \leq \varepsilon$. In the case $\varepsilon = 0$, such point x^* is said Υ-*stationary*.

Notice that the inclusion $x^* \in \mathrm{dom}\, p$ is implicitly required to have $\Upsilon(x^*)$ nonempty. Furthermore, in the exact case, namely $\varepsilon = 0$, the requirement for Υ-stationarity reduces to $0 \in \Upsilon(x^*)$, by closedness of Υ [46, Theorem 10.13].

From the definition of Υ in (4.7), it appears that an ε-Υ-stationary point $x^* \in \mathbb{R}^n$ can be associated to a (possibly non-unique) *certificate* $z^* \in \mathbf{O}(x^*)$ and a *pair* (x^*, z^*) that satisfy

$$\text{dist}(0, \Upsilon(x^*)) = \min_{v \in \Upsilon(x^*)} \|v\| = \min_{\substack{v \in \mathbb{R}^n \\ z \in \mathbf{O}(x^*)}} \{\|v\| \mid (v, 0) \in \partial P(x^*, z)\}$$

$$\leq \min_{v \in \mathbb{R}^n} \{\|v\| \mid (v, 0) \in \partial P(x^*, z^*)\} \leq \varepsilon. \qquad (4.9)$$

If any such pair (x^*, z^*) is found, the upper bound in (4.9) certificates the ε-Υ-stationarity of x^*.

The following result establishes that (approximate) M-stationarity provides a stronger qualification than (approximate) Υ-stationarity, meaning that the former is a sufficient condition, not necessary in general, for the latter. The sought implication readily follows from the upper bound in (4.8) on the Υ-stationarity measure.

Proposition 4.1. *Let $\varepsilon \geq 0$ be fixed. Then, relative to the minimization of p, any ε-M-stationary point is also ε-Υ-stationary.*

The converse implication does not hold in general, but it does under some (possibly local) convexity assumptions; see, e.g., [46, Theorem 2.26]. Similar considerations are valid with regard to the following result, which shows that (approximate) Υ-stationarity relative to p is a stronger qualification than (approximate) M-stationarity relative to P.

Proposition 4.2. *Let $\varepsilon \geq 0$ be fixed and (x^*, z^*) be any ε-Υ-stationary pair associated to p. Then, relative to the minimization of P, (x^*, z^*) is ε-M-stationary.*

Proof. By ε-Υ-stationarity relative to p, the pair (x^*, z^*) necessarily satisfies $x^* \in \text{dom}\, p$ and $z^* \in \mathbf{O}(x^*)$. Then, exploiting the upper bound in (4.9), we can argue that

$$\text{dist}(0, \partial P(x^*, z^*)) = \min_{\substack{v \in \mathbb{R}^n \\ u \in \mathbb{R}^m}} \{\|(v, u)\| \mid (v, u) \in \partial P(x^*, z^*)\}$$

$$\leq \min_{v \in \mathbb{R}^n} \{\|v\| \mid (v, 0) \in \partial P(x^*, z^*)\} \leq \varepsilon,$$

where the first inequality is due to the restriction $u := 0$. Showing that (x^*, z^*) is ε-M-stationary for P, since it satisfies $\text{dist}(0, \partial P(x^*, z^*)) \leq \varepsilon$, the assertion is proved. $\qquad\square$

Notice, however, that in general the statement of Proposition 4.2 is not valid for all $z^* \in \mathbf{O}(x^*)$, even if x^* is ε-Υ-stationary for p. In fact, some pairs (x^*, z^*) may fail to satisfy (4.9), and those are not necessarily ε-M-stationary for P.

At this point it appears clear that, in relation to (4.1) and (4.3), the oracle $\mathbf{O}$ refers to $\mathbf{O}_\mu(\cdot, \hat{y})$, for some given $\mu > 0$ and $\hat{y} \in \mathbb{R}^m$. Moreover, we shall focus our attention on the special case of P having the structure

$$P(x, z) := F(x, z) + g(z), \qquad F(x, z) := f(x) + \frac{1}{2\mu}\|c(x) + \mu\hat{y} - z\|^2, \qquad (4.10)$$

where $F: \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is continuously differentiable.

4.3. **Algorithm.** This section details an AL method for the solution of generalized programs of the form (P). In this work we focus on a safeguarded AL scheme inspired by [6, Algorithm 4.1] and investigate its convergence properties. Compared to the classical AL or multiplier penalty approach for the solution of nonlinear programs [4, 11], this variant uses a safeguarded update rule for the Lagrange multipliers and has stronger global convergence properties [32]. Although

we restrict our analysis to this specific algorithm, analogous results can be obtained for others with minor changes, e.g., [23, Algorithm 1].

The overall method is stated in Algorithm 4.1. Although a starting point is not explicitly required, the subproblems at Step 4 should be solved warm-starting from the current primal estimate x^{k-1}, thus exploiting initial guesses. The AL subproblem appears at Step 4, involving the implicit AL function with current penalty parameter μ_k and estimate $\hat{y}^k$. Crucially, Step 4 requires to compute an approximate Υ-stationary *pair* (x^k, z^k), not merely a point x^k, for $\mathscr{L}_{\mu_k}(\cdot, \hat{y}^k)$. As it may be employed to verify approximate stationarity, it seems likely that a certificate z^k for x^k can be obtained with no additional effort — this is the case for the subsolver developed in Section 5.

Algorithm 4.1: Safeguarded implicit AL method for (P)

Data: $\varepsilon > 0$
Result: ε-KKT-stationary point x^*
1 Select $\mu_0 > 0$, $\theta, \kappa \in (0, 1)$, $Y \subset \mathbb{R}^m$ nonempty bounded
2 **for** $k = 0, 1, 2 \ldots$ **do**
3 $\quad$ Select $\hat{y}^k \in Y$ and $\varepsilon_k \geq 0$
4 $\quad$ Compute an ε_k-Υ-stationary pair (x^k, z^k) for $\mathscr{L}_{\mu_k}(\cdot, \hat{y}^k)$
5 $\quad$ Set $y^k \leftarrow \hat{y}^k + \mu_k^{-1}[c(x^k) - z^k]$ and $V^k \leftarrow \|c(x^k) - z^k\|$
6 $\quad$ **if** $\max\{\varepsilon_k, V^k\} \leq \varepsilon$ **then return** $x^* \leftarrow x^k$
7 $\quad$ **if** $k = 0$ **or** $V^k \leq \theta V^{k-1}$ **then** $\mu_{k+1} \leftarrow \mu_k$, **else** $\mu_{k+1} \leftarrow \kappa \mu_k$

The safeguarded Lagrange multiplier estimate $\hat{y}^k$ is drawn from a bounded set $Y \subset \mathbb{R}^m$ at Step 3. Although not necessary, the choice of $\hat{y}^k$ should also depend on the current estimate y^{k-1}. Moreover, the choice of Y can take advantage of *a priori* knowledge of ∂g and its structure, in order to generate better Lagrange multiplier estimates. In practice, it is advisable to choose the safeguarded estimate $\hat{y}^k$ as the projection of the Lagrange multiplier y^{k-1} onto Y, thus effectively adopting the classical approach as long as y^{k-1} remains within Y; we refer to [6, Section 4.1] and [32, Section 3] for a more articulated discussion.

Inspired by the classical first-order Lagrange multiplier estimate, the update rule at Step 5 is designed around (3.5a) and the identity

$$\nabla_x \mathscr{L}_\mu^{\mathrm{S}}(x, z, \hat{y}) = \nabla f(x) + c'(x)^\top \left[\hat{y} + \frac{c(x) - z}{\mu} \right].$$

Given any (x^k, z^k) obtained at Step 4, the specific choice of y^k leads by construction to

$$\|\nabla f(x^k) + c'(x^k)^\top y^k\| \leq \varepsilon_k,$$

effectively connecting inner and outer loops, and providing an upper bound on the dual infeasibility at (x^k, y^k). Step 6 encapsulates suitable termination conditions for returning an approximate KKT-stationary point; these are discussed in Section 4.5. The monotonicity test at Step 7 is adopted to monitor primal infeasibility along the iterates and update the penalty parameter accordingly. Aimed at driving $V^k := \|c(x^k) - z^k\|$ to zero, the penalty parameter μ_k is reduced at Step 7 in case of insufficient decrease.

4.4. **Convergence Analysis.** Throughout our convergence analysis, we assume that Algorithm 4.1 is well-defined, thus requiring that each subproblem at Step 4 admits an approximate stationary point. Moreover, the following statements assume the existence of some accumulation point x^* for a sequence $\{x^k\}_{k \in \mathbb{N}}$ generated by Algorithm 4.1. In general, coercivity or (level) boundedness arguments should be adopted to verify these preconditions.

Due to their practical importance, we focus on *local* (or affordable) solvers, which return stationary points as candidate local minimizers for the subproblems at Step 4; cf. [6, Chapter 6] and [14, 17]. Instead, we do not investigate the case where the subproblems are solved to *global* optimality, whose analysis would pattern the classical results in [6, Chapter 5] and [32]. Since the analysis in this section patterns standard arguments, detailed proofs are moved to Appendix A.

As it is often the case for penalty-type methods in the nonconvex setting, Algorithm 4.1 may generate accumulation points that are infeasible for (P). The following result gives conditions that guarantee feasibility of limit points; cf. [5, Problem 4.12], [17, Proposition 3.2].

Proposition 4.3. *Let Assumption 1.1 hold and consider a sequence $\{x^k\}_{k \in \mathbb{N}}$ of iterates generated by Algorithm 4.1 with $\varepsilon = 0$. Then, each accumulation point x^* of $\{x^k\}_{k \in \mathbb{N}}$ is feasible for* (P) *if (at least) one of the following conditions holds:*

(i) $x^k \xrightarrow{\varphi}_K x^*$ for some subsequence $\{x^k\}_{k \in K}$;
(ii) there exists some $B \in \mathbb{R}$ such that $\mathscr{L}_{\mu_k}(x^k, \hat{y}^k) \leq B$ for all $k \in \mathbb{N}$;
(iii) $\{\mu_k\}_{k \in \mathbb{N}}$ is bounded away from zero and $\mathrm{dom}\, g$ is closed;
(iv) $\{\mu_k\}_{k \in \mathbb{N}}$ is bounded away from zero and $\mathscr{L}_{\mu_k}(x^k, \hat{y}^k) \leq \mathscr{L}_{\mu_k}(x^{k-1}, \hat{y}^k)$ for all $k \in \mathbb{N}$.

