Mathematical Programming (2023) 201:863–896
https://doi.org/10.1007/s10107-022-01922-4

FULL LENGTH PAPER

Series A

Constrained composite optimization and augmented Lagrangian methods

Alberto De Marchi[1] · **Xiaoxi Jia**[2] · **Christian Kanzow**[2] · **Patrick Mehlitz**[3]

Received: 5 April 2022 / Accepted: 29 December 2022 / Published online: 8 February 2023

Abstract
We investigate finite-dimensional constrained structured optimization problems, featuring composite objective functions and set-membership constraints. Offering an expressive yet simple language, this problem class provides a modeling framework for a variety of applications. We study stationarity and regularity concepts, and propose a flexible augmented Lagrangian scheme. We provide a theoretical characterization of the algorithm and its asymptotic properties, deriving convergence results for fully nonconvex problems. It is demonstrated how the inner subproblems can be solved by off-the-shelf proximal methods, notwithstanding the possibility to adopt any solvers, insofar as they return approximate stationary points. Finally, we describe our matrix-free implementation of the proposed algorithm and test it numerically. Illustrative examples show the versatility of constrained composite programs as a modeling tool and expose difficulties arising in this vast problem class.

Keywords Augmented Lagrangian methods · Composite nonconvex optimization · Nonlinear optimization · Nonsmooth optimization

Mathematics Subject Classification 49J53 · 65K05 · 90C30

✉ Alberto De Marchi
alberto.demarchi@unibw.de

1 Department of Aerospace Engineering, Institute of Applied Mathematics and Scientific Computing, Universität der Bundeswehr München, 85577 Neubiberg, Munich, Germany

2 Institute of Mathematics, University of Würzburg, 97074 Würzburg, Germany

3 Institute of Mathematics, Brandenburg University of Technology Cottbus-Senftenberg, 03046 Cottbus, Germany

1 Introduction

In this paper we investigate and develop numerical methods for constrained composite programs, namely finite-dimensional optimization problems of the form

$$\underset{x}{\text{minimize}} \quad q(x) := f(x) + g(x) \qquad \text{subject to} \quad c(x) \in D, \tag{P}$$

where x is the decision variable, f and c are smooth functions, g is proper and lower semicontinuous, and D is a nonempty closed set. We call (P) a constrained composite optimization problem because it contains set-membership constraints and a composite objective function $q := f + g$. Notice that the problem data, namely f, g, c and D, can be nonconvex, the nonsmooth cost term g can be discontinuous and the constraint set D can be disconnected. Thanks to their rich structure and flexibility, constrained composite problems are of interest for modeling in a variety of applications, ranging from optimal and model predictive control [21, 53] to signal processing [19], low-rank and sparse approximation, compressed sensing, cardinality-constrained optimization [10] and disjunctive programming [6], such as problems with complementarity, vanishing and switching constraints [36, 43].

Augmented Lagrangian methods have recently attracted revived and grown interest. Tracing back to the classical work of Hestenes [34] and Powell [48], the augmented Lagrangian framework can tackle large-scale constrained problems. Recent accounts on this topic can be found in [12, 15, 20], among others. Our approach is inspired by the fact that "augmented Lagrangian ideas are independent of the degree of smoothness of the functions that define the problem" [15, §4.1] and lead to a sequence of unconstrained or simply constrained subproblems. Moreover, this framework can handle nonconvex constraints, is often superior to pure penalty methods, enjoys good warm-starting capabilities and allows to avoid ill-conditioning due to a pure penalty approach as well as to deal with constraints without softening them; cf. [53, 56]. In the context of constrained composite programming, the augmented Lagrangian subproblems associated with (P) may, again, be of composite type but possess, if at all, comparatively simple constraints. Exemplary, these subproblems can be solved with the aid of proximal methods, inaugurated by Moreau [45], which can handle nonsmooth, nonconvex and extended real-valued cost functions; cf. [19, 37, 46, 58] for recent contributions.

The close relationship between augmented Lagrangian and proximal methods is well known and traces back to Rockafellar [49]. These approaches have been combined in [25] to deal with unconstrained, composite optimization problems whose nonsmooth term is convex and possibly composed with a linear operator. Following this strategy, the proximal augmented Lagrangian method has been considered for constrained composite programs in [22, Ch. 1], however lacking of sound theoretical support and convergence analysis. A first step for resolving these shortcomings is constituted by proximal gradient methods that can cope with *local* Lipschitz continuity of the smooth cost gradient, only recently investigated in the Euclidean setting, see [24, 37]. By relying on an adaptive stepsize selection rule for the proximal gradient

oracle, these algorithms can be adopted as inner solver for augmented Lagrangian subproblems arising from general nonlinear constraints.

Another issue originates from the following observation. One can reformulate the original problem, by introducing slack variables, in order to have a set-membership constraint with a convex right-hand side; consider this problem equipped with slack variables and the associated augmented Lagrangian function. The proximal augmented Lagrangian function characterizes the latter one on the manifold corresponding to the explicit minimization over the slack variables [25, 49]. This procedure is employed to eliminate the slack variables and, in the convex setting, to obtain a continuously differentiable function. Although the same ideas apply to (P), the resulting proximal augmented Lagrangian does not exhibit this favorable property in the fully nonconvex setting. In particular, this lack of regularity is due to the set-valued projection onto the constraint set D.

The contribution of this work touches several aspects. We investigate the abstract class of constrained composite optimization problems in the fully nonconvex setting and discuss relevant stationarity concepts. Then, we present an algorithm for the numerical solution of these problems and, considering a classical (safeguarded) augmented Lagrangian scheme, we provide a comprehensive yet compact global convergence analysis. Patterning this methodology, analogous algorithms and theoretical results can be derived based on other augmented Lagrangian schemes. Further, we demonstrate that there is no need for special choices of possibly set-valued projections and proximal mappings since we rely on the aforementioned reformulation of (P) with slack variables and keep them within our algorithmic framework. It is carved out that, apart from the higher number of decision variables, this reformulation is nonhazardous. We show that it is possible to adopt off-the-shelf, yet adaptive, proximal gradient methods for solving the augmented Lagrangian subproblems. Finally, some numerical experiments visualize computational features of our algorithmic approach.

The following blanket assumptions are considered throughout, without further mention. Technical definitions are given in Sect. 2.1.

Assumption I 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 with locally Lipschitz continuous derivatives;
- (ii) $g\colon \mathbb{R}^n \to \overline{\mathbb{R}}$ is proper, lower semicontinuous and prox-bounded;
- (iii) $D \subset \mathbb{R}^m$ is a nonempty and closed set.

Notice that the consequential theory remains valid whenever $\mathbb{R}^n$ and $\mathbb{R}^m$ are replaced by finite-dimensional Hilbert spaces $\mathbb{X}$ and $\mathbb{Y}$. Moreover, the local Lipschitz continuity in Assumption I(i) is actually superfluous for the augmented Lagrangian framework, but sufficient to solve the arising inner problems via proximal gradient methods [24, 37].

By Assumptions I(i) and I(ii), the cost function $q := f + g$ has nonempty domain, that is, $\operatorname{dom} q \neq \emptyset$. Similarly, Assumption I(iii) guarantees that it is always possible to project onto the constraint set D. Nevertheless, these conditions do not imply the existence of feasible points for (P); in fact, the projection onto the set $\{x \in \mathbb{R}^n \mid c(x) \in D\}$ induced by the constraints $c(x) \in D$ can be as difficult as the original problem (P). As it is the case in nonlinear programming [15], we will study the minimization

properties of the augmented Lagrangian scheme with respect to some infeasibility measure.

Finally, we should mention that, for our actual implementation, we work under the practical assumption that (only) the following computational oracles are available or simple to evaluate:

- cost function value $f(x)$ and gradient $\nabla f(x)$, given $x \in \operatorname{dom} q$;
- (arbitrary) proximal point $z \in \operatorname{prox}_{\gamma g}(x)$ and function value $g(z)$ therein, given $x \in \mathbb{R}^n$ and $\gamma \in (0, \gamma_g)$, γ_g being the prox-boundedness threshold of g;
- constraint function value $c(x)$ and Jacobian-vector product $\nabla c(x)^\top v$, given $x \in \operatorname{dom} q$ and $v \in \mathbb{R}^m$;
- (arbitrary) projected point $z \in \Pi_D(v)$, given $v \in \mathbb{R}^m$.

Relying only on these oracles, the method considered for our numerical examples is first-order and matrix-free by construction; as such, it involves only simple operations and has low memory footprint.

1.1 Related work

Augmented Lagrangian schemes have been extensively investigated [12, 15, 20, 53], also in the infinite-dimensional setting [4, 16, 38].

Merely lower semicontinuous cost functions have been considered in [26]. Inspired by [31, Alg. 1] and leveraging the idea behind [15, Ex. 4.12], the convergence properties of [26, Alg. 1] hinge on the upper boundedness of the augmented Lagrangian along the iterates ensured by the initialization at a feasible point. Although possible in some cases, in general finding a feasible starting point can be as hard as the original problem. We deviate in this respect, seeking instead a method able to start from any $x^0 \in \mathbb{R}^n$. Nonetheless, if a feasible point is readily available for (P), one can adopt [26, Alg. 1] in its original form, replacing the augmented Lagrangian function and inner solver accordingly. In this case, and possibly assuming lower boundedness of the cost function q, stronger convergence guarantees can be obtained.

Programs with geometric constraints have been studied in [16, 36] and, for the special case of so-called complementarity constraints, in [32]. These have a continuously differentiable cost function f and set-membership constraints of the form $c(x) \in C$, $x \in D$, with D as in Assumption I(iii) and C nonempty, closed and *convex*. As already mentioned, similar structure can be obtained from (P) by introducing slack variables. Moreover, as pointed out in [36, §5.4], considering a lower semicontinuous functional $q := f + g$ does not enlarge the problem class, since there is an equivalent, yet smooth, reformulation in terms of the epigraph of g. These observations imply that constrained composite programs do not generalize the problem class considered in [36]. Nevertheless, the necessary reformulations come at a price: increased problem size due to slack variables and the need for projections onto the epigraph of g. The augmented Lagrangian method we are about to present is designed around (P) in the fully nonconvex setting. Hence, it natively handles nonsmooth cost functions, nonlinear constraints and nonconvex sets, with no need for oracles other than those mentioned above. Analogous considerations hold for [18], dedicated to an augmented

Lagrangian method for non-Lipschitz nonlinear programs, and [39, §6.2], where the solution of the augmented Lagrangian subproblems is not discussed.

The work presented in this paper collects and builds upon some ideas put forward in [22]. However, we consider different stationarity concepts and necessary optimality conditions, not based on the proximal operator as in [22, §1.2], but rather exploiting tools from variational analysis; see [33, 36, 39, 42]. Furthermore, by avoiding the marginalization approach of [22, §1.4] and so maintaining the slack variables explicit, we can offer rigorous convergence guarantees for the subproblems [24, 37], transcending the dubious justifications given in [22, §1.5.4].

2 Notation and fundamentals

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

2.1 Preliminaries

With $\mathbb{R}$ and $\overline{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$ we denote the real and extended real line, respectively. Furthermore, let $\mathbb{R}_+$ and $\mathbb{R}_{++}$ be the nonnegative and positive real numbers, respectively. We use 0 in order to represent the scalar zero as well as the zero vector of appropriate dimension. The vector in $\mathbb{R}^n$ with all elements equal to 1 is denoted by 1_n. 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 $\bar{x} \in \mathbb{R}^n$.

Given a proper and lsc function $h\colon \mathbb{R}^n \to \overline{\mathbb{R}}$ and a point $\bar{x} \in \operatorname{dom} h$, we may avoid to assume h continuous and instead appeal to *h-attentive* convergence of a sequence $\{x^k\}$:

$$x^k \xrightarrow{h} \bar{x} \quad :\Leftrightarrow \quad x^k \to \bar{x} \quad \text{with} \quad h(x^k) \to h(\bar{x}). \tag{2.1}$$

Following [50, Def. 8.3], we denote by $\hat{\partial} h\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ the *regular subdifferential* of h, where

$$v \in \hat{\partial} h(\bar{x}) \quad :\Leftrightarrow \quad \liminf_{\substack{x\to\bar{x}\\ x\neq\bar{x}}} \frac{h(x) - h(\bar{x}) - \langle v, x - \bar{x}\rangle}{\|x - \bar{x}\|} \geq 0. \tag{2.2}$$

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 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 $\bar{x}$ [50, Ex. 8.8]. For formal completeness, we set $\hat{\partial} h(\bar{x}) := \partial h(\bar{x}) := \emptyset$ for each $\bar{x} \notin \operatorname{dom} h$.

With respect to the minimization of h, we say that $x^* \in \operatorname{dom} h$ is *stationary* if $0 \in \partial h(x^*)$, which constitutes a necessary condition for the optimality of x^* [50, Thm 10.1]. Furthermore, we say that $x^* \in \mathbb{R}^n$ is ε-*stationary* for some $\varepsilon \geq 0$ if

$$\exists \eta \in \partial h(x^*)\colon \|\eta\| \leq \varepsilon. \tag{2.3}$$

A mapping $S\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ is *locally bounded* at a point $\bar{x} \in \mathbb{R}^n$ if for some neighborhood V of $\bar{x}$ the set $S(V) \subset \mathbb{R}^m$ is bounded [50, Def. 5.14]; it is called locally bounded (on $\mathbb{R}^n$) if this holds at every $\bar{x} \in \mathbb{R}^n$. If $S(\bar{x})$ is nonempty, we define the *outer limit* of S at $\bar{x}$ by means of

$$\limsup_{x\to\bar{x}} S(x) := \{y \in \mathbb{R}^m \big| \exists x^k \to \bar{x},\ \exists y^k \to y,\ y^k \in S(x^k)\ \forall k \in \mathbb{N}\}$$

and note that this is a closed superset of $S(\bar{x})$ by definition.

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

$$\mathrm{prox}_{\gamma h}(x) := \mathrm{argmin}_z \left\{ h(z) + \frac{1}{2\gamma}\|z - x\|^2 \right\},$$

and 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 $\gamma > 0$. 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 $\mathrm{prox}_{\gamma h}$ is locally bounded, nonempty- and compact-valued [50, Thm 1.25].

