Deutsche
Forschungsgemeinschaft
Priority Programme 1962

Constrained Structured Optimization and Augmented Lagrangian Proximal Methods

Alberto De Marchi, Xiaoxi Jia, Christian Kanzow, Patrick Mehlitz

SPP 1962

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

Preprint Number SPP1962-191

received on April 7, 2022

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

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

Constrained Structured Optimization and Augmented Lagrangian Proximal Methods

Alberto De Marchi[*] Xiaoxi Jia[†] Christian Kanzow[†]
Patrick Mehlitz[‡]

Abstract

We investigate and develop numerical methods for finite-dimensional constrained structured optimization problems. Offering a comprehensive yet simple and expressive language, this problem class provides a modeling framework for a variety of applications. A general and flexible algorithm is proposed that interlaces proximal methods and safeguarded augmented Lagrangian schemes. We provide a theoretical characterization of the algorithm and its asymptotic properties, deriving convergence results for fully nonconvex problems. Adopting a proximal gradient method with an oracle as a formal tool, it is demonstrated how the inner subproblems can be solved by off-the-shelf methods for composite optimization, without introducing slack variables and despite the appearance of set-valued projections. Finally, we describe our open-source matrix-free implementation of the proposed algorithm and test it numerically. Illustrative examples show the versatility of constrained structured programs as a modeling tool, expose difficulties arising in this vast problem class and highlight benefits of the implicit approach developed.

Keywords. Nonsmooth nonconvex optimization, nonlinear programming, augmented Lagrangian methods, proximal algorithms
AMS subject classifications. 49J53, 65K05, 90C30

1 Introduction

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

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

[*] Universität der Bundeswehr München, Department of Aerospace Engineering, 85577 Neubiberg/Munich, Germany. EMAIL alberto.demarchi@unibw.de, ORCID 0000-0002-3545-6898

[†] University of Würzburg, Institute of Mathematics, 97074 Würzburg, Germany

[‡] Brandenburgische Technische Universität Cottbus-Senftenberg, Institute of Mathematics, 03046 Cottbus, Germany. Universität Mannheim, School of Business Informatics and Mathematics, 68159 Mannheim, Germany. ORCID 0000-0002-9355-850X

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 structured optimization problem because it contains set-membership constraints and a structured (or 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 disconnected. Thanks to their rich structure and flexibility, constrained structured problems are of interest for modeling in a variety of applications, ranging from optimal and model predictive control [19, 48], to signal processing [15], low-rank and sparse approximation, compressed sensing, cardinality-constrained optimization [6], and disjunctive programming [5], such as problems with complementarity, vanishing and switching constraints [32, 38].

Augmented Lagrangian and proximal methods have recently attracted revived and grown interest. Tracing back to the classical work of Hestenes [31] and Powell [43], the augmented Lagrangian framework can tackle large-scale constrained problems, whereas nonsmooth and extended-real valued cost functions are easily treated by proximal algorithms, inaugurated by Moreau [40]. Recent accounts on these topics can be found in [8, 11, 16] and [15, 41, 52], 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" [11, §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 and to deal with constraints without softening them; cf. [48, 50]. In the context of constrained structured programming, proximal methods play a key role, since the augmented Lagrangian subproblems for (P) are in the form of structured optimization problems.

The close relationship between augmented Lagrangian and proximal methods is well known and traces back to Rockafellar [44]. These approaches have been combined in [23] to deal with structured 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 structured programs in [20, Chapter 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 [22, 33]. 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 convex constraint set; 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 [23, 44]. This procedure is employed to eliminate the slack variables and, in the convex setting, 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. Hence, careful handling is required to avoid ruining the problem structure, or rather to exploit it, when solving the (proximal) augmented Lagrangian subproblems, as we will show.

Introducing slack variables corresponds to making explicit some traits implicit in the original problem formulation; see [7] for an investigation on the role of implicit variables in optimization. Therein the authors state that "the implicit formulation avoids the appearance of artificial local minimizers [...]. It is, thus, always desirable to explore the inherent problem structure of the original problem instead of making its implicit variables explicit" [7, §6]. Not only implicit formulations yield stronger theoretical results, but also subproblems with fewer decision variables. Thus, being able to harness the proximal augmented Lagrangian approach in the fully nonconvex setting might be beneficial from both theoretical and numerical perspectives.