Considering the particular case of g being the indicator of some set D with nonempty interior, Proposition 4.4 shows that the Lagrange multipliers vanish for constraints that are (strictly) inactive in the limit. This indicates that the update rule at Step 5 of Algorithm 4.1 is plausible with regard to their expected behavior as shifts [6, Section 4].

Proposition 4.4. *Suppose $g \equiv \delta_D$ for some nonempty and closed set $D \subseteq \mathbb{R}^m$. Let Assumption 1.1 hold and consider a sequence $\{x^k\}_{k \in \mathbb{N}}$ of iterates generated by Algorithm 4.1 with $\varepsilon = 0$. Let x^* be an accumulation point of $\{x^k\}_{k \in \mathbb{N}}$ and $\{x^k\}_{k \in K}$ a subsequence such that $x^k \to_K x^*$. If $c(x^*) \in \mathrm{int}\, D$, then $y^k = 0$ for all $k \in K$ large enough.*

Notice that the assertion can be easily refined by exploiting the separable structure of D, if any. For instance, if D is a hyperbox, every step in the proof can be applied componentwise, recovering [6, Theorem 4.1].

The following convergence result provides fundamental theoretical support to Algorithm 4.1. It shows that, under subsequential attentive convergence, accumulation points are AKKT-stationary for (P). For any accumulation point x^* and for any subsequence $\{x^k\}_{k \in K}$ such that $x^k \to_K x^*$, Proposition 4.3(i) assures the feasibility of x^* under the assumption that $x^k \xrightarrow{\varphi}_K x^*$. However, the φ-attentive convergence also entails $g(c^k) \to_K g(c(x^*))$ as $x^k \to_K x^*$, which is trivially satisfied, e.g., if (x^* is feasible and) g is continuous relative to $\mathrm{dom}\, g$ [19]. We refer to [17, Example 3.4] for an illustration of the importance of φ-attentive convergence.

Theorem 4.1. *Let Assumption 1.1 hold and consider a sequence $\{x^k\}_{k \in \mathbb{N}}$ of iterates generated by Algorithm 4.1 with $\varepsilon = 0$ and $\varepsilon_k \to 0$. Let $x^* \in \mathbb{R}^n$ be an accumulation point of $\{x^k\}_{k \in \mathbb{N}}$ and $\{x^k\}_{k \in K}$ a subsequence such that $x^k \xrightarrow{\varphi}_K x^*$. Then, x^* is an AKKT-stationary point for* (P).

Constrained optimization algorithms aim at finding feasible points and minimizing the objective function subject to constraints. Employing affordable local optimization techniques, one cannot expect to find feasible points nor global minimizers of any infeasibility measure. Nevertheless, the next result proves that Algorithm 4.1 with bounded $\{\varepsilon_k\}_{k\in\mathbb{N}}$ finds at least stationary points of an infeasibility measure. Notice that this property does not require $\varepsilon_k \to 0$, but only boundedness; cf. [6, Theorem 6.3].

Proposition 4.5. *Let Assumption 1.1 hold and suppose* $\operatorname{dom} g$ *is closed. Consider a sequence* $\{x^k\}_{k\in\mathbb{N}}$ *of iterates generated by Algorithm 4.1 with* $\varepsilon = 0$ *and* $\{\varepsilon_k\}_{k\in\mathbb{N}}$ *bounded. Let* $x^* \in \mathbb{R}^n$ *be an accumulation point of* $\{x^k\}_{k\in\mathbb{N}}$ *and* $\{x^k\}_{k\in K}$ *a subsequence such that* $x^k \to_K x^*$. *Then,* x^* *is an M-stationary point of*

$$\underset{x\in\mathbb{R}^n}{\operatorname{minimize}}\ \operatorname{dist}^2(c(x),\operatorname{dom} g).$$

4.5. **Termination Criteria.** Step 4 of Algorithm 4.1 involves the minimization of the implicit AL function, defined in (4.3). Then, the multiplier update at Step 5 allows one to draw conclusions with respect to the original problem (P), as shown by Theorem 4.1. Moreover, recalling the AKKT-stationarity conditions (3.5), in view of Step 3 one may construct a sequence $\{\varepsilon_k\}_{k\in\mathbb{N}} \subset \mathbb{R}_{++}$ such that $\varepsilon_k \to 0$. Then, given some user-specified tolerances $\varepsilon^{\mathrm{dual}}, \varepsilon^{\mathrm{prim}} > 0$, it is reasonable to declare successful convergence when the conditions

$$\varepsilon_k \le \varepsilon^{\mathrm{dual}} \qquad \text{and} \qquad V^k := \|c(x^k) - z^k\| \le \varepsilon^{\mathrm{prim}}$$

are satisfied. Theorem 4.1 demonstrates that these termination criteria (the latter, in particular) are satisfied in finitely many iterations if any subsequence of $\{x^k\}_{k\in K}$ accumulates at a feasible point $x^* \in \mathbb{R}^n$. As this might not be the case, a mechanism for (local) infeasibility detection is needed, and usually included in practical implementations.

Given some tolerances, Algorithm 4.1 can be equipped with relaxed conditions on decrease requirements at Step 7 and optimality at Step 4 [5]. At Step 3 the inner tolerance ε_k can stay bounded away from zero, as long as $\varepsilon_k \le \varepsilon^{\mathrm{dual}}$ for large $k \in \mathbb{N}$. Similarly, the condition at Step 7 can be relaxed by adding the (inclusive) possibility that $V^k \le \varepsilon^{\mathrm{prim}}$, thus avoiding unnecessarily strong penalization.

5. INNER PROBLEM AND SOLVER

In this section we elaborate upon the minimization of the implicit AL function, defined in (4.3), focusing on the numerical solution of subproblems (4.5) for some fixed $\mu_k > 0$ and $\hat{y}^k \in \mathbb{R}^m$. Instead of lifting and jointly minimizing the explicit AL function $\mathscr{L}^{\mathrm{S}}_{\mu_k}(\cdot,\cdot,\hat{y}^k)$ in (4.1) with respect to both primal and auxiliary variables, we take a different approach, aimed at minimizing $\mathscr{L}_{\mu_k}(\cdot,\hat{y}^k)$ directly, in order to solve inner problems without auxiliary variables. To do so, we juggle the implicit and explicit formulations, linked by the oracle, as in (4.6).

In Section 5.1, an algorithm fo the minimization of p is discussed, that exploits the available oracles and the structure of P in (4.10). Reminiscent of [14, Section 1.5.4], the scheme compensates for the lack of regularity by pivoting around the oracle's optimality. In fact, a surrogate gradient of p is built based on that of F, the smooth part of P, evaluated at the oracle's response. Then, it turns out that any off-the-shelf algorithm for unconstrained smooth optimization can be directly adopted for the task of minimizing the implicit AL function, in a transparent way.

Throughout this section, let $\mu_k > 0$ and $\hat{y}^k \in \mathbb{R}^m$ be some fixed given parameters. Denote $p \colon \mathbb{R}^n \to \mathbb{R}$ and $F \colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ as the implicit AL and its smooth part, respectively. Also, let $P \colon \mathbb{R}^n \times \mathbb{R}^m \to \overline{\mathbb{R}}$ be the explicit AL; see Section 4.2. Patterning the transition from $\mathscr{L}_\mu^{\mathrm{S}}(\cdot, \cdot, y)$ in (4.1) to $\mathscr{L}_\mu(\cdot, y)$ in (4.3), we suppose that an oracle $\mathbf{O}$ is available that performs the explicit (exact, global) minimization of P with respect to $z \in \mathbb{R}^m$, apparently alluding to (4.2). In relation to the k-th iteration our AL method, we write $\mathbf{O} = \mathbf{O}_{\mu_k}(\cdot, \hat{y}^k)$. Notice that, for all $x \in \mathbb{R}^n$, it is $\mathbf{O}(x) \subseteq \operatorname{dom} g$ and thus $p(x) < \infty$. We collect some basic properties under Assumption 1.1.

Proposition 5.1. *Consider* (4.6)*,* (4.10)*, and let Assumption 1.1 hold. Then,*

   (i) $F \colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ *is continuously differentiable;*
   (ii) $g \colon \mathbb{R}^m \to \overline{\mathbb{R}}$ *is proper, lower semicontinuous, and prox-bounded;*
   (iii) $P \colon \mathbb{R}^n \times \mathbb{R}^m \to \overline{\mathbb{R}}$, $P(x, z)$ *is level-bounded in z locally uniformly in x;*
   (iv) $\mathbf{O} \colon \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ *is locally bounded, nonempty- and compact-valued;*
   (v) $p \colon \mathbb{R}^n \to \mathbb{R}$ *is strictly continuous (i.e. locally Lipschitz).*

Proof. The assertions readily follow from Assumption 1.1(i)–(ii) and the definitions in (4.6) and (4.10). In particular, (iii) is due to prox-boundedness of g and that $\mu_k \le \gamma_g$, along with $\hat{y}^k$ being fixed and smoothness of f and c. Assertion (iv) is a consequence of the smoothness of c by Assumption 1.1(i) and of the properties of the proximal mapping [46, Theorem 1.25]. Assertion (v) follows from the identity $p(\cdot) = f(\cdot) + g^{\mu_k}(c(\cdot) + \mu_k \hat{y}^k)$ and the fact that the Moreau envelope is strictly continuous [46, Example 10.32]. $\qquad\square$

Owing to Assumption 1.1 and Proposition 5.1(v), the objective function p is proper and lsc. However, as the existence of a minimizer for p is not guaranteed a priori, we also consider the following assumption. In general, it is verified under some coercivity or (level) boundedness conditions.

Assumption 5.1. Consider (4.6). Then, $\inf p \in \mathbb{R}$.

Remark 5.1. The problem format (P) could incorporate also an abstract constraint $x \in X$, with the set $X \subseteq \mathbb{R}^n$ being nonempty, closed, and *simple*. Although not included for the sake of a simpler exposition, the results in this paper could be readily extended to account for a *convex* X. As noted in [44], a possibility is to embed $x \in X$ in the template (P) by replacing $c(x)$ by $\bar{c}(x) := (c(x), x)$ and $g(z)$ by $\bar{g}(z, z') := g(z) + \delta_X(z')$. This approach, however, introduces an extra auxiliary component z', with an extra multiplier component y', while relaxing the simple constraint $x \in X$. Following [1] instead, a sharper strategy is to treat $x \in X$ directly as an unrelaxable constraint for the AL subproblem, and not incorporating it into the implicit AL function. Exploiting the projection onto X, the inner solver could generate iterates that satisfy the abstract constraint by design. With potentially nonconvex X, though, it appears nontrivial to formulate a suitable termination criterion for attesting approximate stationarity.