Some tools of variational analysis will be exploited in order to describe the geometry of the nonempty, closed, but not necessarily convex set $D \subset \mathbb{R}^m$, appearing in the formulation of (P). The *projection* mapping Π_D and the *distance* function dist_D are defined by

$$\Pi_D(v) := \mathrm{argmin}_{z\in D} \|z - v\| \quad \text{and} \quad \mathrm{dist}_D(v) := \inf_{z\in D} \|z - v\|.$$

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

The *indicator* function of a set $D \subset \mathbb{R}^m$ is the function $\delta_D\colon \mathbb{R}^m \to \overline{\mathbb{R}}$ defined as $\delta_D(v) = 0$ 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

$$\mathcal{N}_D^{\lim}(z) := \limsup_{v\to z} \ \mathrm{cone}\,(v - \Pi_D(v))\,.$$

For $\tilde{z} \notin D$, we formally set $\mathcal{N}_D^{\lim}(\tilde{z}) := \emptyset$. The limiting normal cone is robust in the following sense:

$$\mathcal{N}_D^{\lim}(z) = \limsup_{v\to z} \mathcal{N}_D^{\lim}(v).$$

Observe that, for all $v, z \in \mathbb{R}^m$, we have the implication

$$z \in \Pi_D(v) \quad \Rightarrow \quad v - z \in \mathcal{N}_D^{\lim}(z), \tag{2.4}$$

and the converse implication holds, exemplary, if D is convex. For any proper and lsc function $h\colon \mathbb{R}^n \to \overline{\mathbb{R}}$ and a point $\bar{x}$ with $h(\bar{x})$ finite, we have

$$\partial h(\bar{x}) = \left\{ v \in \mathbb{R}^n \middle| (v,-1) \in \mathcal{N}^{\lim}_{\operatorname{epi} h}(\bar{x}, h(\bar{x})) \right\}$$

where $\operatorname{epi} h := \{(x,\alpha) \in \mathbb{R}^n \times \mathbb{R} \big| h(x) \le \alpha\}$ denotes the epigraph of h.

Lemma 2.1 *Let $D \subset \mathbb{R}^m$ be nonempty, closed and convex. Furthermore, let $c\colon \mathbb{R}^n \to \mathbb{R}^m$ be continuously differentiable. We consider the function $\vartheta\colon \mathbb{R}^n \to \mathbb{R}$ given by $\vartheta(x) := \frac{1}{2}\operatorname{dist}_D^2(c(x))$ for all $x \in \mathbb{R}^n$. Then, ϑ is continuously differentiable, and for each $\bar{x} \in \mathbb{R}^n$, we have*

$$\nabla\vartheta(\bar{x}) = \nabla c(\bar{x})^\top \big(c(\bar{x}) - \Pi_D(c(\bar{x}))\big).$$

Proof We define $\psi\colon \mathbb{R}^m \to \mathbb{R}$ by means of $\psi(y) := \frac{1}{2}\operatorname{dist}_D^2(y)$ for all $y \in \mathbb{R}^m$ and observe that $\vartheta = \psi \circ c$. Since D is assumed to be convex, ψ is continuously differentiable with gradient $\nabla\psi(\bar{y}) = \bar{y} - \Pi_D(\bar{y})$ for each $\bar{y} \in \mathbb{R}^m$, see [8, Cor. 12.30], and the statements of the lemma follow trivially from the standard chain rule. □

2.2 Stationarity concepts and qualification conditions

We now define some basic concepts and discuss stationarity conditions for (P). As the cost function $q := f + g$ is possibly extended real-valued, feasibility of a point must account for its domain.

Definition 2.2 (*Feasibility*) A point $x^* \in \mathbb{R}^n$ is called *feasible* for (P) if $x^* \in \operatorname{dom} q$ and $c(x^*) \in D$.

Working under the assumption that the constraint set D is nonconvex, a plausible stationarity concept for addressing (P) is that of Mordukhovich-stationarity, which exploits limiting normals to D; cf. [42, §3] and [44, Thm 5.48].

Definition 2.3 (*M-stationarity*) Let $x^* \in \mathbb{R}^n$ be a feasible point for (P). Then, x^* is called a *Mordukhovich-stationary* point of (P) if there exists a multiplier $y^* \in \mathbb{R}^m$ such that

$$-\nabla c(x^*)^\top y^* \in \partial q(x^*), \tag{2.5a}$$

$$y^* \in \mathcal{N}_D^{\lim}(c(x^*)). \tag{2.5b}$$

Notice that these conditions implicitly require the feasibility of x^*, for otherwise the subdifferential and limiting normal cone would be empty. Note that this definition coincides with the usual KKT conditions of (P) if g is smooth and D is a convex set.

Subsequently, we study an asymptotic counterpart of this definition. In case where q is locally Lipschitz continuous, one could apply the notions from [36, §2.2] and [42, §5.1] for that purpose. However, since g is assumed to be merely lsc, we need to adjust these concepts at least slightly.

Definition 2.4 (*AM-stationarity*) Let $x^* \in \mathbb{R}^n$ be a feasible point for (P). Then, x^* is called an *asymptotically M-stationary* point of (P) if there exist sequences $\{x^k\}$, $\{\eta^k\} \subset \mathbb{R}^n$ and $\{y^k\}$, $\{\zeta^k\} \subset \mathbb{R}^m$ such that $x^k \xrightarrow{q} x^*$, $\eta^k \to 0$, $\zeta^k \to 0$ and

$$-\nabla c(x^k)^\top y^k + \eta^k \in \partial q(x^k), \tag{2.6a}$$

$$y^k \in \mathcal{N}_D^{\lim}(c(x^k) + \zeta^k) \tag{2.6b}$$

for all $k \in \mathbb{N}$.

The definition of an AM-stationary point is similar to the notion of an asymptotic KKT point [15], as well as the meaning of the iterates x^k and the Lagrange multipliers y^k. Notice that Definition 2.4 does not require the sequence $\{y^k\}$ to converge. The vector η^k measures the dual infeasibility, namely the inexactness in the stationarity condition (2.6a) at x^k and y^k. The vector ζ^k is introduced to account for the fact that the condition $c(x^k) \in D$ can be violated along the iterates, though it (hopefully) holds asymptotically. As the corresponding (limiting) normal cone $\mathcal{N}_D^{\lim}(c(x^k))$ would be empty in this case, it would not be possible to satisfy the inclusion $y^k \in \mathcal{N}_D^{\lim}(c(x^k))$. The sequence $\{\zeta^k\}$ remedies this issue and gives a measure of primal infeasibility, as we will attest. Finally, the convergence $x^k \xrightarrow{q} x^*$, which is not restrictive in situations where g is continuous (relative to its domain), will be important later on when taking the limit in (2.6a) since we aim to recover the limiting subdifferential of the objective function as stated in (2.3). Let us note that a slightly different notion of asymptotic stationarity has been introduced for rather general optimization problems in Banach spaces in [39, Def. 6.4, Rem. 6.5]. Therein, different primal sequences are used for the objective function and the constraints.

A local minimizer for (P) is M-stationary only under validity of a suitable qualification condition, which, by non-Lipschitzness of g, will depend on the latter function as well, see [33] for a discussion. However, we can show that each local minimizer of (P) is always AM-stationary. Related results can be found in [39, Thm 6.2] and [42, §5.1].

Proposition 2.5 *Let $x^* \in \mathbb{R}^n$ be a local minimizer for* (P)*. Then, x^* is an AM-stationary point for* (P).

Proof By local optimality of x^* for (P), we find some $\varepsilon > 0$ such that $q(x) \geq q(x^*)$ is valid for all $x \in \mathbb{B}_\varepsilon(x^*) := \left\{x \in \mathbb{R}^n \middle| \|x - x^*\| \leq \varepsilon\right\}$ which are feasible for (P). Consequently, x^* is the uniquely determined global minimizer of

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & q(x) + \frac{1}{2}\|x - x^*\|^2 \\ \text{subject to} \quad & c(x) \in D, \quad x \in \mathbb{B}_\varepsilon(x^*). \end{aligned} \tag{2.7}$$

Let us now consider the penalized surrogate problem

$$\begin{aligned} \underset{x,s}{\text{minimize}} \quad & q(x) + \frac{k}{2}\|c(x) - s\|^2 + \frac{1}{2}\|x - x^*\|^2 \\ \text{subject to} \quad & x \in \mathbb{B}_\varepsilon(x^*), \quad s \in D \cap \mathbb{B}_1(c(x^*)) \end{aligned} \tag{P(k)}$$

where $k \in \mathbb{N}$ is arbitrary. Noting that the objective function of this optimization problem is lsc while its feasible set is nonempty and compact, it possesses a global minimizer $(x^k, s^k) \in \mathbb{R}^n \times \mathbb{R}^m$ for each $k \in \mathbb{N}$. Without loss of generality, we assume $x^k \to \tilde{x}$ and $s^k \to \tilde{s}$ for some $\tilde{x} \in \mathbb{B}_\varepsilon(x^*)$ and $\tilde{s} \in D \cap \mathbb{B}_1(c(x^*))$.

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

$$q(x^k) + \frac{k}{2}\|c(x^k) - s^k\|^2 + \frac{1}{2}\|x^k - x^*\|^2 \leq q(x^*) \tag{2.8}$$

for each $k \in \mathbb{N}$. Using lower semicontinuity of q as well as the convergences $c(x^k) \to c(\tilde{x})$ and $s^k \to \tilde{s}$, taking the limit $k \to \infty$ in (2.8) gives $c(\tilde{x}) = \tilde{s} \in D$. Particularly, $\tilde{x}$ is feasible for (2.7). Therefore, the local optimality of x^* implies $q(x^*) \leq q(\tilde{x})$. Furthermore, we find

$$\begin{aligned} q(\tilde{x}) + \frac{1}{2}\|\tilde{x} - x^*\|^2 &\leq \liminf_{k\to\infty} \left(q(x^k) + \frac{k}{2}\|c(x^k) - s^k\|^2 + \frac{1}{2}\|x^k - x^*\|^2 \right) \\ &\leq q(x^*) \leq q(\tilde{x}). \end{aligned}$$

Hence, $\tilde{x} = x^*$, and noting that (2.8) gives $q(x^k) \leq q(x^*)$ for each $k \in \mathbb{N}$,

$$q(x^*) \leq \liminf_{k\to\infty} q(x^k) \leq \limsup_{k\to\infty} q(x^k) \leq q(x^*),$$

i.e., $x^k \xrightarrow{q} x^*$ follows.

Due to $x^k \to x^*$ and $s^k \to c(x^*)$, we may assume without loss of generality that $\{x^k\}$ and $\{s^k\}$ are taken from the interior of $\mathbb{B}_\varepsilon(x^*)$ and $\mathbb{B}_1(c(x^*))$, respectively. Thus, for each $k \in \mathbb{N}$, (x^k, s^k) is an unconstrained local minimizer of

$$(x, s) \mapsto q(x) + \frac{k}{2}\|c(x) - s\|^2 + \frac{1}{2}\|x - x^*\|^2 + \delta_D(s).$$

Let us introduce $\theta\colon \mathbb{R}^n \times \mathbb{R}^m \to \overline{\mathbb{R}}$ by means of $\theta(x, s) := g(x) + \delta_D(s)$ for each pair $(x, s) \in \mathbb{R}^n \times \mathbb{R}^m$. Applying [44, Prop. 1.107 and 1.114], we find

$$(0, 0) \in \big(\nabla f(x^k) + k\,\nabla c(x^k)^\top (c(x^k) - s^k) + x^k - x^*, k(s^k - c(x^k))\big) + \partial\theta(x^k, s^k)$$

for each $k \in \mathbb{N}$. The decoupled structure of θ and [44, Thm 3.36] yield the inclusion $\partial\theta(x^k, s^k) \subset \partial g(x^k) \times \mathcal{N}_D^{\lim}(s^k)$ for each $k \in \mathbb{N}$. Thus, setting $\eta^k := x^* - x^k$, $y^k := k(c(x^k) - s^k)$ and $\zeta^k := s^k - c(x^k)$ for each $k \in \mathbb{N}$ while observing that $\partial q(x^k) = \nabla f(x^k) + \partial g(x^k)$ holds, we have shown that x^* is AM-stationary for (P). □

In order to guarantee that local minimizers for (P) are not only AM- but already M-stationary, the presence of a qualification condition is necessary. The subsequent definition generalizes the constraint qualification from [42, §3.2] to the

non-Lipschitzian setting and is closely related to the so-called *uniform qualification condition* introduced in [39, Def. 6.8].

Definition 2.6 (*AM-regularity*) Let $x^* \in \mathbb{R}^n$ be a feasible point for (P). Define the set-valued mapping $\mathcal{M}\colon \mathbb{R}^n \times \mathbb{R}^m \rightrightarrows \mathbb{R}^n$ by

$$\mathcal{M}(x, z) := \partial g(x) + \nabla c(x)^\top \mathcal{N}_D^{\lim}(c(x) - z).$$

Then, x^* is called *asymptotically M-regular* for (P) if

$$\limsup_{\substack{x \xrightarrow{g} x^* \\ z \to 0}} \mathcal{M}(x, z) \subset \mathcal{M}(x^*, 0).$$

Let us point the reader's attention to the fact that AM-regularity is not a constraint qualification for (P) in the narrower sense since it depends explicitly on the objective function. However, note that AM-regularity of some feasible point $x^* \in \mathbb{R}^n$ for (P) reduces to

$$\limsup_{\substack{x \to x^* \\ z \to 0}} \nabla c(x)^\top \mathcal{N}_D^{\lim}(c(x) - z) \subset \nabla c(x^*)^\top \mathcal{N}_D^{\lim}(c(x^*)) \tag{2.9}$$

whenever g is locally Lipschitz continuous around x^* since $x \rightrightarrows \partial g(x)$ is locally bounded at x^* in this case, see [44, Cor. 1.81]. We also observe that (2.9) corresponds to the concept of AM-regularity which has been used in [36, 42] where q is assumed to be at least locally Lipschitz continuous, and this condition has been shown to serve as a comparatively weak constraint qualification. Sufficient conditions for the validity of the more general qualification condition from Definition 2.6 can be distilled in a similar way as in [39].

As a corollary of Proposition 2.5, we find the following result, along the lines of [39, Prop. 6.9].

Corollary 2.7 *Let $x^* \in \mathbb{R}^n$ be an AM-regular AM-stationary point for* (P)*. Then, x^* is an M-stationary point for* (P)*. Particularly, each AM-regular local minimizer for* (P) *is M-stationary.*

Following the lines of the proofs of [3, Thm 3.2] or [16, Thm 4.6], it is even possible to show that whenever, for each continuously differentiable function f, AM-stationarity of a feasible point $x^* \in \mathbb{R}^n$ of (P) already implies M-stationarity of x^*, then x^* must be AM-regular. Relying on the terminology coined in [3], this means that AM-regularity is the weakest *strict* qualification condition associated with AM-stationarity.