The contribution of this work touches several aspects. We investigate the abstract class of constrained structured 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 convergence analysis. Patterning this methodology, analogous algorithms and theoretical results can be derived based on other augmented Lagrangian schemes. Further, we demonstrate there is no need for slack variables, nor for special choices of possibly set-valued projections and proximal mappings. Finally, we show that it is possible to adopt off-the-shelf, yet adaptive, proximal gradient methods for solving the augmented Lagrangian subproblems.

The following blanket assumptions are considered throughout, without further mention. Technical definitions are given in Section 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 [22, 33].

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 it is always possible to project onto the constraint set D. Nevertheless, these conditions do not imply the existence of feasible points for (P). As it is the case in nonlinear programming [11], we will study the minimization properties of the augmented Lagrangian scheme with respect to some infeasibility measure. Owing to Assumption I*(ii)*, the constraint set D could be assumed convex without loss of generality with respect to (P): one can obtain an equivalent problem by introducing slack variables and including an indicator function in the objective. However, since this reformulation enlarges the problem size, we are going to consider it only as a theoretical tool to derive formal results, while focusing on the fully nonconvex setting. Moreover, 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 presented in the following 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 [8, 11, 16, 48], also in the infinite-dimensional setting [3, 34].

Merely lower semicontinuous cost functions have been considered in [24]. Inspired by [28, Alg. 1] and leveraging the idea behind [11, Ex. 4.12], the convergence properties of [24, 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 [24, 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 [12, 32] and, for the special case of so-called complementarity constraints, in [29]. 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*. A similar structure can be obtained from (P) by introducing slack variables. Moreover, as pointed out in [32, §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 q. These observations imply that constrained structured programs do not generalize the problem class considered in [32]. Nevertheless, the necessary reformulations come at a price: increased problem size due to slack variables and the need for projections onto the epigraph of q. The proximal 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 slack variables nor for oracles other than those mentioned above. Analogous considerations hold for [14], dedicated to an augmented Lagrangian method for non-Lipschitz nonlinear programs, and [35, Section 6.2], where the solution of the augmented Lagrangian subproblems is not discussed.

The work presented in this paper collects and builds upon the ideas put forward in [20]. However, we consider different stationarity concepts and necessary optimality conditions, not based on the proximal operator as in [20, §1.2], but rather exploiting tools from variational analysis; see [30, 32, 35, 37]. Furthermore, patterning [22] and thanks to a detailed analysis of the subproblems, we offer a rigorous convergence analysis, as well as theoretical justifications for adopting accelerated proximal gradient methods; cf. [20, §1.5.4].

2 Setting 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. The vectors in $\mathbb{R}^n$ with all elements equal to 0 or 1 are denoted as 0_n and 1_n; whenever n is clear from context we simply write 0 and 1, respectively. The effective domain of an extended-real-valued function $h\colon \mathbb{R}^n \to \overline{\mathbb{R}}$ is denoted by $\operatorname{dom} h := \{x \in \mathbb{R}^n \mid h(x) < \infty\}$. We say that h is *proper* if $\operatorname{dom} h \neq \emptyset$ and *lower semicontinuous* (lsc) if $h(\bar{x}) \leq \liminf_{x\to\bar{x}} h(x)$ for all $\bar{x} \in \mathbb{R}^n$. For some constant $\tau \in \mathbb{R}$, $\operatorname{lev}_{\leq\tau} h := \{x \in \mathbb{R}^n \mid h(x) \leq \tau\}$ denotes the τ-*sublevel set* associated with h.

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

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

Following [45, 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 $\bar{x}$ [45, Ex. 8.8]. 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^* [45, Thm 10.1].

A mapping $S : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ 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 [45, 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

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

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

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

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

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 $\operatorname{prox}_{\gamma h}$ is locally bounded, nonempty- and compact-valued [45, 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

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

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

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

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

$$z \in \Pi_D(v) \quad \Rightarrow \quad v - z \in \mathcal{N}_D^{\lim}(z), \tag{2.3}$$

whereas the converse does not hold in general. 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

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

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

Lemma 2.1. *Let $D \subset \mathbb{R}^m$ be nonempty and closed. 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, for each $x^* \in \mathbb{R}^n$, we have*

$$\partial\vartheta(x^*) \subseteq \nabla c(x^*)^\top (c(x^*) - \Pi_D(c(x^*))).$$