The possibility to localize the AL subproblem within a compact set X helps in securing Assumption 5.1. Together with properness and lower semicontinuity of p, compactness of $X \subset \mathbb{R}^n$ ensures that, once localized, the minimization of p is well-posed. In the spirit of trust region methods [12], however, it would be necessary to adapt the admissible region X to avoid interfering with the outer AL loop convergence. These considerations are left for a future effort.

5.1. Descent Methods with an Oracle. In this section we detail and discuss an algorithm for the minimization of p, which proceeds by taking steps along directions of descent for p. The

oracle-equipped descent method is stated in Algorithm 5.1, which closely patterns [6, Algorithm 8.1] when variable z is neglected — recovering a descent method for unconstrained smooth optimization. We prove well-definedness of the algorithm and investigate its convergence properties in Section 5.2.

Algorithm 5.1: Nonmonotone Descent Method with an Oracle

Data: $x^0 \in \mathbb{R}^n$, $\varepsilon > 0$
Result: ε-Υ-stationary point $x^* \in \mathbb{R}^n$ with certificate $z^* \in \mathbf{O}(x^*)$

1 Select parameters $\alpha, \beta, \theta \in (0,1)$, $\omega, \nu \in (0,1]$ and set $\Phi_0 \leftarrow p(x^0)$
2 **for** $j = 0, 1, \dots$ **do**
3   Select $z^j \in \mathbf{O}(x^j)$ and set $\nabla p^j \leftarrow \nabla_x F(x^j, z^j)$
4   **if** $\|\nabla p^j\| \leq \varepsilon$ **then return** $(x^*, z^*) \leftarrow (x^j, z^j)$
5   Compute $d^j \in \mathbb{R}^n$ such that

$$\langle \nabla p^j, d^j \rangle \leq -\theta \|\nabla p^j\| \|d^j\| \qquad \text{and} \qquad \|d^j\| \geq \omega \|\nabla p^j\|$$

6   Compute $x^{j+1} \in \mathbb{R}^n$ and the largest $\gamma_j = \beta^{\ell_j}$, $\ell_j \in \mathbb{N}$, such that

$$p(x^{j+1}) \leq p(x^j + \gamma_j d^j) \leq \Phi_j + \alpha \gamma_j \langle \nabla p^j, d^j \rangle$$

7   Set $\Phi_{j+1} \leftarrow (1 - \nu)\Phi_j + \nu p(x^{j+1})$

Some comments are in order. First, the method requires a starting point $x^0 \in \mathbb{R}^n$ for variable x only, not for z. Then, at Step 3, a surrogate gradient $\nabla p^j \in \mathbb{R}^n$ of p at x^j is evaluated: the oracle returns an arbitrary element $z^j \in \mathbf{O}(x^j)$ and, in light of the structure (4.10), we let ∇p^j be the partial derivative $\partial_x P$ at (x^j, z^j). At Step 5 a suitable search direction $d^j \in \mathbb{R}^n$ has to be selected, possibly among many. The simplest choice is the *steepest* direction given by the negative gradient, i.e. $d^j := -\nabla p^j$, which is always valid with respect to the conditions at Step 5, for any choice of $\theta \in (0,1)$ and $\omega \in (0,1]$. Another possibility is to consider Newton-like directions, based on local quadratic approximations. Barzilai-Borwein stepsizes and (limited memory) BFGS directions are popular strategies in this domain; see [47, Section 5.1.2] for an overview. Afterwards, the inner loop at Step 6 seeks an appropriate stepsize $\gamma_j > 0$ such that the update $x^j + \gamma_j d^j$ yields some improvement. In particular, this backtracking linesearch procedure is herein included as a globalization mechanism. Finally, with $\nu \in (0,1)$, the update rule at Step 7 allows to relax the requirement of a monotone decrease for the objective values, thus reducing conservatism of the linesearch and possibly encouraging longer steps and improved practical performance [6, Section 8.2], [16, 47]. However, a monotone behavior for $\{p(x^j)\}_{j \in \mathbb{N}}$ can be enforced with $\nu := 1$.

Crucially, the tentative update and the backtracking condition at Step 6 for adapting the stepsize γ_k are concerned with variable x only. Also, the oracle evaluation at Step 3 can be formally excluded from the algorithm and incorporated into the *surrogate* gradient evaluation of p at x^j. This is possible thanks to the arbitrariness of the element z^j chosen by the oracle $\mathbf{O}$, which allows us to keep variable z implicit. These observations show that Algorithm 5.1 is equivalent to a descent method applied directly to the minimization of p (with a surrogate gradient), thus making the oracle-equipped algorithm a formal tool only, transparent to the user.

5.2. **Convergence Analysis.** In this section we investigate the convergence properties of Algorithm 5.1, exploiting continuous differentiability of F and the oracle's optimality to prove sufficient decrease with respect to x. We are going to show first that Algorithm 5.1 is well defined, namely that each iteration terminates in finite time. This is an essential theoretical and practical condition that an implementable algorithm must satisfy.

Lemma 5.1 (Well definedness). *Let Assumption 1.1 hold and consider the iterates generated by Algorithm 5.1. Then,*

(i) at every iteration, the number of backtrackings at Step 6 is finite;
(ii) at the end of the j-th iteration, one has

$$p(x^{j+1}) \le \Phi_j + \Delta_j \qquad \text{and} \qquad p(x^{j+1}) \le \Phi_{j+1} \le \Phi_j + \nu\Delta_j$$

where $\Delta_j := \alpha\gamma_j\langle\nabla p^j, d^j\rangle < 0$;
(iii) every iterate x^j remains within the sublevel set $\{x \in \mathbb{R}^n \mid p(x) \le \Phi_0 < \infty\}$.

Proof. Recall the basic properties collected in Proposition 5.1.

(i) Let us discard the trivial case by assuming that $\|\nabla p^j\| > \varepsilon$. Since $\theta \in (0,1)$ and $\omega \in (0,1]$, the choice $d^j \leftarrow -\nabla p^j$ is always viable, hence Step 5 is well-defined. Denote $x^{j+} := x^j + \gamma_j d^j$ and $z^{j+} \in \mathbf{O}(x^{j+})$. Then,

$$p(x^{j+}) = F(x^{j+}, z^{j+}) + g(z^{j+}) \le F(x^{j+}, z) + g(z)$$

for all $z \in \mathbb{R}^m$. In particular, $p(x^{j+}) \le F(x^{j+}, z^j) + g(z^j)$. By continuous differentiability of F, combining Steps 3 and 5 yields

$$\lim_{\gamma \to 0} \frac{F(x^j + \gamma d^j, z^j) - F(x^j, z^j)}{\gamma} = \langle\nabla_x F(x^j, z^j), d^j\rangle = \langle\nabla p^j, d^j\rangle < 0.$$

Thus,

$$\lim_{\gamma \to 0} \frac{F(x^j + \gamma d^j, z^j) - F(x^j, z^j)}{\gamma\langle\nabla p^j, d^j\rangle} = 1.$$

Therefore, since $\alpha \in (0,1)$, for some $\gamma_j = \beta^{\ell_j} > 0$ small enough, we have that

$$\frac{F(x^j + \gamma_j d^j, z^j) - F(x^j, z^j)}{\gamma_j\langle\nabla p^j, d^j\rangle} \ge \alpha.$$

As $\langle\nabla p^j, d^j\rangle < 0$ by Step 5 and $\gamma_j > 0$, we deduce that

$$F(x^{j+}, z^j) \le F(x^j, z^j) + \alpha\gamma_j\langle\nabla p^j, d^j\rangle.$$

Now, invoking the oracle to get $z^{j+} \in \mathbf{O}(x^{j+})$, we obtain

$$p(x^{j+}) \le F(x^{j+}, z^j) + g(z^j)$$
$$\le F(x^j, z^j) + g(z^j) + \alpha\gamma_j\langle\nabla p^j, d^j\rangle = p(x^j) + \alpha\gamma_j\langle\nabla p^j, d^j\rangle.$$

Then, finite termination of the inner loop follows from the bound $\Phi_j \ge p(x^j)$ valid for all $j \in \mathbb{N}$, which is obtained by induction as in [16, Lemma 4.1(i)]: for $j = 0$ it is $\Phi_j = p(x^j)$ by construction; if $j > 0$, the bound is due to Step 7.

(ii) The first inequality follows by combining (i) with the failure of the condition at Step 6. Together with Step 7, this gives the lower bound

$$\Phi_{j+1} = (1-\nu)\Phi_j + \nu p(x^{j+1}) \geq (1-\nu)[p(x^{j+1}) - \Delta_j] + \nu p(x^{j+1})$$
$$= p(x^{j+1}) - (1-\nu)\Delta_j \geq p(x^{j+1})$$

as well as the upper bound

$$\Phi_{j+1} = (1-\nu)\Phi_j + \nu p(x^{j+1}) \leq (1-\nu)\Phi_j + \nu(\Phi_j + \Delta_j) = \Phi_j + \nu\Delta_j.$$

(iii) That $\Phi_0 = p(x^0) < \infty$ follows from the fact that $z^0 \in \mathbf{O}(x^0) \in \operatorname{dom} g \neq \emptyset$. Then, the assertion is a direct consequence of (ii), by telescoping $p(x^j) \leq \Phi_j \leq \ldots \leq \Phi_0$. □

We next consider an asymptotic analysis of the algorithm.

Lemma 5.2. *Let Assumptions 1.1 and 5.1 hold and consider a sequence of iterates $\{x^j\}_{j\in\mathbb{N}}$ generated by Algorithm 5.1 with $\varepsilon = 0$. Then, the sequences $\{\Phi_j\}_{j\in\mathbb{N}}$ and $\{p(x^j)\}_{j\in\mathbb{N}}$ converge to a finite value $p_* \geq \inf p$, the former from above.*

Proof. Regarding $\{\Phi_j\}_{j\in\mathbb{N}}$, the assertion follows from Lemma 5.1(ii) and Assumption 5.1. Then, rearranging the update rule at Step 7, one can observe that $p(x^{j+1}) = \nu^{-1}(\Phi_{j+1} - \Phi_j) + \Phi_j$ holds. Taking the limit for $j \to \infty$, by $\nu > 0$ the convergence of $\{\Phi_j\}_{j\in\mathbb{N}}$ implies that of $\{p(x^j)\}_{j\in\mathbb{N}}$ to the same finite value p_*. □

According to Lemma 5.1(ii) and Lemma 5.2, the merit and objective values tend to decrease along the iterations and, if $\inf p > -\infty$, settle at some finite value p_*. The following Theorem 5.1 provides a guarantee of global convergence to Υ-stationary points, namely that any accumulation point of the iterates $\{x^j\}_{j\in\mathbb{N}}$ generated by Algorithm 5.1 is Υ-stationary relative to the minimization of p, regardless of the initialization x^0 and algorithmic parameters.