3 Augmented Lagrangian method

Constrained minimization problems such as (P) are amenable to be addressed by means of augmented Lagrangian methods. Introducing the slack variable $s \in \mathbb{R}^m$, (P) can be rewritten as

$$\underset{x,\, s}{\text{minimize}} \quad q(x) \qquad \text{subject to} \quad c(x) - s = 0, \quad s \in D. \tag{P$_S$}$$

Notice that (P$_S$) is a particular problem in the form of (P). Moreover, if g is smooth, and thus so is q, then (P$_S$) falls into the problem class analyzed in [36]. Note that $x^* \in \mathbb{R}^n$ is a global (local) minimizer of (P) if and only if $(x^*, c(x^*))$ is a global (local) minimizer of (P$_S$). Similarly, the M-stationary points of (P) and (P$_S$) correspond to each other. An elementary calculation additionally reveals that even the AM-stationary points of (P) and (P$_S$) can be identified with each other.

Lemma 3.1 *A feasible point $x^* \in \mathbb{R}^n$ of (P) is AM-stationary for (P) if and only if $(x^*, c(x^*))$ is AM-stationary for (P$_S$).*

Proof The implication $\Rightarrow$ is obvious, so let us only prove the converse one. If $(x^*, c(x^*))$ is AM-stationary for (P$_S$), we find sequences $\{x^k\}, \{\eta_1^k\} \subset \mathbb{R}^n$ and $\{s^k\}, \{y_1^k\}, \{y_2^k\}, \{\eta_2^k\}, \{\zeta_1^k\}, \{\zeta_2^k\} \subset \mathbb{R}^m$ such that $x^k \overset{q}{\to} x^*$, $s^k \to c(x^*)$, $\eta_i^k \to 0$, $\zeta_i^k \to 0$, $i = 1, 2$, and

$$-\nabla c(x^k)^\top y_1^k + \eta_1^k \in \partial q(x^k), \tag{3.1a}$$

$$y_1^k - y_2^k + \eta_2^k = 0, \tag{3.1b}$$

$$c(x^k) - s^k + \zeta_1^k = 0, \tag{3.1c}$$

$$y_2^k \in \mathcal{N}_D^{\lim}(s^k + \zeta_2^k) \tag{3.1d}$$

for all $k \in \mathbb{N}$, where we already used the Cartesian product rule for the limiting normal cone, cf. [44, Prop. 1.2], in order to split

$$(y_1^k, y_2^k) \in \mathcal{N}_{\{0\}\times D}^{\lim}(c(x^k) - s^k + \zeta_1^k, s^k + \zeta_2^k)$$

into (3.1c) and (3.1d). Now, for each $k \in \mathbb{N}$, set $y^k := y_2^k$, $\eta^k := \nabla c(x^k)^\top \eta_2^k + \eta_1^k$ and $\zeta^k := s^k - c(x^k) + \zeta_2^k$. Then, (2.6a) follows from (3.1a) and (3.1b). Furthermore, (2.6b) can be distilled from (3.1d). The convergence $\eta^k \to 0$ is clear from continuous differentiability of c, and $\zeta^k \to 0$ follows from $c(x^k) - s^k \to 0$ which is a consequence of the continuity of c (or (3.1c)). □

Summarizing the above observations, the way we incorporated the slack variable in (P$_S$) does not change the solution and stationarity behavior when compared with (P). In light of [11], where similar issues are discussed in a much broader context, this is remarkable. We use the lifted reformulation (P$_S$) as a theoretical tool to develop our approach for solving (P) and investigate its properties. For some penalty parameter $\mu > 0$, let us define the μ-augmented Lagrangian function $\mathcal{L}_\mu^S : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^m \to \overline{\mathbb{R}}$ associated to (P$_S$) as

$$\begin{aligned}\mathcal{L}_\mu^S(x, s, y) &:= q(x) + \delta_D(s) + \langle y, c(x) - s\rangle + \frac{1}{2\mu}\|c(x) - s\|^2 \\ &= q(x) + \delta_D(s) + \frac{1}{2\mu}\|c(x) + \mu y - s\|^2 - \frac{\mu}{2}\|y\|^2. \end{aligned} \tag{3.2}$$

Observe that, by adopting the indicator δ_D, the constraint $s \in D$ is considered hard, in the sense that it must be satisfied exactly. These simple, nonrelaxable lower-level constraints have been discussed, e.g., in [1, 15, 20, 36]. For later use, let us compute the subdifferential of $\mathcal{L}_\mu^{\mathrm{S}}$ with respect to the variables x and s:

$$\partial_x \mathcal{L}_\mu^{\mathrm{S}}(x, s, y) = \partial q(x) + \frac{1}{\mu} \nabla c(x)^\top [c(x) + \mu y - s], \tag{3.3a}$$

$$\partial_s \mathcal{L}_\mu^{\mathrm{S}}(x, s, y) = \mathcal{N}_D^{\lim}(s) - \frac{1}{\mu} [c(x) + \mu y - s]. \tag{3.3b}$$

The algorithm we are about to present requires, at each inner iteration, the (approximate) minimization of $\mathcal{L}_\mu^{\mathrm{S}}(\cdot, \cdot, y)$, given some $\mu > 0$ and $y \in \mathbb{R}^m$, while in each outer iteration, μ and y are updated. This nested-loops structure naturally arises in the augmented Lagrangian framework, as it does more generally in nonlinear programming.

A similar method can be obtained by exploiting the structure arising from the original problem (P) in order to eliminate the slack variable s, on the vein of the proximal augmented Lagrangian approach [22, 25]. Given some $\mu > 0$, $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$, the explicit minimization of $\mathcal{L}_\mu^{\mathrm{S}}(x, \cdot, y)$ is readily obtained and yields a set-valued mapping:

$$\operatorname{argmin}_s \mathcal{L}_\mu^{\mathrm{S}}(x, s, y) = \Pi_D \left(c(x) + \mu y \right). \tag{3.4}$$

Evaluating the augmented Lagrangian on the set corresponding to the explicit minimization over the slack variable s, we obtain the (single-valued) augmented Lagrangian function $\mathcal{L}_\mu : \mathbb{R}^n \times \mathbb{R}^m \to \overline{\mathbb{R}}$ associated to (P):

$$\mathcal{L}_\mu(x, y) := \min_s \mathcal{L}_\mu^{\mathrm{S}}(x, s, y) = q(x) + \frac{1}{2\mu} \operatorname{dist}_D^2(c(x) + \mu y) - \frac{\mu}{2} \|y\|^2. \tag{3.5}$$

Then, one may consider replacing the minimization of $\mathcal{L}_\mu^{\mathrm{S}}(\cdot, \cdot, y)$ with that of $\mathcal{L}_\mu(\cdot, y)$. Following the lines of [11, §4.1], one can easily check that the problems $\min \mathcal{L}_\mu(\cdot, y)$ and $\min \mathcal{L}_\mu^{\mathrm{S}}(\cdot, \cdot, y)$ are equivalent in the sense that x^* is a local (global) minimizer of $\min \mathcal{L}_\mu(\cdot, y)$ if and only if (x^*, s^*), for each $s^* \in \operatorname{argmin} \mathcal{L}_\mu^{\mathrm{S}}(x^*, \cdot, y)$, is a local (global) minimizer of $\mathcal{L}_\mu^{S}(\cdot, \cdot, y)$; cf. (3.4). However, we highlight that the term $\operatorname{dist}_D^2 : \mathbb{R}^m \to \mathbb{R}$ is not continuously differentiable in general, as the projection onto D is a set-valued mapping, thus making this approach difficult in practice.

Remark 3.1 Whenever D is a convex set, the augmented Lagrangian function $\mathcal{L}_\mu$ from (3.5) is a continuously differentiable function with a locally Lipschitz continuous gradient; cf. Lemma 2.1. Following the literature, see e.g. [1, 15, 25], one can directly augment the corresponding set-membership constraints within the corresponding augmented Lagrangian framework without the need of an additional slack variable. In practical implementations of an augmented Lagrangian framework addressing (P), it is, thus, recommendable to treat only the difficult set-membership constraints with a nonconvex right-hand side with the aid of the lifting approach discussed here. The remaining set-membership constraints can either be augmented without slacks

Algorithm 1 Augmented Lagrangian method for (P)

INITIALIZE Select $\mu_0 > 0, \theta, \kappa \in (0, 1)$ and $Y \subset \mathbb{R}^m$ nonempty bounded

For $k = 0, 1, 2 \ldots$

1.1: Select $\hat{y}^k \in Y$ and $\varepsilon_k \geq 0$

1.2: Compute an ε_k-stationary point $(x^k, s^k) \in \mathbb{R}^n \times D$ of $\mathcal{L}^{\mathrm{S}}_{\mu_k}(\cdot, \cdot, \hat{y}^k)$

1.3: Set $y^k \leftarrow \hat{y}^k + [c(x^k) - s^k]/\mu_k$

1.4: IF $k = 0$ or $\|c(x^k) - s^k\| \leq \theta \|c(x^{k-1}) - s^{k-1}\|$ THEN

1.5: Set $\mu_{k+1} \leftarrow \mu_k$

1.6: ELSE

1.7: Select $\mu_{k+1} \in (0, \kappa\mu_k]$

or remain explicitly in the constraint set of the augmented Lagrangian subproblems if simple enough (like box constraints).

The following Sect. 3.1 contains a detailed statement of our algorithmic framework, whose convergence analysis is presented in Sect. 3.2. Then, suitable termination criteria are discussed in Sect. 3.3. In Sect. 3.4 we consider the numerical solution of the augmented Lagrangian subproblems.

3.1 Algorithm

This section presents an augmented Lagrangian method for the solution of constrained composite programs of the form (P), under Assumption 1. As the augmented Lagrangian constitutes a framework, rather than a single algorithm, several methods have been presented in the past decades, expressing the foundational ideas in different flavors. Some prominent contributions are those in [12, 15, 20, 31, 38, 53], and for primal-dual methods [30]. In the following, we focus on a safeguarded augmented Lagrangian scheme inspired by [15, Alg. 4.1] and investigate its convergence properties. Compared to the classical augmented Lagrangian or multiplier penalty approach for the solution of nonlinear programs [12], this variant uses a safeguarded update rule for the Lagrange multipliers and has stronger global convergence properties. Although we restrict our analysis to this specific algorithm, analogous results can be obtained for others with minor changes. The overall method is stated in Algorithm 1 and corresponds to the popular augmented Lagrangian solver Algencan from [1] applied to ($\mathrm{P_S}$). Let us mention, however, that the analysis in [1] does neither cover composite objective functions $q := f + g$ nor constraints of the form $c(x) \in D$ with potentially nonconvex constraint set D.

First of all, a primal-dual starting point is not explicitly required. In practice, however, the subproblems at step 1.2 should be solved starting from the current primal estimate x^{k-1} paired with some s^{k-1}, preferably an element of $\Pi_D(c(x^{k-1}) + \mu_k \hat{y}^k)$ as suggested by (3.4), thus exploiting initial guesses. The safeguarded dual estimate $\hat{y}^k$ is drawn from a bounded set $Y \subset \mathbb{R}^m$ at step 1.1. Although not necessary, the choice of $\hat{y}^k$ should also depend on the current dual estimate y^{k-1}. Moreover, the choice of Y can take advantage of *a priori* knowledge of D and its structure, in order to generate better dual estimates. For instance, if $D \subset \mathbb{R}^m$ is compact and convex, we may select

$Y = [-y_{\min}, y_{\max}]^m$ for some $y_{\min}, y_{\max} > 0$, whereas if $D = \mathbb{R}^m_+$, we may more accurately choose $Y = [-y_{\min}, 0]^m$; cf. [36, 53]. In practice, it is advisable to choose the safeguarded multiplier 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.

The augmented Lagrangian functions and subproblems discussed above appear at step 1.2. Section 3.4 is devoted to the numerical solution of the subproblems, discussing several approaches. The subproblems are usually solved only approximately, in some sense, for the sake of computational efficiency. More precisely, the subproblem solver needs to be able to find ε-stationary points of $\mathcal{L}^{\mathrm{S}}_\mu(\cdot, \cdot, y)$ for arbitrarily small $\varepsilon > 0$ $\mu > 0$ and $y \in Y$.

Step 1.3 entails the classical first-order Lagrange multiplier estimate. The update rule is designed around (3.3a) and leads to the inclusion (2.6a) for the primal-dual estimate (x^k, y^k). The monotonicity test at step 1.4 is adopted to monitor primal infeasibility along the iterates. The penalty parameter is reduced at step 1.7 in case of insufficient decrease, effectively implementing a simple feedback strategy to drive $\|c(x^k) - s^k\|$ to zero.

Before proceeding to the convergence analysis, we highlight a different interpretation of the method. As first observed in [49], the augmented Lagrangian method on the primal problem has an associated proximal point method on the dual problem. Introducing the auxiliary variable $r \in \mathbb{R}^m$, we rewrite the augmented Lagrangian subproblem $\min \mathcal{L}^{\mathrm{S}}_\mu(\cdot, \cdot, y)$ as

$$\underset{x,\, s,\, r}{\text{minimize}} \quad q(x) + \delta_D(s) + \frac{1}{2\mu}\|r - \mu y\|^2 \quad \text{subject to} \quad c(x) - s + r = 0$$

and then, by eliminating the slack variable s, as

$$\underset{x,\, r}{\text{minimize}} \quad q(x) + \frac{1}{2\mu}\|r - \mu y\|^2 \quad \text{subject to} \quad c(x) + r \in D.$$

The latter reformulation amounts to a proximal dual regularization of (P) and corresponds to a lifted representation of $\min \mathcal{L}_\mu(\cdot, y)$, where $\mathcal{L}_\mu$ is given in (3.5), thus showing that the approach effectively consists in solving a sequence of subproblems, each one being a proximally regularized version of (P). Yielding feasible and more regular subproblems, this (proximal) regularization strategy has been explored and exploited in different contexts; some recent works are, e.g., [23, 41, 47].

3.2 Convergence analysis

Throughout our convergence analysis, we assume that Algorithm 1 is well-defined, thus requiring that each subproblem at step 1.2 admits an approximate stationary point. Moreover, the following statements assume the existence of some accumulation point x^* or (x^*, s^*) for a sequence $\{x^k\}$ or $\{(x^k, s^k)\}$, respectively, generated by Algorithm 1. In general, coercivity or (level) boundedness arguments should be adopted to verify this precondition; cf. Proposition 3.2 as well.