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 $\vartheta = \psi \circ c$. From [39, Thm 1.110] and [45, Ex. 8.53], we find $\partial\psi(y^*) = y^* - \Pi_D(y^*)$ for all $y^* \in \mathbb{R}^m$. Thus, the subdifferential chain rule from [39, Thm 3.41] yields the claim. □

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 the 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. [37, §3] and [39, 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*

$$0 \in \partial q(x^*) + \nabla c(x^*)^\top y^* \tag{2.4a}$$

$$y^* \in \mathcal{N}_D^{lim}(c(x^*)). \tag{2.4b}$$

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 [32, §2.2] and [37, §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*

$$\eta^k \in \partial q(x^k) + \nabla c(x^k)^\top y^k \tag{2.5a}$$

$$y^k \in \mathcal{N}_D^{lim}(c(x^k) + \zeta^k) \tag{2.5b}$$

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

The definition of an AM-stationary point is similar to the notion of an asymptotic KKT (AKKT) point [11], 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 (2.5a) 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, will be important later on when taking the limit in (2.5a) 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 [35, Definition 6.4, Remark 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 [30] for a discussion. However, we can show that each local minimizer of (P) is always AM-stationary. Related results can be found in [35, Thm 6.2] and [37, §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

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & q(x) + \frac{1}{2}\|x - x^*\|^2 \\ \text{subject to} \quad & c(x) \in D, \quad x \in \mathbb{B}_\varepsilon(x^*). \end{aligned} \tag{2.6}$$

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 lower semicontinuous 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

$$q(x^k) + \frac{k}{2}\|c(x^k) - s^k\|^2 + \frac{1}{2}\|x^k - x^*\|^2 \leq q(x^*) \tag{2.7}$$

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 (2.7) gives $c(\tilde{x}) = \tilde{s} \in D$. Particularly, $\tilde{x}$ is feasible for (2.6). Therefore, the local optimality of x^* implies $q(x^*) \leq q(\tilde{x})$. Furthermore, we find

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

Hence, $\tilde{x} = x^*$, and noting that (2.7) 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 $\vartheta\colon \mathbb{R}^n \times \mathbb{R}^m \to \overline{\mathbb{R}}$ by means of $\vartheta(x, s) := g(x) + \delta_D(s)$ for each pair $(x, s) \in \mathbb{R}^n \times \mathbb{R}^m$. Applying [39, Prop. 1.107, 1.114], we find

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

for each $k \in \mathbb{N}$. The decoupled structure of ϑ and [39, Thm 3.36] yield the inclusion $\partial\vartheta(x^k, s^k) \subset \partial g(x^k) \times \mathcal{N}_D^{\lim}(s^k)$ 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 [37, §3.2] to the non-Lipschitzian setting and is closely related to the so-called *uniform qualification condition* introduced in [35, Definition 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*

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

Note that AM-regularity of some feasible point $x^* \in \mathbb{R}^n$ for (P) reduces to

$$\limsup_{\substack{x \to x^* \\ z \to 0}} \nabla c(x)^\top \mathcal{N}_D^{\lim}(c(x) - z) \subset \nabla c(x^*)^\top \mathcal{N}_D^{\lim}(c(x^*)) \tag{2.8}$$

whenever g is locally Lipschitz continuous around x^* since $x \rightrightarrows \partial g(x)$ is locally bounded at x^* in this case, see [39, Corollary 1.81]. We also observe that (2.8) corresponds to the concept of AM-regularity which has been used in [32, 37] 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 [35].

As a corollary of Proposition 2.5, we find the following result, along the lines of [35, Proposition 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.*

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 equivalently rewritten as

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

Notice that (P$_\text{S}$) is a particular problem in the form of (P). Moreover, if g is smooth, and thus so is q, then (P$_\text{S}$) falls into the problem class analyzed in [32]. We use the lifted reformulation (P$_\text{S}$) 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 $\mathcal{L}_\mu^\text{S} \colon \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^m \to \overline{\mathbb{R}}$ associated to (P$_\text{S}$) as

$$\begin{aligned}\mathcal{L}_\mu^\text{S}(x, s, y) &:= q(x) + \delta_D(s) + \langle y, c(x) - s\rangle + \frac{1}{2\mu}\|c(x) - s\|^2 \\ &= q(x) + \delta_D(s) + \frac{1}{2\mu}\|c(x) + \mu y - s\|^2 - \frac{\mu}{2}\|y\|^2. \end{aligned} \tag{3.1}$$

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, 11, 16, 32].