Theorem 5.1. *Let Assumptions 1.1 and 5.1 hold and consider a sequence of iterates $\{x^j\}_{j\in\mathbb{N}}$ generated by Algorithm 5.1 with $\varepsilon = 0$. Let $x^* \in \mathbb{R}^n$ be an accumulation point of $\{x^j\}_{j\in\mathbb{N}}$. Then, x^* is Υ-stationary for p, namely there exists $z^* \in \mathbf{O}(x^*)$ such that $\nabla_x F(x^*, z^*) = 0$.*

Proof. If Algorithm 5.1 terminates after finitely many iterations returning (x^j, z^j), then it must be that $\|\nabla p^j\| = 0$, hence $\nabla_x F(x^j, z^j) = 0$. Since $z^j \in \mathbf{O}(x^j)$ for all $j \in \mathbb{N}$, then x^j is Υ-stationary.

Now suppose instead that Algorithm 5.1 takes infinitely many iterations. Let $\{x^j\}_{j\in J}$ be a subsequence such that $x^j \to_J x^*$, for some $J := \{j_0, j_1, j_2 \ldots\} \subseteq \mathbb{N}$. Then, continuity of p by Proposition 5.1(v) gives $p(x^j) \to_J p(x^*)$. Owing to the linesearch condition at Step 6, since $j_{i+1} \geq j_i + 1$, we have that

$$p(x^{j_{i+1}}) \leq p(x^{j_i+1}) \leq \Phi_{j_i} + \alpha\gamma_{j_i}\langle \nabla p^{j_i}, d^{j_i}\rangle < \Phi_{j_i}.$$

Then, Lemma 5.2 implies that $\gamma_j\langle \nabla p^j, d^j\rangle \to_J 0$, since $\alpha > 0$, and, due to Step 5, it must be that $\gamma_j\|\nabla p^j\|\|d^j\| \to_J 0$. As any term in the limit could vanish, there are three (non exclusive) cases. Now we argue that, up to extracting a subsequence, it is $\nabla p^j \to_J 0$ in every case. First, the claim is trivial if ∇p^j vanishes. Second, if $d^j \to_J 0$, conditions at Step 5 imply that ∇p^j vanishes too, giving the claim. Third, if $\gamma_j \to_J 0$, we proceed by showing that there is a longer stepsize $\hat{\gamma}_j$ that also tends to zero, along which the function did not decrease enough, and that this is possible only if the gradient vanishes.

Let us start by noticing the failure of attempts at Step 6. The previous (hence rejected) stepsize $\hat{\gamma}_j := \beta^{-1}\gamma_j = \beta^{\ell_j-1}$ gives

$$p(x^j + \hat{\gamma}_j d^j) > \Phi_j + \alpha\hat{\gamma}_i \langle \nabla p^j, d^j \rangle \geq p(x^j) + \alpha\hat{\gamma}_j \langle \nabla p^j, d^j \rangle, \tag{5.1}$$

where the last inequality follows from the bound $\Phi_j \geq p(x^j)$ of Lemma 5.1(ii). In the case $\gamma_j \to_J 0$, it must be $\hat{\gamma}_j \to_J 0$ too, since $\ell_j \to_J 0$. Rearranging terms and using the structure of P in (4.10), the minimizing property of any $z^j \in \mathbf{O}(x^j)$ yields

$$\alpha \langle \nabla p^j, d^j \rangle < \frac{p(x^j + \hat{\gamma}_j d^j) - p(x^j)}{\hat{\gamma}_j} \leq \frac{F(x^j + \hat{\gamma}_j d^j, z^j) - F(x^j, z^j)}{\hat{\gamma}_j}.$$

Defining $s^j := \hat{\gamma}_j d^j$ for all $j \in \mathbb{N}$, we have that $\|s^j\| \to_J 0$. Possibly extracting a subsequence and redefining J, let $s^* \in \mathbb{R}^n$ be such that $s^j/\|s^j\| \to_J s^*$. Now, for all $j \in J$, by the mean value theorem there exists $\zeta_j \in [0,1]$ such that

$$\begin{aligned}
\langle \nabla_x F(x^j + \zeta_j s^j, z^j), s^j \rangle &= F(x^j + s^j, z^j) - F(x^j, z^j) \\
&\geq p(x^j + s^j) - p(x^j) \\
&> \alpha \langle \nabla p^j, s^j \rangle = \alpha \langle \nabla_x F(x^j, z^j), s^j \rangle,
\end{aligned}$$

where the inequalities are due to the oracle's optimality and (5.1). Dividing by $\|s^j\|$ and rearranging give

$$\frac{\langle \nabla_x F(x^j + \zeta_j s^j, z^j) - \alpha \nabla_x F(x^j, z^j), s^j \rangle}{\|s^j\|} > 0.$$

Owing to $\zeta_j s^j \to_J 0$ and continuity of ∇F, taking the limit yields

$$\limsup_{j \to_J \infty} \frac{\langle \nabla_x F(x^j + \zeta_j s^j, z^j) - \alpha \nabla_x F(x^j, z^j), s^j \rangle}{\|s^j\|} = \limsup_{j \to_J \infty} (1 - \alpha) \frac{\langle \nabla_x F(x^j, z^j), s^j \rangle}{\|s^j\|} \leq 0,$$

where the inequality is due to $\langle \nabla_x F(x^j, z^j), s^j \rangle = \hat{\gamma}_j \langle \nabla p^j, d^j \rangle < 0$ for all $j \in \mathbb{N}$, by Step 5, and $\alpha < 1$. Recalling that $s^j/\|s^j\| \to_J s^*$, the inequalities above demonstrate that it is necessarily $\langle \nabla_x F(x^j, z^j), s^* \rangle = \langle \nabla p^j, s^* \rangle \to_J 0$. Then, by the conditions at Step 5, it must be $\nabla p^j \to_J 0$.

Finally, owing to $\nabla p^j := \nabla_x F(x^j, z^j) \to_J 0$ in all cases, $z^j \in \mathbf{O}(x^j)$ for all $j \in J$, $x^j \to_J x^*$, and continuity of ∇F, we conclude that $\nabla_x F(x^*, z^*) = 0$ for some $z^* \in \mathbf{O}(x^*)$, proving the claim that x^* is Υ-stationary for p. $\qquad\square$

5.3. Termination Criteria.
Algorithm 5.1 would run indefinitely without Step 4, generating an infinite sequence of iterates. Suitable termination criteria need to be inserted for stopping and returning an iterate that, in some sense, is satisfactory relatively to the minimization of p. Thus, we equip Algorithm 5.1 with the termination condition $\|\nabla_x F(x^j, z^j)\| \leq \varepsilon$, which encodes ε-Υ-stationarity of x^j relative to p, since $z^j \in \mathbf{O}(x^j)$ for all $j \in \mathbb{N}$. Thanks to Theorem 5.1 we can now easily infer termination of Algorithm 5.1 in finitely many steps for any tolerance $\varepsilon > 0$.

Corollary 5.1. *Let Assumption 1.1 hold and consider a sequence of iterates $\{x^j, z^j\}_{j \in \mathbb{N}}$ generated by Algorithm 5.1 with $\varepsilon > 0$. Suppose that the sequence $\{x^j\}_{j \in \mathbb{N}}$ admits an accumulation point. Then, in finitely many steps an ε-Υ-stationary point x^* is returned with certificate $z^* \in \mathbf{O}(x^*)$.*

Proof. That the algorithm terminates in finitely many iterates, say j_* many, follows from Theorem 5.1, since $\|\nabla_x F(x^j, z^j)\| = \|\nabla p^j\| \to_J 0$ along any convergent subsequence $\{x^j\}_{j \in J}$. Since $z^j \in \mathbf{O}(x^j)$ for all $j \in \mathbb{N}$, it follows that the output $x^* := x^{j_*}$ with certificate $z^* := z^{j_*}$ satisfies $z^* \in \mathbf{O}(x^*)$. Since the magnitude of $\nabla_x F(x^*, z^*)$ is no more than ε by the termination criterion, the pair (x^*, z^*) is ε-Υ-stationary for p. $\square$

Although this result confirms that Algorithm 5.1's output complies with the requirements of the outer AL framework at Step 4 of Algorithm 4.1, note that the existence of limit points is assumed, and not guaranteed, in Corollary 5.1. This can be satisfied by employing, in general, (level) boundedness or coercivity arguments; see Assumption 5.1 and Remark 5.1 as well.

6. NUMERICAL EXAMPLES

This section showcases the flexibility offered by the (nonconvex) generalized programming framework (P) and presents some numerical results that support the implicit approach developed in Section 4. We consider two illustrative examples formulated as constrained structured optimization problems and compare Algorithm 4.1 against ALS from [17], an AL solver based on an explicit treatment of auxiliary variables.

We have implemented the proposed approach in the Augmented Lagrangian Proximal Solver (ALPS), as part of the open-source software package Bazinga.jl [13], which includes ALS as well as the examples discussed below. For the sake of a fair comparison, the implementation of ALPS closely patterns that of ALS; for more details the interested reader may refer to the source code and [17]. In particular, to obtain comparable results, the AL subproblems are solved using the PANOC$^+$ solver with LBFGS acceleration [20]. As a proximal gradient method, PANOC$^+$ can handle the two-term structure of the explicit AL function; it is in fact the default inner solver in ALS [17]. However, PANOC$^+$ can act as a gradient-based descent method too, and thus can be adopted as inner solver within ALPS, in view of Section 5.1.

Despite the fact that solutions and stationary points remain unaffected by the lifting from (P) to (P$_S$), as shown by [17, Lemma 3.1], there can be a significant gap in terms of objective values attained in practice. Based on the arguments in [3] and the numerical results in Section 6.2, we are drawn to suspect that the theoretical advantages of implicit formulations can be materialized by the oracle's action. But the advantages of ALPS over ALS could go beyond this aspect, as for both examples the implicit approach seems consistently effective in reducing the number of iterations and execution time.