Due to their practical importance, we focus on affordable, or *local*, solvers, which return merely stationary points, for the subproblems at step 1.2. Instead, we do not present results on the case where the subproblems are solved to *global* optimality. The analysis would follow the classical results in [15, Ch. 5] and [38], see [39, §6.2] as well. In summary, feasible problems would lead to feasible accumulation points that are global minima, in case of existence. For infeasible problems, infeasibility would be minimized and the objective cost minimum for the minimal infeasibility.

Like all penalty-type methods in the nonconvex setting, Algorithm 1 may generate accumulation points that are infeasible for (P). Patterning standard arguments, the following result gives conditions that guarantee feasibility of limit points; cf. [14, Ex. 4.12], [36, Prop. 4.1].

Proposition 3.2 *Let Assumption 1 hold and consider a sequence* $\{(x^k, s^k)\}$ *of iterates generated by Algorithm 1. Then, each accumulation point* x^* *of* $\{x^k\}$ *is feasible for* (P) *if one of the following conditions holds:*

(i) $\{\mu_k\}$ *is bounded away from zero, or*
(ii) there exists some $B \in \mathbb{R}$ *such that* $\mathcal{L}^S_{\mu_k}(x^k, s^k, \hat{y}^k) \leq B$ *for all* $k \in \mathbb{N}$.

In both situations, $(x^*, c(x^*))$ *is an accumulation point of* $\{(x^k, s^k)\}$ *which is feasible to* ($\mathrm{P_S}$).

Proof Let $x^* \in \mathbb{R}^n$ be an arbitrary accumulation point of $\{x^k\}$ and $\{x^k\}_K$ a subsequence such that $x^k \to_K x^*$. We need to show $c(x^*) \in D$ under two circumstances.

(i) If $\{\mu_k\}$ is bounded away from zero, the conditions at steps 1.4 and 1.7 of Algorithm 1 imply that $\|c(x^k) - s^k\| \to 0$ for $k \to \infty$. By the upper bound $\|c(x^k) - s^k\| \geq \mathrm{dist}_D(c(x^k))$ for all $k \in \mathbb{N}$, due to $s^k \in D$, taking the limit $k \to_K \infty$ and continuity yield $\mathrm{dist}_D(c(x^*)) = 0$, hence $c(x^*) \in D$, i.e., x^* is feasible to (P). Further, $s^k \to_K c(x^*)$ holds.

(ii) In case where $\{\mu_k\}$ is bounded away from zero, we can rely on the already proven first statement. Thus, let us assume that $\mu_k \to 0$. By assumption, we have

$$B \geq \mathcal{L}^{\mathrm{S}}_{\mu_k}(x^k, s^k, \hat{y}^k) = q(x^k) + \frac{1}{2\mu_k}\|c(x^k) + \mu_k \hat{y}^k - s^k\|^2 - \frac{\mu_k}{2}\|\hat{y}^k\|^2 \quad (3.6)$$

and $s^k \in D$ for all $k \in \mathbb{N}$. Rearranging terms yields the inequality

$$q(x^k) + \frac{1}{2\mu_k}\|c(x^k) + \mu_k \hat{y}^k - s^k\|^2 \leq B + \frac{\mu_k}{2}\|\hat{y}^k\|^2$$

for all $k \in \mathbb{N}$. Taking the lower limit $k \to_K \infty$ while respecting that q is lsc and $\{\hat{y}^k\}$ is bounded gives $x^* \in \mathrm{dom}\, q$. Particularly, $\{q(x^k)\}_K$ is bounded from below. Rearranging (3.6) yields

$$\|c(x^k) + \mu_k \hat{y}^k - s^k\|^2 \leq 2\mu_k(B - q(x^k)) + \|\mu_k \hat{y}^k\|^2,$$

and taking the upper limit $k \to_K \infty$ yields $\|c(x^k) - s^k\| \to_K 0$, again by boundedness of $\{\hat{y}^k\}$ and $\mu_k \to 0$. On the other hand, $c(x^k) \to_K c(x^*)$ follows by

continuity, and this gives $s^k \to_K c(x^*)$, since D is closed and $s^k \in D$ for all $k \in \mathbb{N}$. Hence, $(x^*, c(x^*))$ is feasible to (P_S), i.e., x^* is feasible to (P).

The final statement of the proposition follows from the above arguments. □

The following convergence result provides fundamental theoretical support to Algorithm 1. It shows that, under subsequential attentive convergence, any feasible accumulation point is an AM-stationary point for (P).

Theorem 3.3 *Let Assumption 1 hold and consider a sequence* $\{(x^k, s^k)\}$ *of iterates generated by Algorithm 1 with* $\varepsilon_k \to 0$. *Let* $(x^*, c(x^*))$ *be an accumulation point of* $\{(x^k, s^k)\}$ *feasible to (*P_S) and $\{(x^k, s^k)\}_K$ a subsequence such that $x^k \overset{q}{\to}_K x^*$ and $s^k \to_K c(x^*)$. Then, x^* is an AM-stationary point for (P).

Proof Define $\zeta^k := s^k - c(x^k)$ for all $k \in \mathbb{N}$. Then, from steps 1.2 and 1.3 of Algorithm 1, we have that

$$-\nabla c(x^k)^\top y^k + \xi^k \in \partial q(x^k), \tag{3.7}$$

$$y^k + \nu^k \in \mathcal{N}_D^{\lim}(c(x^k) + \zeta^k) \tag{3.8}$$

for some $\xi^k \in \mathbb{R}^n$, $\|\xi^k\| \le \varepsilon_k$, and $\nu^k \in \mathbb{R}^m$, $\|\nu^k\| \le \varepsilon_k$; cf. (2.3) and (3.3). Set $\lambda^k := y^k + \nu^k$ and $\eta^k := \nabla c(x^k)^\top \nu^k + \xi^k$ for all $k \in \mathbb{N}$.

We claim that the four subsequences $\{x^k\}_K$, $\{\eta^k\}_K$, $\{\lambda^k\}_K$ and $\{\zeta^k\}_K$ satisfy the properties in Definition 2.4 and therefore show that x^* is an AM-stationary point for (P).

By construction, we have $x^k \overset{q}{\to}_K x^*$ as well as $-\nabla c(x^k)^\top \lambda^k + \eta^k \in \partial q(x^k)$ and $\lambda^k \in \mathcal{N}_D^{\lim}(c(x^k) + \zeta^k)$ for each $k \in \mathbb{N}$. Continuous differentiability of c and $\|\xi^k\|, \|\nu^k\| \le \varepsilon_k$ give $\|\eta^k\| \to_K 0$. Finally, $\zeta^k \to_K 0$ follows from $s^k \to_K c(x^*)$, $x^k \to_K x^*$ and continuity of c.

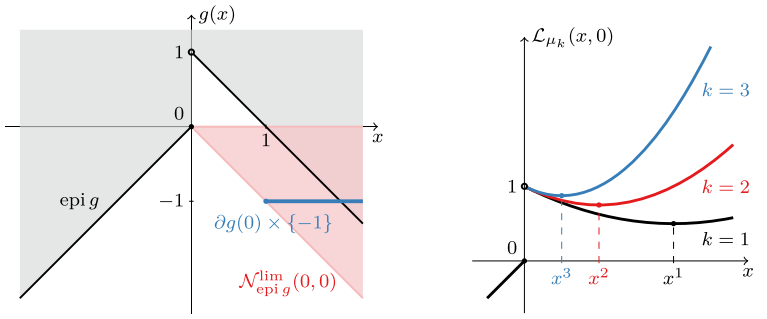
Overall, this proves that x^* is an AM-stationary point for (P). □

The additional assumption $x^k \overset{q}{\to}_K x^*$ in Theorem 3.3 is trivially satisfied if g is continuous on its domain since all iterates of Algorithm 1 belong to dom g. However, the following one-dimensional example illustrates how this additional requirement appears to be indispensable in a discontinuous setting.

Example 3.4 We consider $n := m := 1$ and set $D := (-\infty, 0]$,

$$f(x) := 0, \qquad g(x) := \begin{cases} x & \text{if } x \le 0, \\ 1 - x & \text{otherwise,} \end{cases} \qquad c(x) := x.$$

Note that g is merely lsc at $x^* := 0$, and that $\partial g(x^*) = [1, \infty)$; cf. Fig. 1a. Although x^* is the global maximizer of the associated problem (P), x^* is not an M-stationary point. Since $\nabla f(x^*) = 0$, $\nabla c(x^*) = 1$ and $\mathcal{N}_D^{\lim}(c(x^*)) = \mathbb{R}_+$, there is no $y^* \in \mathcal{N}_D^{\lim}(c(x^*))$ such that $0 \in \nabla f(x^*) + \partial g(x^*) + \nabla c(x^*)^\top y^*$. Indeed, x^* is not even AM-stationary. Possibly discarding early iterates, any sequence $\{x^k\}$ such that $x^k \overset{q}{\to} x^*$



(a) *Computation of* $\partial g(0)$. **(b)** *Iterates* x^k *for* $k \in \{1,2,3\}$.

Fig. 1 Visualizations for Example 3.4

satisfies $x^k \leq 0$ for each $k \in \mathbb{N}$. Hence, we find $\partial q(x^k) \subset [1, \infty)$, $\nabla c(x^k) = 1$ and $\mathcal{N}_D^{\lim}(c(x^k) + \zeta^k) \subset \mathbb{R}_+$ for each $\zeta^k \in \mathbb{R}^m$ and $k \in \mathbb{N}$, showing that the distance between 0 and the set $\partial q(x^k) + \nabla c(x^k)^\top \mathcal{N}_D^{\lim}(c(x^k) + \zeta^k)$ is at least 1.

We apply Algorithm 1 with $Y := \{0\}$, $\mu_0 := 1$, $\theta := 1/4$ and $\kappa := 1/2$. This may yield sequences $\{x^k\}$, $\{s^k\}$ and $\{\mu_k\}$ given by $x^0 := \mu_0$, $s^0 := 0$, $x^k := \mu_k := 2^{1-k}$ and $s^k := 0$ for each $k \in \mathbb{N}$, $k \geq 1$; cf. Figure 1b. Hence, we have $x^k \to x^*$ and, crucially, not $x^k \overset{q}{\to} x^*$.

The next result readily follows from Corollary 2.7 and Theorem 3.3.

Corollary 3.5 *Let Assumption 1 hold and consider a sequence* $\{(x^k, s^k)\}$ *of iterates generated by Algorithm 1 with* $\varepsilon_k \to 0$. *Let* $(x^*, c(x^*))$ *be an accumulation point of* $\{(x^k, s^k)\}$ *feasible to* (P$_S$) and $\{(x^k, s^k)\}_K$ a subsequence such that $x^k \overset{q}{\to}_K x^*$ and $s^k \to_K c(x^*)$. Furthermore, assume that x^* is AM-regular for (P). Then, x^* is an M-stationary point for (P).

We note that related results have been obtained in [18, Thm 3.1] and [39, Cor. 6.16]. In [18], however, the authors in most cases overlooked the issue of attentive convergence in the definition of the limiting subdifferential for discontinuous functions so that their findings are not reliable.

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 global minimizers of any infeasibility measure. Nevertheless, the next result proves that Algorithm 1 with bounded $\{\varepsilon_k\}$ finds stationary points of an infeasibility measure. Notice that this property does not require $\varepsilon_k \to 0$, but only boundedness; cf. [15, Thm 6.3].

Proposition 3.6 *Let Assumption 1 hold and consider a sequence* $\{(x^k, s^k)\}$ *of iterates generated by Algorithm 1 with* $\{\varepsilon_k\}$ *bounded. Let* (x^*, s^*) *be an accumulation point of* $\{(x^k, s^k)\}$ *and* $\{(x^k, s^k)\}_K$ *a subsequence such that* $x^k \overset{q}{\to}_K x^*$ *and* $s^k \to_K s^*$. *Then,* $(x^*, q(x^*), s^*)$ *is an M-stationary point of the feasibility problem*

$$\underset{(x,\alpha,s)\in \text{epi}\, q \times D}{\text{minimize}} \quad \tfrac{1}{2}\|c(x) - s\|^2. \tag{3.9}$$

If q is locally Lipschitz continuous at x^*, then x^* is an M-stationary point of the constraint violation

$$\underset{(x,s)\in\mathbb{R}^n\times D}{\text{minimize}} \quad \tfrac{1}{2}\|c(x)-s\|^2. \tag{3.10}$$

Proof. By Proposition 3.2*(i)*, if $\{\mu_k\}$ is bounded away from zero, x^* is feasible for (P) and $s^* = c(x^*) \in D$. Thus, $(x^*, q(x^*), c(x^*))$ is a global minimizer of (3.9) and $(x^*, c(x^*))$ is a global minimizer of (3.10). By continuous differentiability of the objective function, M-stationarity with respect to both problems follows, see [44, Prop. 5.1]. Hence, it remains to consider the case $\mu_k \to 0$.

Owing to step 1.2 of Algorithm 1, for all $k \in \mathbb{N}$ it is

$$\xi^k \in \partial q(x^k) + \nabla c(x^k)^\top \left[\hat{y}^k + (c(x^k)-s^k)/\mu_k\right], \tag{3.11a}$$

$$\nu^k \in -\left[\hat{y}^k + (c(x^k)-s^k)/\mu_k\right] + \mathcal{N}_D^{\lim}(s^k) \tag{3.11b}$$

for some $\xi^k \in \mathbb{R}^n$, $\|\xi^k\| \le \varepsilon_k$, and $\nu^k \in \mathbb{R}^m$, $\|\nu^k\| \le \varepsilon_k$; cf. (3.3). Particularly, (3.11a) gives us

$$(\xi^k - \nabla c(x^k)^\top[\hat{y}^k + (c(x^k)-s^k)/\mu_k], -1) \in \mathcal{N}^{\lim}_{\operatorname{epi} q}(x^k, q(x^k)).$$

Multiplying by $\mu_k > 0$ and exploiting that $\mathcal{N}^{\lim}_{\operatorname{epi} q}(x^k, q(x^k))$ is a cone, we have

$$(\mu_k\xi^k - \nabla c(x^k)^\top[c(x^k) + \mu_k\hat{y}^k - s^k], -\mu_k) \in \mathcal{N}^{\lim}_{\operatorname{epi} q}(x^k, q(x^k)). \tag{3.12}$$

Furthermore, (3.11b) yields

$$\mu_k(\nu^k + \hat{y}^k) + c(x^k) - s^k \in \mathcal{N}_D^{\lim}(s^k) \tag{3.13}$$

since $\mathcal{N}_D^{\lim}(s^k)$ is a cone. Taking the limit $k \to_K \infty$ in (3.12) and (3.13), the robustness of the limiting normal cone, $x^k \overset{q}{\to}_K x^*$ and boundedness of $\{\hat{y}^k\}$, $\{\xi^k\}$ and $\{\nu^k\}$ yield

$$\begin{aligned}(-\nabla c(x^*)^\top[c(x^*)-s^*], 0) &\in \mathcal{N}^{\lim}_{\operatorname{epi} q}(x^*, q(x^*)),\\ c(x^*) - s^* &\in \mathcal{N}_D^{\lim}(s^*).\end{aligned} \tag{3.14}$$

Keeping the Cartesian product rule for the computation of limiting normals in mind, see [44, Prop. 1.2], $(x^*, q(x^*), s^*)$ is an M-stationary point of (3.9).