We now exploit 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 [20, 23]. Given some $\mu > 0$ and $y \in \mathbb{R}^m$, the explicit minimization of $\mathcal{L}_\mu^\text{S}(x, \cdot, y)$ yields the set-valued mapping $\mathcal{S}_\mu(\cdot, y) \colon \mathbb{R}^n \rightrightarrows \mathbb{R}^m$,

$$\mathcal{S}_\mu(x, y) := \operatorname*{arg\,min}_s \mathcal{L}_\mu^\text{S}(x, s, y) = \Pi_D\left(c(x) + \mu y\right). \tag{3.2}$$

Injecting back into $\mathcal{L}_\mu^\text{S}(x, \cdot, y)$ any arbitrary element of $\mathcal{S}_\mu(x, y)$, namely 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 $\mathcal{L}_\mu \colon \mathbb{R}^n \times \mathbb{R}^m \to \overline{\mathbb{R}}$ associated to (P):

$$\mathcal{L}_\mu(x, y) := \min_s \mathcal{L}_\mu^\text{S}(x, s, y) = q(x) + \frac{1}{2\mu}\operatorname{dist}_D^2(c(x) + \mu y) - \frac{\mu}{2}\|y\|^2. \tag{3.3}$$

We highlight that the term $\operatorname{dist}_D^2 \colon \mathbb{R}^m \to \mathbb{R}$ is not continuously differentiable in general, as the projection onto D is a set-valued mapping.

The algorithm we are about to present requires, at each (outer) iteration, the minimization of $\mathcal{L}_\mu(\cdot, y)$, given some $\mu > 0$ and $y \in \mathbb{R}^m$. 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 considering the joint minimization of $\mathcal{L}_\mu^{\mathrm{S}}(\cdot,\cdot,y)$, with respect to both, primal and slack variables. 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 arbitrary $s^* \in \mathcal{S}_\mu(x^*,y)$, is a local (global) minimizer of (P_S); cf. (3.2).

The subproblems are usually solved only approximately, in some sense, for the sake of computational efficiency. Given some tolerance $\varepsilon \geq 0$, we say that $x^* \in \mathbb{R}^n$ is an ε-approximate stationary (or ε-stationary) point for $\mathcal{L}_\mu(\cdot,y)$ if

$$\exists \eta \in \mathbb{R}^n,\ \|\eta\| \leq \varepsilon : \eta \in \partial_x \mathcal{L}_\mu(x^*,y). \tag{3.4}$$

By the local Lipschitzness of the squared distance function, [39, Thm 3.36] and Lemma 2.1, we have the inclusion

$$\partial_x \mathcal{L}_\mu(x^*,y) \subset \partial q(x^*) + \frac{1}{\mu}\nabla c(x^*)^\top[c(x^*) + \mu y - \Pi_D(c(x^*) + \mu y)]. \tag{3.5}$$

Furthermore, x^* is ε-stationary for $\mathcal{L}_\mu(\cdot,y)$ if (and only if) there exists $s^* \in \mathcal{S}_\mu(x^*,y)$ such that x^* is ε-stationary for $\mathcal{L}_\mu^{\mathrm{S}}(\cdot,s^*,y)$, since it is

$$\partial_x \mathcal{L}_\mu^{\mathrm{S}}(x^*,s^*,y) = \partial q(x^*) + \frac{1}{\mu}\nabla c(x^*)^\top[c(x^*) + \mu y - s^*]. \tag{3.6}$$

This shows that the two subproblem formulations are intimately connected and, in particular, that a slack variable s^* is needed in that it provides a certificate of (approximate) stationarity for some primal variable x^*.

The following Section 3.1 contains a detailed statement of our algorithmic framework, whose convergence analysis is presented in Section 3.2. Then, suitable termination criteria are discussed in Section 3.3. In Section 4 we consider the numerical solution of the subproblems and propose a method to solve $\min \mathcal{L}_\mu(\cdot,y)$ directly, that is, without slack variables.