6.1. Nonsmooth Rosenbrock and either-or constraints. Let us consider the two-dimensional optimization problem of [17, Section 4.2]: it involves a nonsmooth Rosenbrock-like objective function and either-or constraints. The original formulation in disjunctive form reads

$$\begin{aligned}
\underset{x \in \mathbb{R}^2}{\text{minimize}} \quad & f(x) + |x_1| & (6.1) \\
\text{subject to} \quad & x_2 \leq -x_1 \quad \vee \quad x_2 \geq x_1
\end{aligned}$$

where $f(x) := 10(x_2 + 1 - (x_1 + 1)^2)^2$. The feasible set is nonconvex but simple to project onto, and the problem admits a unique (global) minimizer $x^* = (0,0)$, where objective function and feasible region have a kink; see Figure 1. We cast (6.1) into the form of (P) with $n = 2$, $m = 3$,

and by defining the data functions

$$c(x) := (x_1, -x_1 - x_2, -x1 + x_2)$$
$$g(c) := |c_1| + \delta_{D_{EO}}(c_2, c_3)$$

where the (nonconvex) either-or set $D_{EO} \subset \mathbb{R}^2$ is described by

$$D_{EO} := \{(a,b) \in \mathbb{R}^2 \,|\, a \geq 0\} \,\cup\, \{(a,b) \in \mathbb{R}^2 \,|\, b \geq 0\};$$

the (set-valued) projection $\Pi_{D_{EO}}$ can be easily evaluated.

We consider a uniform grid of $41^2 = 1681$ starting points $x^0 \in [-5,5]^2$. Both ALS and ALPS solve all the problem instances, approximately reaching the optimal $x^* = (0,0)$ in all cases. Descriptive statistics for comparing the performance of ALS and ALPS are reported in Table 1. A close look at the runtime distribution depicted in Figure 1 indicates that both solvers struggle with some (relatively few) instances for which the other one maintains its performance. Nevertheless, ALPS is consistently faster than ALS, in terms of both iterations and runtime; for most instances, it is at least three times better.
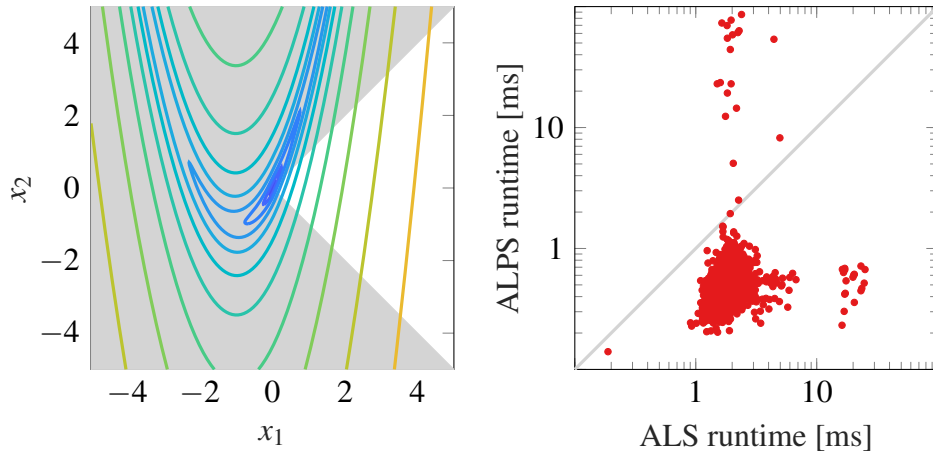

FIGURE 1. Left: feasible set (gray background) and contour lines for the objective of (6.1). Right: performance comparison in terms of runtime, with each marker associated to a starting point and the solid line having unitary slope.

TABLE 1. Performance comparison of ALPS and ALS on (6.1). Statistical summary using different quantiles for the number of inner iterations and runtime.

	total inner iterations				
	q(1%)	q(25%)	q(50%)	q(75%)	q(99%)
ALPS	24	37	44	52	248
ALS	84	109	122	137	308
	runtime [ms]				
	q(1%)	q(25%)	q(50%)	q(75%)	q(99%)
ALPS	0.25	0.39	0.48	0.58	6.30
ALS	1.15	1.60	1.82	2.05	8.59

6.2. **Problem with Vanishing Constraints.** Let us consider the following two-dimensional problem, which arises from truss topology optimization and is a well-known academic problem, thoroughly discussed in [31, Section 9.5.1]:

$$\begin{aligned}
&\underset{x \in \mathbb{R}^2}{\text{minimize}} \quad 4x_1 + 2x_2 &&&&(6.2)\\
&\text{subject to} \quad x_1 \geq 0, \qquad x_2 \geq 0,\\
&\hphantom{\text{subject to} \quad} x_1 > 0 \quad \Longrightarrow \quad x_1 + x_2 \geq 5\sqrt{2},\\
&\hphantom{\text{subject to} \quad} x_2 > 0 \quad \Longrightarrow \quad x_1 + x_2 \geq 5.
\end{aligned}$$

Here nonnegativity constraints and logical implications give rise to the so called vanishing constraints. The feasible set of (6.2) consists of the union of an unbounded polyhedron, a line segment, and an isolated point, as depicted in Figure 2:

$$\{(x_1,x_2) \in \mathbb{R}^2 \,|\, x_1 + x_2 \geq 5\sqrt{2}, \ x_1, x_2 \geq 0\} \ \cup \ \{(0,x_2) \in \mathbb{R}^2 \,|\, 5 \leq x_2 \leq 5\sqrt{2}\} \ \cup \ \{(0,0)\}.$$

Problem (6.2) admits a global minimum at $x^\star = (0,0)$ and a local minimum at $x^\diamond = (0,5)$.

There exist several ways to cast (6.2) into the form of (P). We are going to consider three different formulations, with increasing degree of explicitness for treating constraints: implicit, intermediate (or partially explicit, avoiding auxiliary variable components for trivial constraints), and (fully) explicit. Let us define the cost function $f(x) := 4x_1 + 2x_2$ and denote by $D_{\mathrm{VC}} \subset \mathbb{R}^2$ the (nonconvex) set associated to a vanishing constraint:

$$D_{\mathrm{VC}} := \{(a,b) \in \mathbb{R}^2 \,|\, a \geq 0, ab \geq 0\},$$

whose (set-valued) projection $\Pi_{D_{\mathrm{VC}}}$ can be easily evaluated.

- An *implicit* representation of (6.2) in the form (P) is given by the data functions

$$c(x) := (x_1, \, x_1 + x_2 - 5\sqrt{2}, \, x_2, \, x_1 + x_2 - 5)$$
$$g(c) := \delta_{D_{\mathrm{VC}}}(c_1, c_2) + \delta_{D_{\mathrm{VC}}}(c_3, c_4).$$

  With size $(n,m) = (2,4)$, this seems the most compact formulation for (6.2).

- An *explicit* reformulation of (6.2) as (P) is obtained by introducing auxiliary variables as in (P$_{\mathrm{S}}$), leading to a problem with size $(n,m) = (6,8)$. Notice that this explicit reformulation is carried out internally by ALS when fed with the implicit formulation, as a lifted representation of the original problem [17, Section 3].

- An *intermediate* model of (6.2) into the form of (P) is found by avoiding auxiliary variable components for trivial constraints. With data functions

$$c(x) := (x_1, \, x_1 + x_2 - 5\sqrt{2} - x_3, \, x_2, \, x_1 + x_2 - 5 - x_4, \, x_3, \, x_4)$$
$$g(c) := \delta_{D_{\mathrm{VC}}}(c_1, c_5) + \delta_{D_{\mathrm{VC}}}(c_3, c_6) + \delta_0(c_2) + \delta_0(c_4),$$

  this formulation has size $(n,m) = (4,6)$.

We execute ALPS on all problem representations. The outcome of ALS [17] on the implicit formulation is also recorded for comparison. We consider a grid of $51^2 = 2601$ starting points x^0 in $[-5,20]^2$; auxiliary variables are initialized so as to satisfy the linear constraints arising from the reformulation. For all instances both solvers return with a successful status, always indicating either the global $x^\star$ or local $x^\diamond$ minimizer. In Figure 2 we depict the feasible region of (6.2) and display these results in more detail; Table 2 collects relevant statistics. Since the

set D_{VC} is nonconvex, the iterates of Algorithm 4.1 depend on the arbitrary choices made when the projection onto D_{VC} is not a singleton: by selecting different projections, we observed analogous numerical results, albeit not identical.
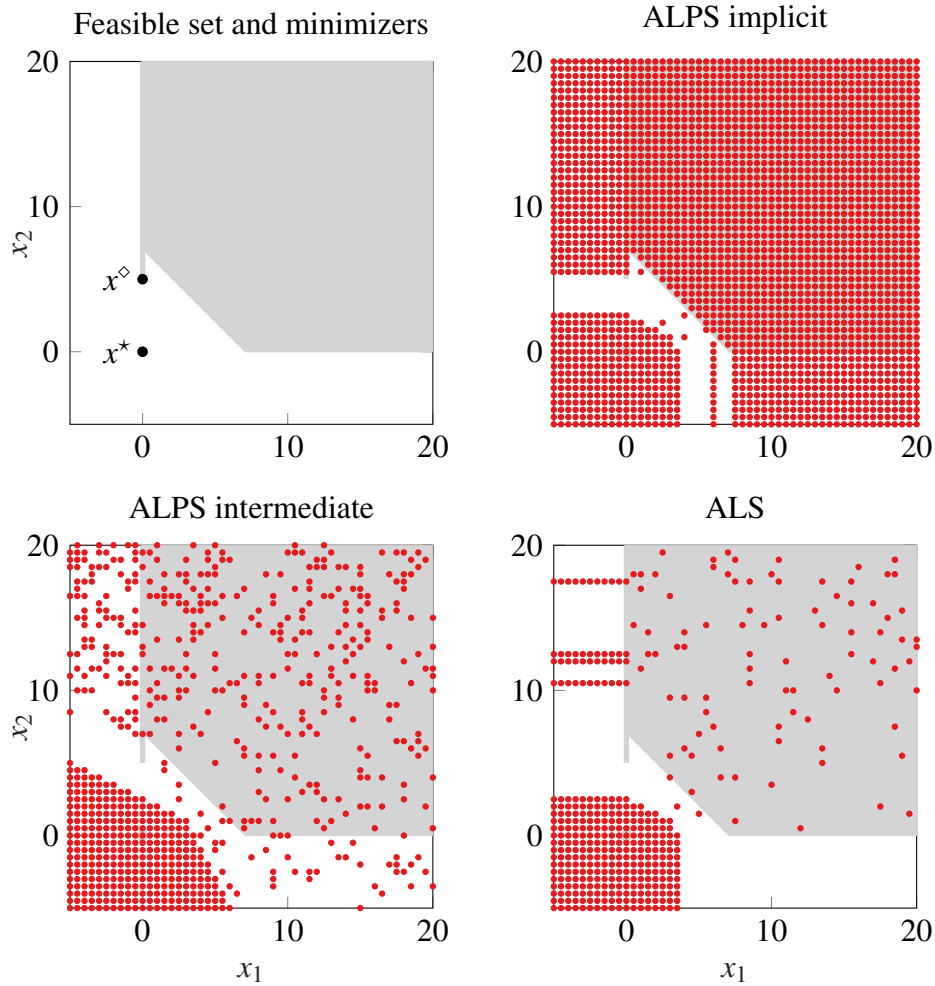

FIGURE 2. Comparison of solutions obtained using different methods and formulations. Feasible set (gray background) and grid of (2601 uniformly spaced) starting points and corresponding termination mark depending on the returned solution: global minimizer $x^\star = (0,0)$ (red circle) or local minimizer $x^\diamond = (0,5)$ (omitted mark).