Finally, assume that q is locally Lipschitz continuous at x^*. Then, due to [44, Cor. 1.81], we have

$$(y^*, 0) \in \mathcal{N}^{\lim}_{\operatorname{epi} q}(x^*, q(x^*)) \quad \Rightarrow \quad y^* = 0,$$

so that the above arguments already show M-stationarity of (x^*, s^*) for (3.10). □

In case where D is convex, the assertion of Proposition 3.6 can be slightly strengthened.

Corollary 3.7 *Let D be convex, let Assumption 1 hold and consider a sequence $\{(x^k, s^k)\}$ of iterates generated by Algorithm 1 with $\{\varepsilon_k\}$ bounded. Let (x^*, s^*) be an accumulation point of $\{(x^k, s^k)\}$ and $\{(x^k, s^k)\}_K$ a subsequence such that $x^k \xrightarrow{q}_K x^*$ and $s^k \to_K s^*$. Then, $(x^*, q(x^*))$ is an M-stationary point of the feasibility problem*

$$\underset{(x,\alpha)\in \operatorname{epi} q}{\text{minimize}} \quad \tfrac{1}{2}\operatorname{dist}_D^2(c(x)).$$

If q is locally Lipschitz continuous at x^*, then x^* is an M-stationary point of the constraint violation

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \quad \tfrac{1}{2}\operatorname{dist}_D^2(c(x)).$$

Proof We proceed as in the proof of Proposition 3.6 in order to come up with (3.14). By convexity of D, $c(x^*) - s^* \in \mathcal{N}_D^{\lim}(s^*)$ is equivalent to $s^* \in \Pi_D(c(x^*))$. Thus, the assertion follows from Lemma 2.1. □

3.3 Termination criteria

Step 1.2 involves the minimization of the augmented Lagrangian function defined in (3.5). Then, the dual update at step 1.3 allows to draw conclusions with respect to the original problem (P), as Theorem 3.3 shows that accumulation points of sequences generated by Algorithm 1 are AM-stationary under mild assumptions.

Owing to (3.7)–(3.8) and recalling the AM-stationarity conditions (2.6), one may select a null sequence $\{\varepsilon^k\} \subset \mathbb{R}_{++}$ at step 1.1. Then, given some user-defined tolerances $\varepsilon^{\text{dual}}, \varepsilon^{\text{prim}} > 0$, it is reasonable to declare successful convergence when the conditions

$$\varepsilon^k \le \varepsilon^{\text{dual}} \quad \text{and} \quad \|c(x^k) - s^k\| \le \varepsilon^{\text{prim}}$$

are satisfied. Theorem 3.3 demonstrates that these termination criteria (the latter, in particular) are satisfied in finitely many iterations if any subsequence of $\{(x^k, s^k)\}$ accumulates at a feasible point $(x^*, c(x^*))$ of ($\mathrm{P_S}$). As this might not be the case, a mechanism for (local) infeasibility detection is needed, and usually included in practical implementations; see [5, 17].

Given some tolerances, Algorithm 1 can be equipped with relaxed conditions on decrease requirements at step 1.4 and optimality at step 1.2. At step 1.1 the inner tolerance ε^k can stay bounded away from zero, as long as $\varepsilon^k \le \varepsilon^{\text{dual}}$ for large $k \in \mathbb{N}$. Similarly, the condition at step 1.4 can be relaxed by adding the (inclusive) possibility that $\|c(x^k) - s^k\| \le \varepsilon^{\text{prim}}$. Finally, at step 1.5 a nonmonotone update is allowed, namely the penalty parameter can be increased, as long as some watchdog procedures are in place to avoid cycling [14].

3.4 Inner problem and solver

In this section we elaborate upon step 1.2 of Algorithm 1 that aims at minimizing the augmented Lagrangian function $\mathcal{L}_\mu^{\mathrm{S}}(\cdot, \cdot, y)$ defined in (3.2). To this end, let us take a closer look at the structure of this subproblem.

Using the decomposition $\mathcal{L}_\mu^{\mathrm{S}}(\cdot,\cdot,y) = f^{\mathrm{S}}(\cdot,\cdot)+g^{\mathrm{S}}(\cdot,\cdot)$ with component functions $f^{\mathrm{S}}\colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ and $g^{\mathrm{S}}\colon \mathbb{R}^n \times \mathbb{R}^m \to \overline{\mathbb{R}}$ given by

$$f^{\mathrm{S}}(x,s) := f(x) + \frac{1}{2\mu}\|c(x)+\mu y - s\|^2 - \frac{\mu}{2}\|y\|^2, \tag{3.15}$$

$$g^{\mathrm{S}}(x,s) := g(x) + \delta_D(s), \tag{3.16}$$

one immediately sees that this split recovers the classical setting of an *unconstrained* composite optimization problem with f^{S} being continuously differentiable, while g^{S} is merely lsc, but of a particular structure. In principle, proximal gradient-type methods can therefore be applied as approximate solvers for our subproblems, see [9] for an introduction of this class of methods. A standing assumption of the corresponding convergence theory in [9] and all previous works on proximal gradient-type methods, however, is a global Lipschitz condition regarding the gradient of the smooth part f^{S}. Note that this gradient is given by

$$\nabla f^{\mathrm{S}}(x,s) = \begin{bmatrix} \nabla f(x) + \frac{1}{\mu}\nabla c(x)^\top [c(x)+\mu y - s] \\ -\frac{1}{\mu}[c(x)+\mu y - s] \end{bmatrix}.$$

Observe that our standing assumptions from Assumption 1 imply that this gradient is locally Lipschitz continuous, but they do not guarantee global Lipschitzness in general. Fortunately, some recent contributions on proximal gradient-type methods show that these methods also work under suitable assumptions if the smooth term has a locally Lipschitz gradient only; cf. [7, 24, 37] for more details. Consequently, these proximal gradient-type methods offer a viable way to solve the augmented Lagrangian subproblems, even for fully nonconvex problems. Let us also mention that, at least in [24, 37], it has been verified that accumulation points of sequences generated by proximal gradient-type methods are stationary while along the associated subsequence, the iterates are ε_k-stationary for a null sequence $\{\varepsilon_k\}$. This requirement is essential in Algorithm 1.

For a practical implementation of these proximal methods, it is advantageous to exploit the particular structure of the nonsmooth term g^{S}. In fact, due to the separability of g^{S} with respect to x and s, it follows that the corresponding proximal mapping is easily computable. More precisely, one obtains

$$\mathrm{prox}_{\gamma g^{\mathrm{S}}}(x,s) = \begin{bmatrix} \mathrm{prox}_{\gamma g}(x) \\ \Pi_D(s) \end{bmatrix}$$

for any $\gamma \in (0, \gamma_g)$.

Though the proximal-type approach is used in our numerical setting (see the next section for some more details), we stress that there exist other candidates for the numerical solution of the resulting augmented Lagrangian subproblems. To this end, recall that the previous discussion looked at these subproblems as an unconstrained composite optimization problem. Alternatively, we may view these subproblems from the point of view of machine learning, where (essentially) the same class of optimization

problems is solved by (possibly) different techniques. We refer the interested reader to [54, 60] for a survey of optimization methods for machine learning and data analysis problems. These techniques might be applicable very successfully at least in certain situations. For example, if the smooth term f^{S} is convex (the gradient does not have to be globally Lipschitz), whereas the nonsmooth term g^{S} is still just assumed to be lsc (and not necessarily convex), it is possible to adapt the idea of cutting plane methods to this setting by applying the cutting plane technique to f^{S} only, whereas one does not change the nonsmooth term. The resulting subproblems then use a piecewise affine lower bound for the function f^{S} and add the (possibly complicated) function g^{S}. Of course, and similar to the proximal gradient-type approaches, these subproblems need to be easily solvable for the overall augmented Lagrangian method to be efficient, and this, in general, is true only for particular classes of problems; cf. Sect. 4.

4 Numerical examples

This section presents a numerical implementation of Algorithm 1 and discusses its behavior on some illustrative examples, showcasing the flexibility offered by the constrained composite programming framework. In particular, we consider challenging problems where the cost function is nonsmooth and nonconvex or where the constraints are inherently nonconvex by a disjunctive structure of the respective set D. In Sect. 4.2 we demonstrate the benefit of accelerated proximal-gradient methods for solving the subproblems by means of a simple two-dimensional problem where a nonsmooth variant of the Rosenbrock function is minimized over a set of combinatorial structure. Next, Sect. 4.3 is dedicated to a binary optimal control problem with nonlinear dynamics, free final time and switching costs, where we display and discuss weaknesses of our approach. Section 4.4 deals with a test collection of portfolio optimization problems from [28] which are equipped with a nonconvex sparsity-promoting term in the objective function. Finally, in Sect. 4.5 we address a class of matrix recovery problems discussed e.g. in [52] where the rank of the unknown matrix has to be minimized.

4.1 Implementation

We have implemented the proposed Augmented Lagrangian Solver (ALS) as part of an open-source software package in the Julia language [13]. ALS can solve constrained composite problems of the form (P) and is available online at https://github.com/aldma/Bazinga.jl, together with the examples presented in the following sections. ALS can be used to solve, in the sense of Sect. 3.3, a wide spectrum of optimization problems, requiring only first-order primitives, i.e., gradient, proximal mapping and projections. By default, ALS invokes PANOC$^+$ [24] for solving the augmented Lagrangian subproblems at step 1.2 of Algorithm 1, possibly inexactly and up to stationarity, using the implementation offered by ProximalAlgorithms.jl [55]; see Appendix A for more details. The method is implemented matrix-free, that is, the constraint Jacobian ∇c does not need to be explicitly formed as only Jacobian-vector products $\nabla c(x)^\top v$ are required.

The solver requires the data functions f, g, c and constraint set D specified as objects returning the oracles discussed at the end of Sect. 1. Further, the initialization requires a primal-dual starting point $(x^{\text{init}}, y^{\text{init}}) \in \mathbb{R}^n \times \mathbb{R}^m$. The default safeguarding set Y in $\mathbb{R}^m$ is $Y = [-y_{\max}, y_{\max}]^m$, with $y_{\max} = 10^{20}$, and the safeguarded dual estimate $\hat{y}^k$ at step 1.1 is chosen as the projection of y^{k-1} onto Y; of y^{init} for $k = 0$. User override of this oracle allows for tailored choices of Y, possibly exploiting the structure of D [53].

ALS initializes Algorithm 1 by overwriting x^{init} with an arbitrary element of $\text{prox}_{\gamma g}(x^{\text{init}}) \subset \text{dom}\, q$, where $\gamma = \epsilon_M$ and ϵ_M denotes the machine epsilon of a given floating-point system. The examples presented in the following are in double precision (Float64), so $\epsilon_M \approx 2.22 \cdot 10^{-16}$. The inner tolerances ε_k at step 1.1 are constructed as a sequence of decreasing values, defined by the recurrence

$$\varepsilon_{k+1} = \max\{\kappa_\varepsilon \varepsilon_k, \varepsilon^{\text{dual}}\},$$

starting from $\varepsilon_0 := (\varepsilon^{\text{dual}})^{\frac{1}{3}}$ and given some $\varepsilon^{\text{dual}}, \kappa_\varepsilon \in (0, 1)$ [14]. The initial penalty parameter μ_0 is automatically chosen by default, similarly to [15, Eq. 12.1]. Given $x^{\text{init}} \in \text{dom}\, q$, we evaluate the constraints $c^{\text{init}} := c(x^{\text{init}})$, select an arbitrary element $s^{\text{init}} \in \Pi_D(c^{\text{init}})$ and compute the vector $\Delta^{\text{init}} := c^{\text{init}} - s^{\text{init}}$. Then, the vector $\mu_0 \in \mathbb{R}^m$ of penalty parameters is selected componentwise as follows:

$$(\mu_0)_i := \max\left\{10^{-8}, \min\left\{\frac{1}{10}\frac{\max\{1, (\Delta_i^{\text{init}})^2/2\}}{\max\{1, q(x^{\text{init}})\}}, 10^8\right\}\right\},$$

effectively scaling the contribution of each constraint [15, 20]. Then, according to the overall feasibility-complementarity of the iterate, the penalty parameters are updated in unison at step 1.7, since using a different penalty parameter for each constraint is theoretically worse than using a common parameter [2, §3.4]; we set $\mu_{k+1} := \kappa_\mu \mu_k$, for some fixed $\kappa_\mu \in (0, 1)$. At the kth iteration, the subsolver at step 1.2 is warm-started from the previous estimate $(x^{k-1}, s^{k-1}) \in \text{dom}\, q \times D$; from $(x^{\text{init}}, s^{\text{init}})$ for $k = 0$.

The default parameters in ALS are $\theta = 0.8$, $\kappa_\mu = 0.5$ and $\kappa_\varepsilon = 0.1$, termination tolerances $\varepsilon^{\text{prim}} = \varepsilon^{\text{dual}} = 10^{-6}$ and a maximum number of (outer) iterations, whose default value is 100.