3.1 Algorithm

This section presents an augmented Lagrangian method for the solution of constrained structured programs of the form (P), under Assumption I. 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 [8, 11, 16, 28, 34, 48], and for primal-dual methods [27]. In the following, we focus on a safeguarded augmented Lagrangian scheme inspired by [11, Alg. 4.1] and investigate its convergence properties. Compared to the classical augmented Lagrangian or multiplier penalty approach for the solution of nonlinear programs [8], 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.

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 $k = 0, 1, 2 \dots$
1.1:  Select $\hat{y}^k \in Y$ and $\varepsilon_k \geq 0$
1.2:  Compute an $\varepsilon_k$-approximate stationary point $x^k$ of $\mathcal{L}_{\mu_k}(\cdot, \hat{y}^k)$
1.3:  Select $s^k \in \mathcal{S}_{\mu_k}(x^k, \hat{y}^k)$ such that $x^k$ is $\varepsilon_k$-stationary for $\mathcal{L}^{\mathrm{S}}_{\mu_k}(\cdot, s^k, \hat{y}^k)$
1.4:  Set $y^k \leftarrow \hat{y}^k + [c(x^k) - s^k]/\mu_k$
1.5:  IF $k = 0$ or $\|c(x^k) - s^k\| \leq \theta \|c(x^{k-1}) - s^{k-1}\|$ THEN
1.6:      Set $\mu_{k+1} \leftarrow \mu_k$
1.7:  ELSE
1.8:      Select $\mu_{k+1} \in (0, \kappa\mu_k]$
```

The overall method is stated in Algorithm 1. 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}, 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. [32, 48]. 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 steps 1.2 and 1.3. Considering the proximal augmented Lagrangian function $\mathcal{L}_\mu$ instead of the lifted $\mathcal{L}^{\mathrm{S}}_\mu$ not only results in smaller subproblems, but also guarantees the exact, global optimality of the slack variable, since the inclusion $s \in \mathcal{S}_\mu(x, y)$ is satisfied by construction. Although a suitable s^k can be obtained at step 1.2, as it is implicitly employed to verify approximate stationarity, we prefer to display x^k and s^k as computed at different steps, thus highlighting that the subproblems at step 1.2 involve the primal variable x only. Conversely, step 1.3 requires no additional effort in practice. Section 4 is devoted to the numerical solution of the subproblems, discussing several approaches. Note that step 1.3 simplifies significantly if D is convex. In this case, one simply needs to find the (uniquely determined) element in $\mathcal{S}_\mu(x^k, \hat{y}^k)$ by evaluating the projection operator.

Step 1.4 entails the classical first-order Lagrange multiplier estimate. The update rule is designed around (3.6) and leads to the inclusion (2.5a) for the primal-dual estimate (x^k, y^k). The monotonicity test at step 1.5 is adopted to monitor primal infeasibility along the iterates. The penalty parameter is reduced at step 1.8 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 [44], 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

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

and then, by eliminating the slack variable s, as

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

The latter reformulation amounts to a proximal dual regularization of (P) and corresponds to a lifted representation of $\min \mathcal{L}_{\mu}(\cdot, y)$, 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., [21, 36, 42].

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^* for a sequence $\{x^k\}$ generated by Algorithm 1. In general, coercivity or (level) boundedness arguments should be adopted to verify this precondition.

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 [11, Chapter 5] and [34], see [35, §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.