Considering the solution's quality, the global minimizer $x^\star = (0,0)$ is most probably returned by ALPS with *implicit* problem formulation (93%). With the *intermediate* formulation, ALPS seems to consistently find $x^\star$ only when starting nearby (28%), and even more so with the fully *explicit* formulation (19%) or with ALS (16%), as indicated by Figure 2. Regarding the (median) computational effort, ALPS with implicit formulation exhibits the best and most consistent performance, about eight times better than the second best.

TABLE 2. Comparison of solutions obtained and performances using different methods and formulations. Statistics on the number of (cumulative inner) iterations, execution time, and how often the global minimizer is returned.

		ALPS *implicit*	ALPS *intermediate*	ALPS *explicit*	ALS
total inner	q(1%)	10	12	12	10
iterations	q(50%)	10	77	117	117
	q(75%)	10	87	142	140
	q(99%)	98	120	187	183
runtime	q(1%)	0.15	0.16	0.16	0.21
[ms]	q(50%)	0.16	1.33	1.64	2.69
	q(75%)	0.18	1.62	2.06	3.39
	q(99%)	11.26	5.42	3.15	17.83
global	abs	2413	721	496	405
minimizer	rel	92.77%	27.72%	19.07%	15.57%

7. CONCLUDING REMARKS

The theoretical investigation of generalized optimization in the fully nonconvex setting has led to a significant extension of the augmented Lagrangian (AL) framework. With the help of a lifted reformulation, KKT-type stationarity conditions have been derived in Lagrangian terms and complemented with asymptotic and regularity notions. The implicit AL function and subproblems have been defined, marginalizing their explicit counterparts with respect to auxiliary variables, seemingly sacrificing some regularity due to set-valuedness. Then, rooted on a parametric interpretation, we devised a suitable stationarity concept for characterizing the subproblems' solutions, weaker than M-stationarity for the implicit AL function, but strong enough for investigating asymptotic properties. We considered a safeguarded AL algorithm and established opportune convergence guarantees, on par with those in nonlinear programming.

Embracing an implicit philosophy was not only a theoretical curiosity, though. We developed and characterized a solver for the implicit AL subproblems, which is capable of generating suitable iterates. Equipped with a proximal oracle, it handles fewer decision variables, maintaining auxiliary variables under the hood. This appears to produce a tighter model and catalyze improved practical performance, compared to more explicit approaches, as witness by numerical results on an illustrative problem.

Several issues and questions remain open. It is a common practice to assume that the AL subproblems are well-posed, possibly based on coercivity or boundedness arguments. In general, however, to overcome this concern one could investigate the combination of AL schemes with trust-region strategies, as hinted at in Remark 5.1. Future works should also elaborate on a characterization of stationary points as saddle points of the implicit AL function, a local duality theory for generalized optimization, and the relationships with proximal point algorithms.

REFERENCES

[1] R. Andreani, E. G. Birgin, J. M. Martínez, M. L. Schuverdt, On augmented Lagrangian methods with general lower–level constraints, SIAM J. Optim. 18 (2008), 1286–1309.

[2] E. Balas, Disjunctive Programming, Springer, Cham, 2018.

[3] M. Benko, P. Mehlitz, On implicit variables in optimization theory, J. Nonsmooth Anal. Optim. 2 (2021), 7215.

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

[5] E. G. Birgin, J. M. Martínez, Augmented Lagrangian method with nonmonotone penalty parameters for constrained optimization, Comput. Optim. Appl. 51 (2012), 941–965.

[6] E. G. Birgin, J. M. Martínez, Practical Augmented Lagrangian Methods for Constrained Optimization, SIAM, Philadelphia, 2014.

[7] J. Bolte, S. Sabach, M. Teboulle, Nonconvex lagrangian-based optimization: Monitoring schemes and global convergence, Math. Oper. Res. 43 (2018), 1210–1232.

[8] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, Distributed optimization and statistical learning via the alternating direction method of multipliers, Foundations and Trends in Machine Learning, 3 (2011), 1–122.

[9] X. Chen, L. Guo, Z. Lu, J. J. Ye, An augmented Lagrangian method for non-Lipschitz nonconvex programming, SIAM J. Numer. Anal. 55 (2017), 168–193.

[10] P. L. Combettes, J.-C. Pesquet, Proximal Splitting Methods in Signal Processing, Springer, New York, 2011.

[11] A. R. Conn, N. I. M. Gould, P. L. Toint, A globally convergent augmented Lagrangian algorithm for optimization with general constraints and simple bounds, SIAM J. Numer. Anal. 28 (1991), 545–572.

[12] A. R. Conn, N. I. M. Gould, P. L. Toint, Trust Region Methods, SIAM, Philadelphia, 2000.

[13] A. De Marchi, Bazinga.jl: A toolbox for constrained structured optimization in Julia, https://github.com/aldma/Bazinga.jl, 2020. Accessed: 2023-01-21.

[14] A. De Marchi, Augmented Lagrangian and Proximal Methods for Constrained Structured Optimization, PhD thesis, University of the Bundeswehr Munich, Neubiberg, Germany, 2021.

[15] A. De Marchi, On a primal-dual Newton proximal method for convex quadratic programs, Comput. Optim. Appl. 81 (2022), 369–395.

[16] A. De Marchi, Proximal gradient methods beyond monotony, J. Nonsmooth Anal. Optim. 4 (2023), 10290.

[17] A. De Marchi, X. Jia, C. Kanzow, P. Mehlitz, Constrained composite optimization and augmented Lagrangian methods, Math. Program. 201 (2023), 863–896.

[18] A. De Marchi, P. Mehlitz, Local properties and augmented Lagrangians in fully nonconvex composite optimization, arXiv:2309.01980v2, 2023.

[19] A. De Marchi, A. Themelis, An interior proximal gradient method for nonconvex optimization, arXiv:2208.00799v2, 2022.

[20] A. De Marchi, A. Themelis, Proximal gradient algorithms under local Lipschitz gradient continuity, J. Optim. Theory Appl. 194 (2022), 771–794.

[21] N. K. Dhingra, S. Z. Khong, M. R. Jovanović, The proximal augmented Lagrangian method for nonsmooth composite optimization, IEEE Trans. Automat. Control 64 (2019), 2861–2868.

[22] D. Drusvyatskiy, A. D. Ioffe, A. S. Lewis, Nonsmooth optimization using Taylor-like models: Error bounds, convergence, and termination criteria, Math. Program. 185 (2021), 357–383.

[23] B. Evens, P. Latafat, A. Themelis, J. Suykens, P. Patrinos. Neural network training as an optimal control problem: An augmented Lagrangian approach, In: 2021 60th IEEE Conference on Decision and Control (CDC), pp. 5136–5143, IEEE, 2021.

[24] P. E. Gill, D. P. Robinson, A primal-dual augmented Lagrangian, Comput. Optim. Appl. 51 (2012), 1–25.

[25] G. N. Grapiglia, Y. X. Yuan, On the complexity of an augmented Lagrangian method for nonconvex optimization, IMA J. Numer. Anal. 41 (2020), 1546–1568.

[26] L. Guo, J. J. Ye, Necessary optimality conditions and exact penalization for non-Lipschitz nonlinear programs, Math. Program. 168 (2018), 571–598.

[27] G. Haeser, M. L. Schuverdtm On approximate KKT condition and its extension to continuous variational inequalities, J. Optim. Theory Appl. 149 (2011), 528–539.

[28] N. Hallak, M. Teboulle, An adaptive Lagrangian-based scheme for nonconvex composite optimization, Math. Oper. Res. 48 (2023), 1811-2382.

[29] B. Hermans, A. Themelis, P. Patrinos, QPALM: a proximal augmented Lagrangian method for nonconvex quadratic programs, Math. Program. Comput. 14 (2022), 497–541.

[30] M. R. Hestenes, Multiplier and gradient methods, J. Optim. Theory Appl. 4 (1969), 303–320.

[31] T. Hoheisel, Mathematical Programs with Vanishing Constraints, PhD thesis, University of Würzburg, Würzburg, 2009.

[32] C. Kanzow, D. Steck, D. Wachsmuth, An augmented Lagrangian method for optimization problems in Banach spaces, SIAM J. Control Optim. 56 (2018), 272–291.

[33] A. Y. Kruger, P. Mehlitz, Optimality conditions, approximate stationarity, and applications — a story beyond Lipschitzness, ESAIM Control Optim. Calc. Var. 28 (2022), 42.

[34] P. Latafat, P. Patrinos, Asymmetric forward-backward-adjoint splitting for solving monotone inclusions involving three operators, Comput. Optim. Appl. 68 (2017), 57–93.

[35] A. S. Lewis, S. J. Wright, A proximal method for composite minimization, Math. Program. 158 (2016), 501–546.

[36] P. Mehlitz, Asymptotic stationarity and regularity for nonsmooth optimization problems, J. Nonsmooth Anal. Optim. 1 (2020), 6575.

[37] B. S. Mordukhovich, Variational Analysis and Generalized Differentiation, Springer, Berlin, 2006.

[38] J. J. Moreau, Proximité et dualité dans un espace hilbertien, Bulletin de la Société Mathématique de France 93 (1965), 273–299.