4.2 Nonsmooth Rosenbrock and either-or constraints

Let us consider a two-dimensional optimization problem involving a nonsmooth Rosenbrock-like objective function and either-or constraints, namely set-membership constraints entailing an inclusive disjunction. It reads

$$\underset{x}{\text{minimize}}\ 10(x_2 + 1 - (x_1 + 1)^2)^2 + |x_1| \quad \text{subject to } x_2 \le -x_1 \ \vee\ x_2 \ge x_1 \quad (4.1)$$

and admits a unique (global) minimizer $x^* = (0, 0)$. The feasible set is nonconvex and connected; see Fig. 2. We cast (4.1) into the form of (P) by defining the data functions

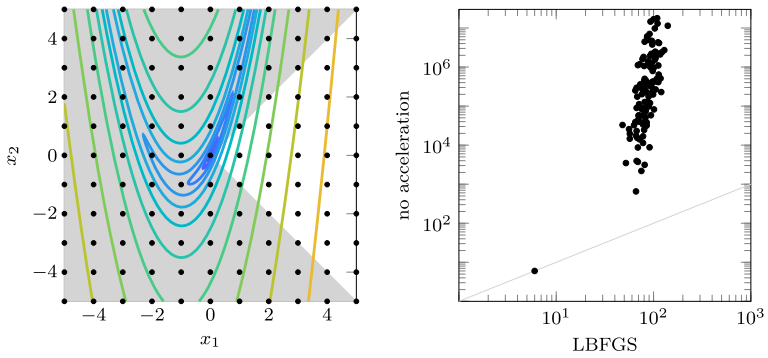


Fig. 2 Setup and results for the illustrative problem (4.1). Left: Feasible region (gray background), objective contour lines, global minimizer $x^\star = (0, 0)$ and grid of starting points. Right: Comparison of inner iterations needed without acceleration against LBFGS acceleration; each mark corresponds to a starting point and the gray line has unitary slope

as

$$f(x) := 10(x_2 + 1 - (x_1 + 1)^2)^2, \qquad g(x) := |x_1|, \qquad c(x) := \begin{pmatrix} -x_1 - x_2 \\ -x_1 + x_2 \end{pmatrix},$$

and let the constraint set be $D := D_{\mathrm{EO}}$, where the (nonconvex) set

$$D_{\mathrm{EO}} := \{(a, b) \big| a \geq 0 \vee b \geq 0\} = \{(a, b) \big| a \geq 0\} \cup \{(a, b) \big| b \geq 0\}$$

describes the either-or constraint.

We consider a uniform grid of $11^2 = 121$ starting points x^0 in $[-5, 5]^2$ and let the initial dual estimate be $y^0 = 0$. Also, we compare the performance of ALS by solving the subproblems using PANOC$^+$ without or with (LBFGS) acceleration; see the last paragraph of Appendix A for more details.

ALS solves all the problem instances, approximately (tolerance 10^{-3} in Euclidean distance) reaching $x^\star = (0, 0)$ in all cases. Figure 2 depicts the feasible region of (4.1), some contour lines of its objective function and the grid of starting points x^0. Over all problems, ALS with no acceleration takes at most 17 870 346 (cumulative) inner iterations to find a solution (median 291 756), whereas with LBFGS directions only 140 inner iterations are needed at most (median 86). A closer look at Fig. 2 indicates that not only the accelerated method usually requires far less iterations, but also that its behavior is more consistent, as the majority of cases spread over a narrow interval. These results support the claim that (quasi-Newton) acceleration techniques can give a mean to cope with bad scaling and ill-conditioning [56, 58].

4.3 Sparse switching time optimization

Constrained composite programming offers a flexible language for modeling a variety of problems. In this section we consider the sparse binary optimal control of Lotka-Volterra dynamics. Known as the fishing problem [51, §6.4], it is typically stated as

$$
\begin{aligned}
\underset{x,u}{\text{minimize}} \quad & \int_0^T \|x(t) - 1\|^2 \mathrm{d}t \\
\text{subject to} \quad & \dot{x}_1(t) = x_1(t)[-c_1 u(t) - x_2(t) + 1] && \text{for a.e. } t \in [0, T], \\
& \dot{x}_2(t) = x_2(t)[-c_2 u(t) + x_1(t) - 1] && \text{for a.e. } t \in [0, T], \\
& x(0) = x_0, \\
& u(t) \in \{0, 1\} && \text{for } t \in [0, T], \qquad (4.2)
\end{aligned}
$$

where final time $T = 12$, initial state $x_0 := (0.5, 0.7)$ and parameters $c_1 = 0.4$, $c_2 = 0.2$ are given and fixed. In order to showcase the peculiar features of (P), we focus on a variant of the fishing problem with switch costs and free, although constrained, final time. First, the problem is reformulated as a finite-dimensional one by adopting the switching time optimization approach, that consists in optimizing the times at which the control input changes, given a fixed sequence of N admissible controls [51, §5.2]. We call switching intervals the time between these switching times and collect them in a vector $\tau \in \mathbb{R}^N$. Clearly, they must take nonnegative values and sum up to the final time T. Furthermore, considering the chattering solution exhibited by the fishing problem [51, §6.5], we introduce switch costs to penalize solutions that show frequent switching of the binary control trajectory, yielding more practical results. Following [21], [22, Ch. 2], switch costs can be interpreted as a regularization term and modeled using the ℓ_0 quasi-norm of the switching intervals, effectively counting how many control inputs in the given control sequence are active. The resulting problem formulation reads

$$
\underset{\tau}{\text{minimize}} \quad f(\tau) + \delta_{\mathbb{R}^N_+}(\tau) + \sigma \|\tau\|_0 \qquad \text{subject to} \quad 1_N^\top \tau \in D. \qquad (4.3)
$$

Here, the smooth cost function f returns the tracking cost, by integrating the dynamics, starting from the initial state, for the given sequence of control inputs and switching intervals. The nonnegativity constraint $\delta_{\mathbb{R}^N_+}$ and sparsity-promoting cost $\sigma\|\cdot\|_0$ form the nonsmooth cost function g in (P); despite g being nonconvex and discontinuous, its proximal mapping can be easily evaluated [21, §3.2]. The nonnegative parameter σ controls the impact of the ℓ_0 regularization and can be interpreted as the switching cost. The only constraint remained explicit is the one on the final time $T := 1_N^\top \tau$. Hence, the constraint set $D \subset \mathbb{R}_+$ is constituted by the admissible values for T.

We consider the binary control sequence $\{0, 1, 0, \ldots, 1\}$ with $N := 24$ intervals. A background time grid with $n = 200$ points is adopted to integrate dynamics and evaluate sensitivities, following the linearization approach of [57]. We solve (4.3) for increasing values of the switching cost parameter $\sigma \in \{10^{-6}, 10^{-5}, \ldots, 10\}$. For the first problem, the initial guess τ^0 corresponds to uniform switching intervals with the final time $T = 12$ usually fixed in (4.2). Then, following a continuation approach, a solution is adopted as initial guess for the subsequent problem, but always with dual estimate $y^0 = 0$. Moreover, we consider two cases for the constraint set D. First, we let $D := [0, 15]$ and ALS returns solutions whose final time reaches values around $T \approx 12$. Then, we consider a second case with the disconnected constraint set $D := [5, 10] \cup [13, 15]$, so to impact on the solution; in this case the returned final times are $T \approx 13$.

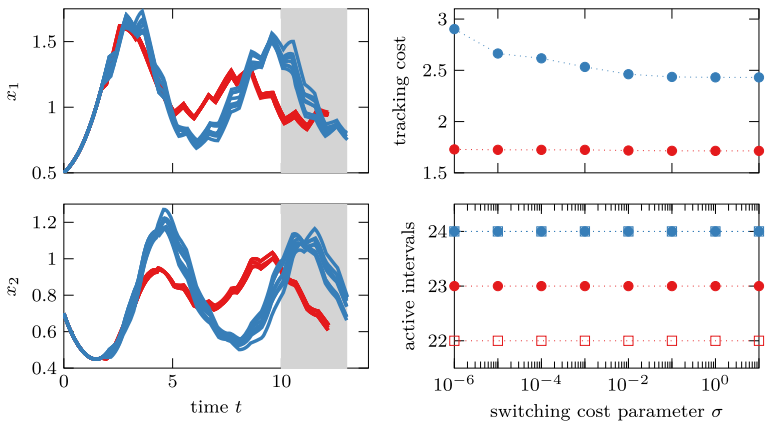


Fig. 3 Results for the illustrative problem (4.3) using switching time optimization with a sequence of 24 binary controls and several values for the switching cost parameter σ. Left: Prohibited region for the final time (gray background) and state trajectories with (blue) or without (red) constraint. Right: Comparison of the resulting tracking cost and number of nonzero variables, corresponding to active intervals (circle). Identical control trajectories can be obtained with fewer active intervals (square), yielding lower switching cost (color figure online)

ALS is able to find reasonable solutions that satisfy the constraints, despite the nonconvexity of the switching time approach [51, Apx B.4], the discrete nature of the sparse regularizer and the constraint set D being disconnected. It should be stressed, however, that there are no guarantees on the quality of these solutions and, in fact, the solutions found by ALS are poor in terms of objective value, as we are about to show.

The state trajectories are depicted in Fig. 3, for both cases, along with a comparison of the tracking cost and number of active intervals against the switching cost parameter σ. First, we observe that the trajectories are not strongly affected, despite the dramatic increase of σ (relative to the tracking cost). Moreover, the solver performs only few iterations, needed to adjust the dual estimate and verify the termination criteria. In practice, the iterates remain trapped around a minimizer with high objective value, and a huge value of σ is required for jumping to a lower objective value. This becomes apparent looking at $\|\tau\|_0$, namely the number of active intervals. Given a sequence of control inputs, several choices of switching intervals can give the same state trajectory, hence the same tracking cost. Among these, we would expect the solver to return one with minimum number of nonzeros. For instance, vectors of switching intervals in the form $(\alpha+\beta, 0, 0, \dots)$ and $(0, 0, \alpha+\beta, \dots)$ should be preferred over $(\alpha, 0, \beta, \dots)$, for they yield the same control trajectory whilst having fewer nonzero elements. The solutions returned by ALS are compared against equivalent although sparser ones in Fig. 3. Clearly, and not surprisingly, the solutions obtained are far from being globally optimal.

4.4 Sparse portfolio optimization

Let us consider portfolio optimization problems in the form

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & \frac{1}{2} x^\top Q x + \alpha \|x\|_0 \\ \text{subject to} \quad & \mu^\top x \geq \varrho, \quad 1_n^\top x = 1, \quad 0 \leq x \leq u. \end{aligned} \tag{4.4}$$

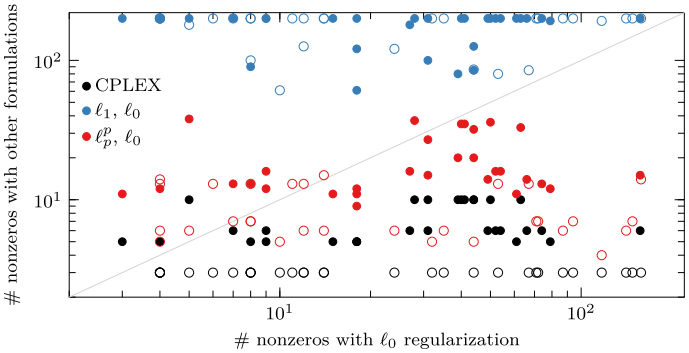


Fig. 4 Results for the portfolio problem (4.4): Comparison of the solutions found with ℓ_0 regularization against those obtained with CPLEX and ℓ_0 warm-started with ℓ_1 or ℓ_p^p, with $p = 0.5$. We depict the number of nonzero entries of the solutions returned for $\alpha = 10$ (dot) and $\alpha = 100$ (circle). The gray line has unitary slope

The problem data $Q \in \mathbb{R}^{n\times n}$ and $\mu \in \mathbb{R}^n$ denote the covariance matrix and the mean of $n \in \mathbb{N}$ possible assets, respectively, while $\varrho \in \mathbb{R}$ is a lower bound for the expected return. Furthermore, $u \in \mathbb{R}^n$ provides an upper bound for the individual assets within the portfolio. Aiming at a sparse portfolio, and in contrast with cardinality-constrained formulations, see e.g. [36], we use the ℓ_0 quasi-norm as a regularization term that penalizes the number of chosen assets within the portfolio.

We reformulate the model in the form of (P) by letting f be the quadratic cost, g the nonsmooth cost and indicator of the bounds, $c\colon \mathbb{R}^n \to \mathbb{R}^m$, $m := 2$, defined by $c(x) := [\mu, 1_n]^\top x$ and $D := [\varrho, \infty) \times \{1\}$.

Through a mixed-integer quadratic program formulation of (4.4), which can be obtained via the theory provided in [27], we compute a solution using CPLEX [35], for comparison. Based on our experiences from Sect. 4.3, we also solve (4.4) using a continuation procedure: the ℓ_0 minimization is warm-started at a primal-dual point found replacing the discontinuous ℓ_0 function with either the norm $\ell_1 := \|\cdot\|_1$ or the p-th power of the ℓ_p quasi-norm, i.e., $\ell_p^p := \|\cdot\|_p^p$ ($p = 0.5$) and solving the corresponding problem. Notice that (4.4) with the ℓ_0- replaced by the ℓ_1-term boils down to a convex quadratic program; in fact, it is $\|x\|_1 = 1$ for each feasible point of (4.4) by the nonnegativity and equality constraints.

The data Q, μ, ϱ and u is taken from the test problem collection [28], which has been created randomly and is available online [29]. Here, we used all 30 test instances of dimension $n := 200$ and the two different values $\alpha \in \{10, 100\}$ for each problem.

The results of our experiments are depicted in Fig. 4. Let us mention that ALS solved all problem instances, in the sense that it returned primal-dual pairs satisfying the termination criteria of Sect. 3.3. Below, we comment on some median values for our experiments with parameters $\alpha = 10/100$: a direct use of ℓ_0 minimization resulted in 10/13 outer and 908/1633 inner iterations, while warm-starting with the continuous ℓ_p^p function required 13/9 outer and 686/1830 inner iterations. Let us point the reader's attention to the fact that the ℓ_p^p-warm-started ℓ_0 minimization did not affect the solution sparsity, i.e., the numbers of nonzero components of the obtained solutions were the same with and without an additional round of ℓ_0 minimization after the ℓ_p^p warm-start.

Although one cannot expect to find a global minimum in general, we recall that the standard ℓ_1 regularization does not work in this example, as confirmed by the poor performance depicted in Fig. 4, whereas the nonconvex ℓ_p^p penalty already leads to very sparse solutions.

4.5 Matrix completion with minimum rank

For some $\ell \in \mathbb{N}$, $\ell \geq 2$, let us consider $N \in \mathbb{N}$ points $x_1, \ldots, x_N \in \mathbb{R}^\ell$ and define a block matrix $X \in \mathbb{R}^{N\times\ell}$ by means of $X := [x_1, x_2, \ldots, x_N]^\top$. Let $\Delta \in \mathbb{R}^{N\times N}$ denote the Euclidean distance matrix associated with these points, given by $\Delta_{ij} := \|x_i - x_j\|^2 = (x_i - x_j)^\top (x_i - x_j)$ for all $i, j \in \mathcal{I} := \{1, \ldots, N\}$. We aim at recovering X based on a partial knowledge of Δ. In particular, we assume that $\Omega \subset \mathcal{I}^2$ is a set of pairs such that only the entries Δ_{ij}, $(i, j) \in \Omega$, of Δ are known.