Proposition 3.1 shows that the Lagrange multiplier vanishes for constraints that are inactive in the limit, independently of the feasibility of the limit point. Notice that the assertion can be easily refined by exploiting the separable structure of D, if available. For instance, considering a hyperbox, every step in the proof can be applied componentwise, recovering the classical result [11, Thm 4.1].

Proposition 3.1. *Let Assumption I hold and consider a sequence $\{x^k\}$ of iterates generated by Algorithm 1. Let x^* be an accumulation point of $\{x^k\}$ and $\{x^k\}_K$ a subsequence such that $x^k \to_K x^*$. If $c(x^*) \in \operatorname{int} D$, then $y^k = 0$ for all $k \in K$ large enough.*

Proof. Let $c(x^*) \in \operatorname{int} D \neq \emptyset$ and consider the two cases:

(i) If $\mu_k \to 0$, by boundedness of $\{\hat{y}^k\}$ and continuity of c, there exists $k_0 \in K$ such that $c(x^k) + \mu_k \hat{y}^k \in \operatorname{int} D$ for all $k \in K$, $k \geq k_0$.

(ii) If $\{\mu_k\}$ is bounded away from zero, it follows from steps 1.5 and 1.8 that $\|c(x^k) - s^k\| \to_K 0$. Then, by continuity of c and $x^k \to_K x^*$, $s^k \to_K c(x^*) \in \operatorname{int} D$ follows. Hence, there exists $k_1 \in K$ such that $s^k \in \operatorname{int} D$ for all $k \in K$, $k \geq k_1$.

In both cases, since $s^k \in \Pi_D(c(x^k) + \mu_k \hat{y}^k)$ by definition, we obtain $s^k = c(x^k) + \mu_k \hat{y}^k$ for all $k \in K$ large enough. Consequently, by step 1.4, it is $y^k = 0$. □

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. [10, Ex. 4.12], [32, Prop. 4.1]. A proof is included in the Additional Material (p. 34).

Proposition 3.2. *Let Assumption I hold and consider a sequence $\{x^k\}$ 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 $\mathcal{L}_{\mu_k}(x^k, \hat{y}^k) \leq B$ for all $k \in \mathbb{N}$.

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

Proof. From steps 1.2 to 1.4 of Algorithm 1, we have that

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

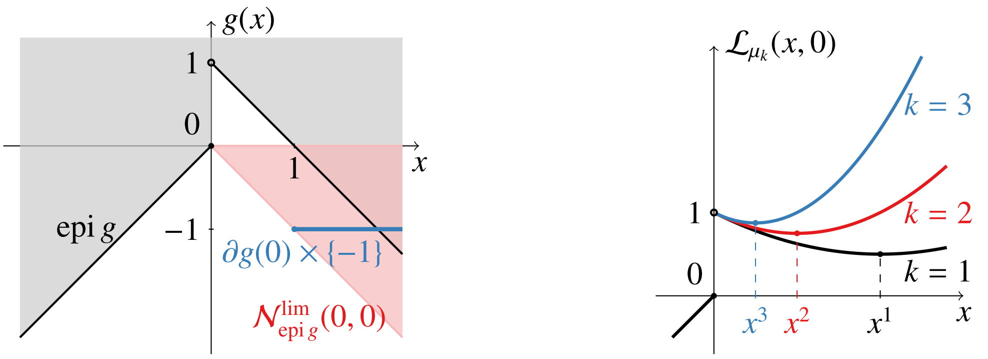
for some $\eta^k \in \mathbb{R}^n$, $\|\eta^k\| \leq \varepsilon_k$. Define $\zeta^k := s^k - c(x^k)$ for all $k \in \mathbb{N}$, where $s^k \in \mathcal{S}_{\mu_k}(x^k, \hat{y}^k)$ by step 1.3. We claim that the four subsequences $\{x^k\}_K$, $\{\eta^k\}_K$, $\{y^k\}_K$ and $\{\zeta^k\}_K$ generated by Algorithm 1 satisfy the properties in Definition 2.4 and therefore show that x^* is an AM-stationary point for (P).

By construction, we have $x^k \xrightarrow{q}_K x^*$ and $\|\eta^k\| \leq \varepsilon_k \to_K 0$. Further, from steps 1.3 and 1.4 of Algorithm 1, we obtain that, for all $k \in \mathbb{N}$,

$$y^k = \frac{c(x^k) + \mu_k \hat{y}^k - s^k}{\mu_k} \in \mathcal{N}_D^{\lim}(s^k) = \mathcal{N}_D^{\lim}(c(x^k) + \zeta^k), \tag{3.8}$$

where the inclusion follows from $s^k \in \mathcal{S}_{\mu_k}(x^k, \hat{y}^k)$, (2.3), and the cone property of $\mathcal{N}_D^{\lim}(s^k)$. It remains to show that $\zeta^k \to_K 0$. To this end, we consider two cases.

(i) If $\{\mu_k\}$ is bounded away from zero, steps 1.5 and 1.8 of Algorithm 1 imply that $\|\zeta^k\| = \|c(x^k) - s^k\| \to 0$ for $k \to \infty$.



(a) *Computation of* $\partial g(0)$.

(b) *Iterates* x^k *for* $k \in \{1,2,3\}$.

Figure 1: *Visualizations for Example 3.4.*

(ii) Instead, if $\mu_k \to 0$, we exploit the continuity of the distance function and, using the triangle inequality, we obtain that

$$\begin{aligned}\|\zeta^k\| &= \|c(x^k) + \mu_k \hat{y}^k - \mu_k \hat{y}^k - s^k\| \\ &\le \|c(x^k) + \mu_k \hat{y}^k - s^k\| + \|\mu_k \hat{y}^k\| \\ &= \operatorname{dist}_D(c(x^k) + \mu_k \hat{y}^k) + \|\mu_k \hat{y}^k\| \to_K 0,\end{aligned}$$

where the limit follows from boundedness of $\{\hat{y}^k\}$ and feasibility of x^*.

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

The additional assumption $x^k \xrightarrow{q}_K x^*$ 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 that this additional requirement is indispensable in a discontinuous setting.