[39] N. Parikh, S. Boyd, Proximal algorithms, Foundations and Trends in Optimization, 1 (2014), 127–239.

[40] M. J. D. Powell, A Method for Nonlinear Constraints in Minimization Problems, Academic Press, 1969.

[41] R. T. Rockafellar, A dual approach to solving nonlinear programming problems by unconstrained optimization, Math. Program. 5 (1973), 354–373.

[42] R. T. Rockafellar, Augmented Lagrangians and applications of the proximal point algorithm in convex programming, Math. Oper. Res. 1 (1976), 97–116.

[43] R. T. Rockafellar, Extended nonlinear programming. In G. Di Pillo and F. Giannessi, editors, Nonlinear Optimization and Related Topics, Applied Optimization, vol. 36, pp. 381–399, Kluwer, 2000.

[44] R. T. Rockafellar, Augmented Lagrangians and hidden convexity in sufficient conditions for local optimality, Math. Program. 198 (2023), 159–194.

[45] R. T. Rockafellar, Convergence of augmented Lagrangian methods in extensions beyond nonlinear programming, Math. Program. 199 (2023), 375–420.

[46] R. T. Rockafellar, R. J. B. Wets, Variational Analysis, volume 317, Springer, 1998.

[47] A. Themelis, L. Stella, P. Patrinos, Forward-backward envelope for the sum of two nonconvex functions: Further properties and nonmonotone linesearch algorithms, SIAM J. Optim. 28 (2018), 2274-2303.

[48] Y. Yuan, Conditions for convergence of trust region algorithms for nonsmooth optimization, Math. Program. 31 (1985), 220–228.

APPENDIX A. OMITTED PROOFS

Proof of Proposition 3.1. By local optimality of x^* for (P) there exists $\delta > 0$ such that $\varphi(x^*) \leq \varphi(x)$ is valid for all feasible $x \in \mathbb{B}_\delta(x^*)$. Consequently, x^* is the unique global minimizer of the localized problem

$$\underset{x \in \mathbb{B}_\delta(x^*)}{\text{minimize}} \quad \varphi(x) + \frac{1}{2}\|x - x^*\|^2. \tag{A.1}$$

Let us now consider the penalized surrogate problem

$$\underset{\substack{x \in \mathbb{B}_\delta(x^*) \\ z \in \mathbb{B}_1(c(x^*))}}{\text{minimize}} \quad f(x) + g(z) + \frac{\rho_k}{2}\|c(x) - z\|^2 + \frac{1}{2}\|x - x^*\|^2 \tag{P(k)}$$

where $k \in \mathbb{N}$ is arbitrary, $\rho_k > 0$, and the sequence $\{\rho_k\}_{k \in \mathbb{N}}$ satisfies $\rho_k \to \infty$ as $k \to \infty$. Noting that the objective function of this optimization problem is lsc while its feasible set is nonempty (by feasibility of x^*) and compact, it possesses a global minimizer (x^k, z^k) for each $k \in \mathbb{N}$. Therefore, without loss of generality, up to extracting a subsequence, we assume $x^k \to \widehat{x}$ and $z^k \to \widehat{z}$ for some $\widehat{x} \in \mathbb{B}_\delta(x^*)$ and $\widehat{z} \in \operatorname{dom} g \cap \mathbb{B}_1(c(x^*))$.

We now argue that $\widehat{x} = x^*$ and $\widehat{z} = c(\widehat{x}) = c(x^*)$. To this end, we note that $(x^*, c(x^*))$ is feasible to (P(k)), which yields the estimate

$$f(x^k) + g(z^k) + \frac{\rho_k}{2}\|c(x^k) - z^k\|^2 + \frac{1}{2}\|x^k - x^*\|^2 \leq \varphi(x^*) \qquad \text{(A.2)}$$

for each $k \in \mathbb{N}$. Using lower semicontinuity of g, finiteness of $\varphi(x^*)$ as well as the convergences $c(x^k) \to c(\widehat{x})$ and $z^k \to \widehat{z}$, taking the limit for $k \to \infty$ in (A.2) gives $c(\widehat{x}) = \widehat{z} \in \operatorname{dom} g$. Therefore, $\widehat{x}$ is feasible for (A.1) and local optimality of x^* for (P) implies $\varphi(x^*) \leq \varphi(\widehat{x})$. Furthermore, exploiting (A.2) and the optimality of each $(x^k, z^k) \in \mathbb{R}^n \times \operatorname{dom} g$, we find that $\widehat{x} = x^*$, since

$$\varphi(\widehat{x}) + \frac{1}{2}\|\widehat{x} - x^*\|^2 \leq \liminf_{k \to \infty}\left(f(x^k) + g(z^k) + \frac{\rho_k}{2}\|c(x^k) - z^k\|^2 + \frac{1}{2}\|x^k - x^*\|^2\right)$$
$$\leq \varphi(x^*) \leq \varphi(\widehat{x}).$$

Moreover, noting that (A.2) also gives $f(x^k) + g(z^k) \leq \varphi(x^*)$ for each $k \in \mathbb{N}$, by

$$\varphi(x^*) \leq \liminf_{k \to \infty} f(x^k) + g(z^k) \leq \limsup_{k \to \infty} f(x^k) + g(z^k) \leq \varphi(x^*)$$

we have the attentive convergence of $\{z^k\}_{k \in \mathbb{N}}$ to $\widehat{z} = c(x^*)$, that is, $z^k \xrightarrow{g} c(x^*)$.

Now, due to $x^k \to x^*$ and $z^k \to c(x^*)$, we may assume without loss of generality that $\{x^k\}_{k \in \mathbb{N}}$ and $\{z^k\}_{k \in \mathbb{N}}$ are taken from the interior of $\mathbb{B}_\delta(x^*)$ and $\mathbb{B}_1(c(x^*))$, respectively, as this is eventually the case. Thus, for each $k \in \mathbb{N}$, (x^k, z^k) is an unconstrained (that is, with the constraints being inactive) minimizer of

$$(x, z) \mapsto f(x) + g(z) + \frac{\rho_k}{2}\|c(x) - z\|^2 + \frac{1}{2}\|x - x^*\|^2,$$

whose relevant necessary conditions for optimality read

$$0 = \nabla f(x^k) + \rho_k c'(x^k)^\top [c(x^k) - z^k] + x^k - x^*,$$
$$0 \in \partial g(z^k) - \rho_k [c(x^k) - z^k].$$

Thus, setting $y^k := \rho_k[c(x^k) - z^k]$ for each $k \in \mathbb{N}$, the conditions in (3.5) are a consequence of $x^k \to x^*$ and $z^k \to c(x^*)$. Overall, this shows that x^* is AKKT-stationary for (P). $\qquad\square$

Proof of Proposition 4.3. Let $x^* \in \mathbb{R}^n$ be an arbitrary accumulation point of $\{x^k\}$ and $\{x^k\}_{k \in K}$ a subsequence such that $x^k \to_K x^*$. We shall argue that x^* is feasible for (P), that is, $c(x^*) \in \operatorname{dom} g$. Before considering each assertion, we point out that, if $\{\mu_k\}$ is bounded away from zero, the conditions at Step 7 of Algorithm 4.1 imply that $V^k := \|c(x^k) - z^k\| \to 0$ for $k \to \infty$. Thus, since $x^k \to_K x^*$, by continuity arguments we have also $z^k \to_K c(x^*)$. Finally, let us conveniently denote by $\Delta_Y \in \mathbb{R}$ the finite size of the bounded set $Y \subset \mathbb{R}^m$ initialized in Algorithm 4.1, such that $\|\widehat{y}\| \leq \Delta_Y$ for all $\widehat{y} \in Y$.

(i) The φ-attentive convergence of $\{x^k\}_{k \in K}$ to x^* requires that $x^* \in \operatorname{dom} \varphi$. Hence, by $\varphi := g \circ c$ and continuity of c, it must be $c(x^*) \in \operatorname{dom} g$.

(ii) In view of the preliminary observation regarding $\{\mu_k\}$ bounded away from zero, it remains to consider the case $\mu_k \to 0$ for showing that $z^k \to_K c(x^*)$. Using (4.3) and $z^k \in \mathbf{O}_{\mu_k}(x^k, \hat{y}^k)$, the upper bound yields

$$B \geq \mathscr{L}_{\mu_k}(x^k, \hat{y}^k) = f(x^k) + g(z^k) + \frac{1}{2\mu_k}\|c(x^k) + \mu_k \hat{y}^k - z^k\|^2 - \frac{\mu_k}{2}\|\hat{y}^k\|^2$$

for all $k \in \mathbb{N}$. Rearranging terms and considering $\mu_k \to 0$, we get

$$\|c(x^k) + \mu_k \hat{y}^k - z^k\|^2 \leq 2\mu_k \left[B + \frac{\mu_k}{2}\|\hat{y}^k\|^2 - f(x^k) - g(z^k)\right] \to_K 0,$$

where the limit follows from boundedness of $\{\hat{y}^k\}_{k \in \mathbb{N}}$ and $\{z^k\}_{k \in K}$, the latter remaining bounded by $x^k \to_K x^*$ and local boundedness of the proximal mapping. Using once again that $\mu_k \hat{y}^k \to 0$, the limit implies $c(x^k) - z^k \to_K 0$. Thus, since $x^k \to_K x^*$, by continuity arguments it is $z^k \to_K c(x^*)$. Now, owing to $x^k \to_K x^*$, boundedness of $\{\hat{y}^k\}_{k \in \mathbb{N}}$, and $\mu_k \leq \mu_0$, from the upper bound condition we also obtain that $\{g(z^k)\}_{k \in K}$ remains bounded above:

$$\limsup_{k \to_K \infty} g(z^k) \leq B + \limsup_{k \to_K \infty} \frac{\mu_k}{2}\|\hat{y}^k\|^2 - f(x^k) \leq B + \frac{\mu_0}{2}\Delta_Y^2 - f(x^*) < \infty.$$

Combined with $z^k \to_K c(x^*)$ and lower semicontinuity of g, this implies that $c(x^*) \in \operatorname{dom} g$.

(iii) As $\{\mu_k\}_{k \in \mathbb{N}}$ remains bounded away from zero, we have that $z^k \to_K c(x^*)$. Then, since $z^k \in \mathbf{O}_{\mu_k}(x^k, \hat{y}^k) \subseteq \operatorname{dom} g$ for all $k \in \mathbb{N}$ by Step 4, it follows from $\operatorname{dom} g$ being closed that $c(x^*) \in \operatorname{dom} g$.