Following [52], we lift the problem by introducing a symmetric matrix $B := XX^\top$ whose rank is, by construction, smaller than or equal to ℓ. Hence, we seek a matrix $B \in \mathbb{R}^{N\times N}$ that satisfies the symmetry constraint $B = B^\top$ and the distance constraints associated with the observations, i.e., $B_{ii} + B_{jj} - B_{ij} - B_{ji} = \Delta_{ij}$ has to hold for all $(i, j) \in \Omega$. Among these admissible matrices, those with minimum rank are preferred.

Let us consider problems of type

$$\begin{array}{lll} \underset{B}{\text{minimize}} & g(B) & \\ \text{subject to} & B_{ii} + B_{jj} - B_{ij} - B_{ji} = \Delta_{ij} & \forall (i, j) \in \Omega, \\ & B_{ij} = B_{ji} & \forall i, j \in \mathcal{I}, j < i \end{array} \tag{4.5}$$

where the function $g\colon \mathbb{R}^{N\times N} \to \mathbb{R}$ encodes a matrix regularization term. In the following, we consider $g := \text{rank} := \|\sigma(\cdot)\|_0$, the nuclear norm $g := \|\cdot\|_* := \sum_i \sigma_i(\cdot)$ or the p-powered Schatten p-quasi-norm $g := \|\cdot\|_p^p := \sum_i \sigma_i(\cdot)^p$, $p \in (0, 1)$, where $\sigma(A)$ denotes the vector of singular values of a matrix A. In our experiments rank and singular values are numerically evaluated using Julia's LinearAlgebra functions `rank` and `svd`, respectively. Notice in particular that the rank of a matrix A is computed by counting how many singular values of A have magnitude greater than a numerical tolerance whose value depends on the machine precision.

Denoting $m_o := |\Omega|$ and $m_s := N(N-1)/2$ the number of observation and symmetry constraints, respectively, there are $n := N^2$ variables and $m := m_o + m_s$ constraints in (4.5). We reformulate the model in the form of (P) by setting $f := 0$, $D := \{0\}$ and a constraint function $c\colon \mathbb{R}^{N\times N} \to \mathbb{R}^m$ returning the observation and symmetry constraints stacked in vector form.

For our experiments, we chose $N \in \{10, 20\}$, $\ell = 5$, $m_o = \lfloor (n - m_s)/3 \rfloor$, $p = 0.5$ and consider 30 randomly generated instances for each value of N. We generate $X \in \mathbb{R}^{N\times\ell}$ by sampling the standard normal distribution, i.e., $X_{ij} \sim \mathcal{N}(0, 1)$, $(i, j) \in \mathcal{I}^2$, and then compute Δ. Finally, we sample observations by selecting m_o different entries of Δ with uniform probability.

We run our solver ALS with default options, and abstain from setting an iteration limit for the subproblem solver. The initial guess $B^0 \in \mathbb{R}^{N\times N}$ is chosen randomly

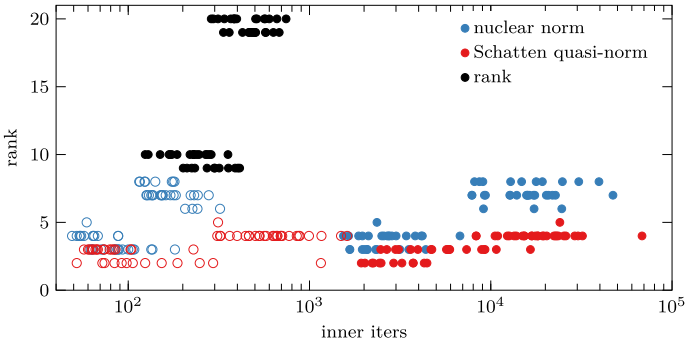


Fig. 5 Results for the matrix recovery problem (4.5): Comparison of (accumulated) inner iteration numbers and rank of the solutions found with different formulations, including warm-started rank minimization (circle)

based on $B^0_{ij} \sim \mathcal{N}(0, 1)$, $(i, j) \in \mathcal{I}^2$, whereas the dual initial guess is fixed to $y^0 := 0$. We invoke ALS directly for solving (4.5) with the different cost functions mentioned above. Additionally, the solutions obtained with nuclear norm and Schatten quasi-norm as cost functions, which are at least continuous, are used as initial guesses for another round of minimization exploiting the discontinuous rank functional.

We depict the results of our experiments in Fig. 5. Minimization based on the (convex) nuclear norm produces matrices with rank between 3 and 8, while the use of the Schatten quasi-norm culminates in solutions having rank between 2 and 5. These findings outperform the direct minimization of the rank which results in matrices of rank between 9 and 20. This behavior is not surprising since (4.5) possesses plenty of non-global minimizers in case where minimization of the discontinuous rank is considered, and ALS can terminate in such solutions. Let us mention that, out of 60 instances, the warm-started rank minimization yields further reduction of the rank in one case after minimization of the Schatten quasi-norm and 11 cases after minimization of the nuclear norm; in all other cases, no deterioration has been observed. In summary, ALS manages to find feasible solutions of (4.5) in all cases, and with adequate objective value in cases where we minimize the nuclear norm or the Schatten quasi-norm. These solutions can be used as initial guesses for a warm-started minimization of the rank via ALS or tailored mixed-integer numerical methods.

5 Conclusions

We presented the class of constrained composite optimization problems and proposed a general-purpose solver based on an augmented Lagrangian method. The (outer) augmented Lagrangian loop generates a sequence of subproblems, each one being a dual proximal regularization of the original, that can be solved, e.g., by off-the-shelf proximal algorithms for composite optimization. Requiring only first-order primitives, such as gradient and proximal mapping oracles, and projections onto the constraint set, the method is matrix-free and allows the seamless integration of routines for special problem structures. The proposed method is easily warm started to reduce the number of iterations and can take advantage of accelerated methods.

We have implemented our algorithm in the open-source Augmented Lagrangian Solver (ALS), disentangled from modeling tools and subproblem solvers. Thanks to its low memory footprint and simple, yet fast and robust iterations, ALS can handle large-scale problems and is suitable for embedded applications. We tested our approach numerically with problems arising in mixed-integer optimal control, sparse portfolio optimization and minimum-rank matrix completion. Illustrative examples showed the flexibility and descriptive power of constrained composite programs and the impact of accelerated methods for solving the inner problems.

Acknowledgements Alberto De Marchi is grateful to Andreas Themelis (Kyushu University), for sharing his insight and rigour, and to Matthias Gerdts (Universität der Bundeswehr München), for the support and guidance. The authors wish to thank two anonymous referees for their detailed comments and constructive suggestions, which significantly shaped and improved the quality of this work.

Funding Open Access funding enabled and organized by Projekt DEAL. Xiaoxi Jia and Christian Kanzow acknowledge support by the German Research Foundation (DFG) within the priority program *Non-smooth and Complementarity-based Distributed Parameter Systems: Simulation and Hierarchical Optimization* (SPP 1962) under Grant Numbers KA 1296/24-2.

Code and data availability The implementation of the solver and the datasets used, generated, and analyzed during the current study are openly available on GitHub at https://github.com/aldma/Bazinga.jl (September 2022).

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article.

A On the subproblem solver

In this appendix, we briefly describe the algorithm PANOC$^+$ from [24], which is used as a subproblem solver in Algorithm 1, and discuss some of its properties.

Let us consider the abstract unconstrained, composite optimization problem

$$\underset{z\in\mathbb{R}^p}{\text{minimize}} \quad \omega(z) := \varphi(z) + \psi(z) \tag{Q}$$

under the following standing assumption.

Assumption II The following hold in (Q):

(i) $\varphi\colon \mathbb{R}^p \to \mathbb{R}$ is continuously differentiable with locally Lipschitz continuous gradient;

(ii) $\psi : \mathbb{R}^p \to \overline{\mathbb{R}}$ is proper, lower semicontinuous and prox-bounded with threshold $\gamma_\psi > 0$;
(iii) $\inf_{z\in\mathbb{R}^p} \omega(z) > -\infty$.

For simplicity of notation, we introduce a set-valued mapping $\mathbf{T}_\gamma : \mathbb{R}^p \rightrightarrows \mathbb{R}^p$ for arbitrary $\gamma \in (0, \gamma_\psi)$ by means of

$$\mathbf{T}_\gamma(z) := \operatorname{prox}_{\gamma\psi}(z - \gamma\nabla\varphi(z)). \tag{A.1}$$

Furthermore, the algorithm makes use of the so-called *forward-backward envelope* (FBE) relative to (Q) with stepsize $\gamma \in (0, \gamma_\psi)$ given by

$$\omega_\gamma^{\mathrm{FB}}(z) := \min_{w\in\mathbb{R}^p} \varphi(z) + \langle\nabla\varphi(z), w - z\rangle + \psi(w) + \tfrac{1}{2\gamma}\|w - z\|^2.$$

Clearly, for any $\bar{z} \in \mathbf{T}_\gamma(z)$, we have

$$\omega_\gamma^{\mathrm{FB}}(z) = \varphi(z) + \langle\nabla\varphi(z), \bar{z} - z\rangle + \psi(\bar{z}) + \tfrac{1}{2\gamma}\|\bar{z} - z\|^2. \tag{A.2}$$

In Algorithm 2, we provide the pseudo code for PANOC$^+$, whose peculiarity is the intricate structure emerging at steps 2.5 and 2.7. The two backtracking linesearches are entangled, concurrently affecting both the direction stepsize τ_k and the proximal stepsize γ_k. These persistent adjustments allow PANOC$^+$ to construct a tighter merit function $\omega_\gamma^{\mathrm{FB}}$ that better captures the (local) landscape of ω, obviating the need for global Lipschitz gradient continuity for the smooth term in (Q).

Algorithm 2 PANOC$^+$ [24]

REQUIRE $z^0 \in \mathbb{R}^p$, $\gamma_0 \in (0, \gamma_\psi)$, $\Delta \geq 0$, $\alpha, \beta \in (0, 1)$, $\varepsilon > 0$
INITIALIZE $k \leftarrow 0$, and start from step 2.4
2.1: $\gamma_k \leftarrow \gamma_{k-1}$
2.2: Select an update direction $d^k \in \mathbb{R}^p$ with $\|d^k\| \leq \Delta\|\bar{z}^{k-1} - z^{k-1}\|$ and set $\tau_k = 1$
2.3: Set $z^k = (1 - \tau_k)\bar{z}^{k-1} + \tau_k(z^{k-1} + d^k)$
2.4: Compute $\bar{z}^k \in \mathbf{T}_{\gamma_k}(z^k)$ and set $\Phi_k := \omega_{\gamma_k}^{\mathrm{FB}}(z^k)$ as in (A.2)
2.5: IF $\varphi(\bar{z}^k) > \varphi(z^k) + \left\langle\nabla\varphi(z^k), \bar{z}^k - z^k\right\rangle + \frac{\alpha}{2\gamma_k}\|\bar{z}^k - z^k\|^2$ THEN
$\gamma_k \leftarrow \gamma_k/2$, and go back to step 2.2 if $k > 0$, or step 2.4 if $k = 0$
2.6: IF $\|\frac{1}{\gamma_k}(\bar{z}^k - z^k) - \nabla\varphi(\bar{z}^k) + \nabla\varphi(z^k)\| \leq \varepsilon$ THEN
RETURN $\bar{z}^k$
2.7: IF $k > 0$ AND $\Phi_k > \Phi_{k-1} - \beta\frac{1-\alpha}{2\gamma_{k-1}}\|\bar{z}^{k-1} - z^{k-1}\|^2$ THEN
$\tau_k \leftarrow \tau_k/2$ and go back to step 2.3
2.8: $k \leftarrow k + 1$ and start the next iteration at step 2.1

The analysis in [24] provides global convergence guarantees for PANOC$^+$ under Assumption II. Let us recall the basic result associated with Algorithm 2 that is important in the context of Algorithm 1. For the reader's convenience, we present a brief proof of the result as it is not explicitly stated in [24].

Proposition A.1 *Let $\{z^k\}$ and $\{\bar{z}^k\}$ be sequences generated by Algorithm 2. Furthermore, let $z^\star$ be an accumulation point of $\{z^k\}$ and $\{z^k\}_K$ a subsequence such that $z^k \to_K z^\star$. Then, $z^\star$ is a stationary point of ω. Additionally, $\bar{z}^k \to_K z^\star$ holds, and for each $\varepsilon > 0$ and any large enough $k \in K$, $\bar{z}^k$ is an ε-stationary point of ω.*

Proof Owing to [24, Thm 4.3], we have $\bar{z}^k \to_K z^\star$, and $\gamma_k = \gamma$ holds for some $\gamma > 0$ and large enough $k \in K$. Furthermore, this result gives boundedness of the expressions

$$\Phi_k := \varphi(z^k) + \left\langle \nabla\varphi(z^k), \bar{z}^k - z^k \right\rangle + \psi(\bar{z}^k) + \tfrac{1}{2\gamma_k}\|\bar{z}^k - z^k\|^2,$$

so that taking the lower limit $k \to_K \infty$ yields $z^\star \in \operatorname{dom}\psi$. Next, step 2.4 of Algorithm 2 yields

$$\begin{aligned}
\omega(z^\star) &\le \liminf_{k\to_K\infty} \Phi_k \\
&\le \liminf_{k\to_K\infty} \left(\varphi(z^k) + \left\langle \nabla\varphi(z^k), z^\star - z^k \right\rangle + \psi(z^\star) + \tfrac{1}{2\gamma_k}\|z^\star - z^k\|^2\right) \\
&\le \limsup_{k\to_K\infty} \left(\varphi(z^k) + \left\langle \nabla\varphi(z^k), z^\star - z^k \right\rangle + \psi(z^\star) + \tfrac{1}{2\gamma_k}\|z^\star - z^k\|^2\right) \\
&= \omega(z^\star),
\end{aligned}$$

giving $\bar{z}^k \overset{\omega}{\to}_K z^\star$ by continuity of φ. Considering the stationarity condition resulting from evaluation of the proximal map $\mathbf{T}_{\gamma_k}$,

$$0 \in \nabla\varphi(z^k) + \partial\psi(\bar{z}^k) + \tfrac{1}{\gamma_k}(\bar{z}^k - z^k)$$

holds for each $k \in K$, giving

$$\tfrac{1}{\gamma_k}(z^k - \bar{z}^k) + \nabla\varphi(\bar{z}^k) - \nabla\varphi(z^k) \in \nabla\varphi(\bar{z}^k) + \partial\psi(\bar{z}^k) = \partial\omega(\bar{z}^k).$$

Taking the limit $k \to_K \infty$ while respecting continuous differentiablity of φ, the result follows. □

Let us mention that slightly weaker convergence guarantees can be obtained for PANOC$^+$ whenever the evaluation of the proximal mapping $\mathbf{T}_{\gamma_k}$ in step 2.4 of Algorithm 2 is done inexactly, see [24, §4] for details.