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

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

Note that g is merely lsc at $x^* := 0$, and that $\partial g(x^*) = [1, \infty)$, cf. Figure 1a. Although x^* is the global maximizer of the associated problem (P), x^* is not an M-stationary point; in fact, x^* is not even AM-stationary. Since $\nabla f(x^*) = 0$, $\nabla c(x^*) = 1$ and $\mathcal{N}^{\lim}_D(c(x^*)) = [0, \infty)$, there is no $y^* \in \mathcal{N}^{\lim}_D(c(x^*))$ such that $0 \in \nabla f(x^*) + \partial g(x^*) + \nabla c(x^*)^\top y^*$.

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

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

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

We note that related results have been obtained in [14, Thm 3.1] and [35, Cor. 6.16]. In [14], 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. [11, Thm 6.3].

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

$$\operatorname*{minimize}_{(x,\alpha)\in\operatorname{epi} q} \quad \operatorname{dist}_D^2(c(x)). \tag{3.9}$$

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

$$\operatorname*{minimize}_{x} \quad \operatorname{dist}_D^2(c(x)). \tag{3.10}$$

Proof. By Proposition 3.2*(i)*, if $\{\mu_k\}$ is bounded away from zero, then each accumulation point x^* is feasible for (P), namely it is a global minimizer of the problems (3.9) as well as (3.10) and an M-stationary point thereof by Lipschitzianity of the objective function, see [39, Prop 5.3]. Hence, it remains to consider the case $\mu_k \to 0$.

Owing to steps 1.2 and 1.3 of Algorithm 1, for all $k \in K$ it is $s^k \in \mathcal{S}_{\mu_k}(x^k, \hat{y}^k) = \Pi_D(c(x^k) + \mu_k \hat{y}^k)$ and

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

for some $\eta^k \in \mathbb{R}^n$, $\|\eta^k\| \le \varepsilon_k$; cf. (3.6). This gives us

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

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

$$(\mu_k \eta^k - \nabla c(x^k)^\top [c(x^k) + \mu_k \hat{y}^k - s^k], -\mu_k) \in \mathcal{N}^{\lim}_{\operatorname{epi} q}(x^k, q(x^k)). \tag{3.11}$$

We note that by boundedness of $\{c(x^k) + \mu_k \hat{y}^k\}_K$, $\{s^k\}_K$ is also bounded by definition of the projection. Hence, we may assume $s^k \to_K s^*$ for some $s^* \in \mathbb{R}^m$ without loss of generality. Taking the limit $k \to_K \infty$ in (3.11), the robustness of the limiting normal cone and $x^k \xrightarrow{q}_K x^*$ yield

$$(-\nabla c(x^*)^\top [c(x^*) - s^*], 0) \in \mathcal{N}^{\lim}_{\operatorname{epi} q}(x^*, q(x^*)).$$

Observing that the graph of the set-valued projection Π_D is closed, we also have $s^* \in \Pi_D(c(x^*))$. Taking Lemma 2.1 into account, $(x^*, q(x^*))$ is an M-stationary point of (3.9).

Finally, assume that q is locally Lipschitz continuous at x^*. Then, due to [39, 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^* for (3.10). □

3.3 Termination Criteria

Steps 1.2 and 1.3 involve the minimization of the augmented Lagrangian function defined in (3.3). Then, the dual update at step 1.4 allows to draw conclusions with respect to the original problem (P), as shown by Theorem 3.3.

Owing to (3.7)–(3.8) and recalling the AM-stationarity conditions (2.5), one may select a zero 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

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

are satisfied. Theorem 3.3 demonstrates that these termination criteria (the latter, in particular) are satisfied in finitely many iterations if any subsequence of $\{x^k\}$ accumulates at a feasible point x^*. As this might not be the case, a mechanism for (local) infeasibility detection is needed, and usually included in practical implementations; see [4, 13].