(iv) As $\{\mu_k\}_{k \in \mathbb{N}}$ remains bounded away from zero, we have that $z^k \to_K c(x^*)$ and, owing to the conditions at Step 7, $\{\mu_k\}_{k \in \mathbb{N}}$ is asymptotically constant, namely for some finite index $k_* \in \mathbb{N}$ it is $\mu_k = \mu_* > 0$ for all $k \geq k_*$. In turn, the primal infeasibility measure V^k decreases linearly, as it must be $V^k \leq \theta V^{k-1}$ for all $k \geq k_*$. Now, rearranging (4.3) and exploiting the explicit minimization over z, we obtain the upper bound

$$\mathscr{L}_{\mu_*}(x^{k-1}, \hat{y}^k) = f(x^{k-1}) + \min_z \left\{g(z) + \langle \hat{y}^k, c(x^{k-1}) - z \rangle + \frac{1}{2\mu_*}\|c(x^{k-1}) - z\|^2\right\}$$

$$\leq f(x^{k-1}) + g(z^{k-1}) + \langle \hat{y}^k, c(x^{k-1}) - z^{k-1} \rangle + \frac{1}{2\mu_*}\|c(x^{k-1}) - z^{k-1}\|^2$$

$$= \mathscr{L}_{\mu_*}(x^{k-1}, \hat{y}^{k-1}) + \langle \hat{y}^k - \hat{y}^{k-1}, c(x^{k-1}) - z^{k-1} \rangle$$

for all $k > k_*$. Invoking the Cauchy-Schwarz inequality, the descent condition gives

$$\mathscr{L}_{\mu_*}(x^k, \hat{y}^k) \leq \mathscr{L}_{\mu_*}(x^{k-1}, \hat{y}^k) \leq \mathscr{L}_{\mu_*}(x^{k-1}, \hat{y}^{k-1}) + 2\Delta_Y V^{k-1}$$

for all $k > k_*$, where we used that $\|\hat{y}^k - \hat{y}^{k-1}\| \leq 2\Delta_Y$. Then, telescoping this argument, the linear decrease of V^k yields

$$\mathscr{L}_{\mu_*}(x^{k+\ell}, \hat{y}^{k+\ell}) - \mathscr{L}_{\mu_*}(x^{k-1}, \hat{y}^{k-1}) \leq 2\Delta_Y \sum_{m=0}^{\ell} V^{k+m-1} \leq 2\Delta_Y V^{k-1} \sum_{m=0}^{\ell} \theta^\ell$$

for all $k > k_*$ and $\ell \in \mathbb{N}$. Hence, since $\theta \in (0,1)$, for any arbitrary but fixed $k > k_*$ taking the limit results in

$$\limsup_{\ell \to \infty} \mathscr{L}_{\mu_*}(x^{k+\ell}, \hat{y}^{k+\ell}) \leq \mathscr{L}_{\mu_*}(x^{k-1}, \hat{y}^{k-1}) + \frac{2\Delta_\Upsilon}{1-\theta} V^{k-1} < \infty,$$

showing that, up to discarding early iterates, there is an upper bound to $\mathscr{L}_{\mu_k}(x^k, \hat{y}^k)$. Then, feasibility of x^* follows as in (ii). $\square$

Proof of Proposition 4.4. Let $c(x^*) \in \text{int}\, D \neq \emptyset$ and notice that the proximal mapping of δ_D coincides with the projection onto D, namely $\text{prox}_{\gamma \delta_D} = \Pi_D$ for all $\gamma > 0$. Consider the two cases.

- If $\mu_k \to 0$, by boundedness of $\{\hat{y}^k\}_{k \in \mathbb{N}}$ and continuity of c, there exists $k_1 \in K$ such that $c(x^k) + \mu_k \hat{y}^k \in \text{int}\, D$ for all $k \in K$, $k \geq k_1$.
- If $\{\mu_k\}_{k \in \mathbb{N}}$ is bounded away from zero, it follows from Step 7 that $c(x^k) - z^k \to_K 0$. Then, by continuity of c and $x^k \to_K x^*$, $z^k \to_K c(x^*) \in \text{int}\, D$ follows. Hence, there exists $k_2 \in K$ such that $z^k \in \text{int}\, D$ for all $k \in K$, $k \geq k_2$.

In both cases, since $z^k \in \mathbf{O}_{\mu_k}(x^k, \hat{y}^k) = \Pi_D(c(x^k) + \mu_k \hat{y}^k)$ and $z^k \in \text{int}\, D$ for all $k \in K$ large enough, it must be $z^k = c(x^k) + \mu_k \hat{y}^k$ and, consequently, $y^k = 0$ by Step 5. $\square$

Proof of Theorem 4.1. We claim that the subsequences $\{x^k\}_{k \in K}$, $\{y^k\}_{k \in K}$ and $\{z^k\}_{k \in K}$ generated by Algorithm 4.1 satisfy the properties in Definition 3.3 and therefore show that x^* is an AKKT-stationary point for (P). First of all, by Proposition 4.3(i), the φ-attentive convergence of $\{x^k\}_{k \in K}$ implies that x^* is feasible for (P). From Steps 4 and 5 of Algorithm 4.1, we have by ε_k-Υ-stationarity of x^k with certificate z^k that

$$\eta^k := \nabla f(x^k) + c'(x^k)^\top y^k = \nabla_x \mathscr{L}^S_{\mu_k}(x^k, z^k, \hat{y}^k)$$

satisfies $\|\eta^k\| \leq \varepsilon_k$. Then, by construction, we have $\eta^k \to_K 0$. Further, from Step 5 we obtain that, for all $k \in \mathbb{N}$,

$$y^k = \hat{y}^k + \frac{c(x^k) - z^k}{\mu_k} \in \partial g(z^k),$$

where the inclusion follows from $z^k \in \mathbf{O}_{\mu_k}(x^k, \hat{y}^k)$ (in fact, from the necessary condition for the optimality of z^k). It remains to show that $c(x^k) - z^k \to_K 0$. To this end, we consider the two cases.

- If $\{\mu_k\}_{k \in \mathbb{N}}$ is bounded away from zero, the conditions at Step 7 imply that $V^k := \|c(x^k) - z^k\| \to 0$, hence in particular $c(x^k) - z^k \to_K 0$.
- If instead $\mu_k \to 0$, we exploit continuity of c and boundedness of $\{\hat{y}^k\}_{k \in \mathbb{N}}$. These properties result in $c(x^k) + \mu_k \hat{y}^k \to_K c(x^*) \in \text{dom}\, g$, where the inclusion follows from feasibility of x^*. Then, since $z^k \in \mathbf{O}_{\mu_k}(x^k, \hat{y}^k)$ for all $k \in \mathbb{N}$, from (4.2) and [19, Lemma 3.3] we have that

$$z^k \in \text{prox}_{\mu_k g}(c(x^k) + \mu_k \hat{y}^k) \to_K c(x^*)$$

while $\mu_k \to 0$. Thus, the limit $c(x^k) - z^k \to_K 0$ follows.

Overall, this proves that x^* is an AKKT-stationary point for (P). $\square$

Proof of Proposition 4.5. By Proposition 4.3(iii) and dom g being closed, if $\{\mu_k\}_{k\in\mathbb{N}}$ is bounded away from zero, then each accumulation point x^* is feasible for (P), namely it is a global minimizer of the feasibility problem, and so it is an M-stationary point thereof by strict continuity of the objective function [37, Proposition 5.3]. Hence, it remains to consider the case $\mu_k \to 0$.

Let us rewrite the feasibility problem with the equivalent lifted representation

$$\underset{\substack{x\in\mathbb{R}^n \\ (z,\alpha)\in\text{epi}\,g}}{\text{minimize}} \quad \frac{1}{2}\|c(x) - z\|^2 \tag{A.3}$$

obtained by epigraphical reformulation. The inclusion $(z,\alpha) \in \text{epi}\,g$ requires that $g(z) \leq \alpha < \infty$, thus encoding $z \in \text{dom}\,g$. M-stationarity of x^* for (A.3) requires the existence of some $z^* \in \mathbb{R}^m$ such that

$$c'(x^*)^\top [c(x^*) - z^*] = 0, \tag{A.4a}$$

$$(c(x^*) - z^*, 0) \in N_{\text{epi}\,g}^{\lim}(z^*, g(z^*)). \tag{A.4b}$$

Owing to Step 4 of Algorithm 4.1 it is $\|\eta^k\| \leq \varepsilon_k$, where

$$\eta^k := \nabla f(x^k) + \frac{1}{\mu_k} c'(x^k)^\top \left[c(x^k) + \mu_k \hat{y}^k - z^k \right],$$

for all $k \in K$. Taking the limit of $\mu_k \eta^k$ for $k \to_K \infty$, while using $\mu_k \to 0$ and boundedness of $\{\eta^k\}_{k\in\mathbb{N}}$ and $\{\nabla f(x^k)\}_{k\in K}$, it follows that (A.4a) is satisfied. Furthermore, by the minimizing property of $z^k \in \mathbf{O}_{\mu_k}(x^k, \hat{y}^k) = \text{prox}_{\mu_k g}(c(x^k) + \mu_k \hat{y}^k)$, it is $z^k \in \text{dom}\,g$ and

$$\frac{1}{\mu_k} \left[c(x^k) + \mu_k \hat{y}^k - z^k \right] \in \partial g(z^k)$$

for all $k \in \mathbb{N}$. By boundedness of $\{c(x^k) + \mu_k \hat{y}^k\}_{k\in K}$ and local boundedness of the proximal mapping, $\{z^k\}_{k\in K}$ remains bounded too. Hence, since $z^k \in \text{dom}\,g$ for all $k \in \mathbb{N}$ and dom g is closed, we may assume without loss of generality that $z^k \to_K z^*$ for some $z^* \in \text{dom}\,g$. Using the relationship stated in (3.2), multiplying by $\mu_k > 0$ and exploiting that $N_{\text{epi}\,g}^{\lim}(z^k, g(z^k))$ is a cone, we have

$$(c(x^k) + \mu_k \hat{y}^k - z^k, -\mu_k) \in N_{\text{epi}\,g}^{\lim}(z^k, g(z^k))$$

for all $k \in \mathbb{N}$. Taking the limit for $k \to_K \infty$, the robustness of the limiting normal cone [37, Proposition 1.20] and the convergence $z^k \to_K z^* \in \text{dom}\,g$ yield (A.4b), concluding the proof. $\quad\square$