Finally, in light of Sect. 4, we shall comment on the acceleration mechanism in PANOC$^+$. Although robust to arbitrary choices of (bounded) directions d^k, the practical performance of Algorithm 2 is strongly affected by the specific selection; we refer to [58, §4.3] for an overview on some potential update directions. In the numerical experiments, we consider two strategies for executing step 2.2 of Algorithm 2. First, we may select $d^k := \bar{z}^{k-1} - z^{k-1}$, so that $z^k = \bar{z}^{k-1}$, effectively reducing the algorithm to an adaptive proximal gradient method, without any acceleration [24, §4.4]. Second, as a baseline, we use the default acceleration strategy in ProximalAlgorithms.jl,

namely LBFGS directions with memory 5. Inspired by quasi-Newton methods, these are recursively constructed by keeping memory of pairs $z^{k+1} - z^k$ and $r^{k+1} - r^k$, with $r^k := z^k - \bar{z}^k$, and retrieving $d^k := -H^k r^k$ by simply performing scalar products [40]. Herein, the linear operator H_k mimics the (inverse) fixed-point residual mapping associated to the splitting scheme in a neighborhood of z^k [56, 59]. Notice that, as the geometry of the residual mapping depends on the proximal stepsize, (the memory of) the LBFGS approximation is reset every time the stepsize is adapted [24, §3.1].

References

1. Andreani, R., Birgin, E.G., Martínez, J.M., Schuverdt, M.L.: On augmented Lagrangian methods with general lower-level constraints. SIAM J. Optim. **18**(4), 1286–1309 (2008). https://doi.org/10.1137/060654797
2. Andreani, R., Haeser, G., Mito, L.M., Ramos, A., Secchin, L.D.: On the best achievable quality of limit points of augmented Lagrangian schemes. Numer. Algorithms **90**(2), 851–877 (2022). https://doi.org/10.1007/s11075-021-01212-8
3. Andreani, R., Martínez, J.M., Ramos, A., Silva, P.J.S.: A cone-continuity constraint qualification and algorithmic consequences. SIAM J. Optim. **26**(1), 96–110 (2016). https://doi.org/10.1137/15M1008488
4. Antil, H., Kouri, D.P., Ridzal, D.: ALESQP: An augmented Lagrangian equality-constrained SQP method for optimization with general constraints. http://www.optimization-online.org/DB_HTML/2021/01/8232.html (2020)
5. Armand, P., Tran, N.N.: Rapid infeasibility detection in a mixed logarithmic barrier-augmented Lagrangian method for nonlinear optimization. Optim. Methods Softw. **34**(5), 991–1013 (2019). https://doi.org/10.1080/10556788.2018.1528250
6. Balas, E.: Disjunctive Programming. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00148-3
7. Bauschke, H.H., Bolte, J., Teboulle, M.: A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications. Math. Oper. Res. **42**(2), 330–348 (2017). https://doi.org/10.1287/moor.2016.0817
8. Bauschke, H.H., Combettes, P.L.: Convex Analysis and Monotone Operator Theory in Hilbert Spaces. Springer, New York (2011). https://doi.org/10.1007/978-1-4419-9467-7
9. Beck, A.: First-Order Methods in Optimization. Society for Industrial and Applied Mathematics, Philadelphia, PA (2017). https://doi.org/10.1137/1.9781611974997
10. Beck, A., Hallak, N.: Optimization problems involving group sparsity terms. Math. Program. **178**(1), 39–67 (2019). https://doi.org/10.1007/s10107-018-1277-1
11. Benko, M., Mehlitz, P.: On implicit variables in optimization theory. J. Nonsmooth Anal. Optim. **2**, 7215 (2021). https://doi.org/10.46298/jnsao-2021-7215
12. Bertsekas, D.P.: Constrained Optimization and Lagrange Multiplier Methods. Athena Scientific, Nashua (1996)
13. Bezanson, J., Edelman, A., Karpinski, S., Shah, V.B.: Julia: A fresh approach to numerical computing. SIAM Rev. **59**(1), 65–98 (2017). https://doi.org/10.1137/141000671
14. Birgin, E.G., Martínez, J.M.: Augmented Lagrangian method with nonmonotone penalty parameters for constrained optimization. Comput. Optim. Appl. **51**(3), 941–965 (2012). https://doi.org/10.1007/s10589-011-9396-0
15. Birgin, E.G., Martínez, J.M.: Practical Augmented Lagrangian Methods for Constrained Optimization. Society for Industrial and Applied Mathematics, Philadelphia, PA (2014)
16. Börgens, E., Kanzow, C., Mehlitz, P., Wachsmuth, G.: New constraint qualifications for optimization problems in Banach spaces based on asymptotic KKT conditions. SIAM J. Optim. **30**(4), 2956–2982 (2020). https://doi.org/10.1137/19M1306804
17. Burke, J.V., Curtis, F.E., Wang, H.: A sequential quadratic optimization algorithm with rapid infeasibility detection. SIAM J. Optim. **24**(2), 839–872 (2014). https://doi.org/10.1137/120880045
18. Chen, X., Guo, L., Lu, Z., Ye, J.J.: An augmented Lagrangian method for non-Lipschitz nonconvex programming. SIAM J. Numer. Anal. **55**, 168–193 (2017). https://doi.org/10.1137/15M1052834

19. Combettes, P.L., Pesquet, J.C.: Proximal Splitting Methods in Signal Processing, pp. 185–212. Springer, New York (2011)
20. Conn, A.R., Gould, N.I.M., Toint, P.L.: A globally convergent augmented Lagrangian algorithm for optimization with general constraints and simple bounds. SIAM J. Numer. Anal. **28**(2), 545–572 (1991). https://doi.org/10.1137/0728030
21. De Marchi, A.: Constrained and sparse switching times optimization via augmented Lagrangian proximal methods. In: 2020 American Control Conference (ACC), pp. 3633–3638 (2020). https://doi.org/10.23919/ACC45564.2020.9147892
22. De Marchi, A.: Augmented Lagrangian and proximal methods for constrained structured optimization. Ph.D. thesis, Universität der Bundeswehr München (2021). https://doi.org/10.5281/zenodo.4972536
23. De Marchi, A.: On a primal-dual Newton proximal method for convex quadratic programs. Comput. Optim. Appl. **81**, 369–395 (2022). https://doi.org/10.1007/s10589-021-00342-y
24. De Marchi, A., Themelis, A.: Proximal gradient algorithms under local Lipschitz gradient continuity. J. Optim. Theory Appl. **194**(3), 771–794 (2022). https://doi.org/10.1007/s10957-022-02048-5
25. Dhingra, N.K., Khong, S.Z., Jovanović, M.R.: The proximal augmented Lagrangian method for nonsmooth composite optimization. IEEE Trans. Autom. Control **64**(7), 2861–2868 (2019). https://doi.org/10.1109/TAC.2018.2867589
26. Evens, B., Latafat, P., Themelis, A., Suykens, J., Patrinos, P.: Neural network training as an optimal control problem: An augmented Lagrangian approach. In: 60th IEEE Conference on Decision and Control (CDC), pp. 5136–5143 (2021). https://doi.org/10.1109/CDC45484.2021.9682842
27. Feng, M., Mitchell, J.E., Pang, J.S., Shen, X., Wächter, A.: Complementarity formulations of ℓ_0-norm optimization problems. Pac. J. Optim. **14**(2), 273–305 (2018)
28. Frangioni, A., Gentile, C.: SDP diagonalizations and perspective cuts for a class of nonseparable MIQP. Oper. Res. Lett. **35**(2), 181–185 (2007). https://doi.org/10.1016/j.orl.2006.03.008
29. Frangioni, A., Gentile, C.: The Mean-Variance portfolio problem. https://commalab.di.unipi.it/datasets/MV/ (2021). Accessed 20 Sep 2022
30. Gill, P.E., Robinson, D.P.: A primal–dual augmented Lagrangian. Comput. Optim. Appl. **51**(1), 1–25 (2012). https://doi.org/10.1007/s10589-010-9339-1
31. Grapiglia, G.N., Yuan, Y.: On the complexity of an augmented Lagrangian method for nonconvex optimization. IMA J. Numer. Anal. **41**(2), 1546–1568 (2020). https://doi.org/10.1093/imanum/draa021
32. Guo, L., Deng, Z.: A new augmented Lagrangian method for MPCCs—theoretical and numerical comparison with existing augmented Lagrangian methods. Math. Oper. Res. **47**(2), 1229–1246 (2022). https://doi.org/10.1287/moor.2021.1165
33. Guo, L., Ye, J.J.: Necessary optimality conditions and exact penalization for non-Lipschitz nonlinear programs. Math. Program. **168**(1), 571–598 (2018). https://doi.org/10.1007/s10107-017-1112-0
34. Hestenes, M.R.: Multiplier and gradient methods. J. Optim. Theory Appl. **4**(5), 303–320 (1969). https://doi.org/10.1007/BF00927673
35. IBM ILOG CPLEX: V12. 1: User's Manual for CPLEX. International Business Machines Corporation **46**(53), 157 (2009)
36. Jia, X., Kanzow, C., Mehlitz, P., Wachsmuth, G.: An augmented Lagrangian method for optimization problems with structured geometric constraints. Math. Program. (2022). https://doi.org/10.1007/s10107-022-01870-z
37. Kanzow, C., Mehlitz, P.: Convergence properties of monotone and nonmonotone proximal gradient methods revisited. J. Optim. Theory Appl. **195**(2), 624–646 (2022). https://doi.org/10.1007/s10957-022-02101-3
38. Kanzow, C., Steck, D., Wachsmuth, D.: An augmented Lagrangian method for optimization problems in Banach spaces. SIAM J. Control. Optim. **56**(1), 272–291 (2018). https://doi.org/10.1137/16M1107103
39. Kruger, A.Y., Mehlitz, P.: Optimality conditions, approximate stationarity, and applications—a story beyond Lipschitzness. ESAIM: Control Optim. Calc. Var. **28**, 42 (2022). https://doi.org/10.1051/cocv/2022024
40. Liu, D.C., Nocedal, J.: On the limited memory BFGS method for large scale optimization. Math. Program. **45**, 503–528 (1989). https://doi.org/10.1007/BF01589116
41. Ma, D., Judd, K.L., Orban, D., Saunders, M.A.: Stabilized optimization via an NCL algorithm. In: Al-Baali, M., Grandinetti, L., Purnama, A. (eds.) Numerical Analysis and Optimization, pp. 173–191. Springer, Berlin (2018). https://doi.org/10.1007/978-3-319-90026-1_8
42. Mehlitz, P.: Asymptotic stationarity and regularity for nonsmooth optimization problems. J. Nonsmooth Anal. Optim. **1**, 6575 (2020). https://doi.org/10.46298/jnsao-2020-6575

43. Mehlitz, P.: A comparison of first-order methods for the numerical solution of or-constrained optimization problems. Comput. Optim. Appl. **76**, 233–275 (2020). https://doi.org/10.1007/s10589-020-00169-z
44. Mordukhovich, B.S.: Variational Analysis and Generalized Differentiation, Part I: Basic Theory, Part II: Applications. Springer, Berlin (2006)
45. Moreau, J.J.: Proximité et dualité dans un espace hilbertien. Bulletin de la Société Mathématique de France **93**, 273–299 (1965). https://doi.org/10.24033/bsmf.1625
46. Parikh, N., Boyd, S.: Proximal algorithms. Found. Trends Optim. **1**(3), 127–239 (2014). https://doi.org/10.1561/2400000003
47. Potschka, A., Bock, H.G.: A sequential homotopy method for mathematical programming problems. Math. Program. **187**(1), 459–486 (2021). https://doi.org/10.1007/s10107-020-01488-z
48. Powell, M.J.D.: A Method for Nonlinear Constraints in Minimization Problems, pp. 283–298. Academic Press, London (1969)
49. Rockafellar, R.T.: Augmented Lagrangians and applications of the proximal point algorithm in convex programming. Math. Oper. Res. **1**(2), 97–116 (1976). https://doi.org/10.1287/moor.1.2.97
50. Rockafellar, R.T., Wets, R.J.B.: Variational Analysis, vol. 317. Springer, Berlin (1998)
51. Sager, S.: Numerical methods for mixed-integer optimal control problems. Ph.D. thesis, University of Heidelberg (2005). Interdisciplinary Center for Scientific Computing
52. Shen, X., Mitchell, J.E.: A penalty method for rank minimization problems in symmetric matrices. Comput. Optim. Appl. **71**(2), 353–380 (2018). https://doi.org/10.1007/s10589-018-0010-6
53. Sopasakis, P., Fresk, E., Patrinos, P.: OpEn: Code generation for embedded nonconvex optimization. IFAC-PapersOnLine **53**(2), 6548–6554 (2020). https://doi.org/10.1016/j.ifacol.2020.12.071. (**21st IFAC World Congress**)
54. Sra, S., Nowozin, S., Wright, S.J.: Optimization for Machine Learning. Neural Information Processing Series. MIT Press, Cambridge (2011)
55. Stella, L.: ProximalAlgorithms.jl: Proximal algorithms for nonsmooth optimization in Julia. https://github.com/JuliaFirstOrder/ProximalAlgorithms.jl
56. Stella, L., Themelis, A., Sopasakis, P., Patrinos, P.: A simple and efficient algorithm for nonlinear model predictive control. In: 56th IEEE Conference on Decision and Control (CDC), pp. 1939–1944 (2017). https://doi.org/10.1109/CDC.2017.8263933
57. Stellato, B., Ober-Blöbaum, S., Goulart, P.J.: Second-order switching time optimization for switched dynamical systems. IEEE Trans. Autom. Control **62**(10), 5407–5414 (2017). https://doi.org/10.1109/TAC.2017.2697681
58. Themelis, A.: Proximal algorithms for structured nonconvex optimization. Ph.D. thesis, KU Leuven, Arenberg Doctoral School, Faculty of Engineering Science (2018)
59. Themelis, A., Stella, L., Patrinos, P.: Forward-backward envelope for the sum of two nonconvex functions: Further properties and nonmonotone linesearch algorithms. SIAM J. Optim. **28**(3), 2274–2303 (2018). https://doi.org/10.1137/16M1080240
60. Wright, S.J., Recht, B.: Optimization for Data Analysis. Cambridge University Press, Cambridge (2022). https://doi.org/10.1017/9781009004282