Given some tolerances, Algorithm 1 can be equipped with relaxed conditions on decrease requirements at step 1.5 and optimality at step 1.2. At step 1.1 the inner tolerance ε^k can stay bounded away from zero, as long as $\varepsilon^k \leq \varepsilon^{\text{dual}}$ for large $k \in \mathbb{N}$. Similarly, the condition at step 1.5 can be relaxed by adding the (inclusive) possibility that $\|c(x^k) - s^k\| \leq \varepsilon^{\text{prim}}$. Finally, at step 1.6 a nonmonotone update is allowed, namely the penalty parameter can be increased, as long as some watchdog procedures are in place to avoid cycling [10].

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 $\mathcal{L}_\mu(\cdot, y)$ defined in (3.3). As mentioned in Section 3.1, steps 1.2 and 1.3 could be combined and performed by jointly minimizing $\mathcal{L}^{\mathrm{S}}_\mu(\cdot, \cdot, y)$, that is, with respect to both primal and slack variables. Owing to the structure of $\mathcal{L}^{\mathrm{S}}_\mu(\cdot, \cdot, y)$, defined in (3.1), any suitable method for composite optimization could be adopted for this task.

We take a different approach, in order to solve subproblems without slack variables, and aim at solving $\min \mathcal{L}_\mu(\cdot, y)$ using any algorithm for two-terms composite optimization, ruling out three-terms splitting methods. Moreover, to avoid ruining the problem structure, we split terms in $\mathcal{L}_\mu(\cdot, y)$ by leaving the nonsmooth cost function g alone, and not paired with $\operatorname{dist}^2_D(\cdot)$. Consequently, the proximal mapping oracle of g can be readily exploited. However, this pairing yields a second term $f + \operatorname{dist}^2_D(c(\cdot) + \mu y)/\mu$ that is not continuously differentiable, since the projection onto D is possibly set-valued, thus hindering the direct application of proximal gradient algorithms as presented in [22, 33]. In Section 4.1 we show it is indeed possible to adopt any adaptive proximal gradient method to minimize $\mathcal{L}_\mu(\cdot, y)$, in a remarkably transparent way.

4.1 Proximal Gradient with an Oracle

This section is dedicated to showing that the subproblems at step 1.2 of Algorithm 1 can be solved by any suitable off-the-shelf method for two-terms composite optimization. Here, by suitable solvers we mean those that can cope with a smooth cost term having only locally Lipschitz continuous gradient, see [33] as well. The proposed algorithm heavily relies on PANOC$^+$ [22], retaining its lightweight iterations and convergence guarantees, although the overall approach is more general and of interest in its own right. Accordingly, one may replace PANOC$^+$ in the following discussion with any suitable proximal gradient method.

Let $\mu > 0$ and $y \in \mathbb{R}^m$ be fixed and denote $\psi\colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ the smooth part of $\mathcal{L}_\mu^{\mathrm{S}}(\cdot,\cdot,y)$ and $h\colon \mathbb{R}^m \to \overline{\mathbb{R}}$ the indicator onto the constraint set, namely according to (3.1)

$$\psi(x,s) := f(x) + \frac{1}{2\mu}\|c(x) + \mu y - s\|^2 - \frac{\mu}{2}\|y\|^2, \qquad h(s) := \delta_D(s). \tag{4.1}$$

Then, the joint minimization of $\mathcal{L}_\mu^{\mathrm{S}}(\cdot,\cdot,y)$ can be written as

$$\operatorname*{minimize}_{x,\,s} \quad \varphi(x,s), \qquad \text{where} \quad \varphi(x,s) := \psi(x,s) + g(x) + h(s). \tag{4.2}$$

Concurrently, one can consider the cost function as effectively dependent on x only. Patterning the transition from $\mathcal{L}_\mu^{\mathrm{S}}(\cdot,\cdot,y)$ to $\mathcal{L}_\mu(\cdot,y)$ in Section 3, we suppose that an oracle $\mathcal{O}$ is available that performs the explicit (exact, global) minimization of φ with respect to $s \in \mathbb{R}^m$, apparently alluding to (3.2). Let $\mathcal{O}\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ be the set-valued mapping defined by

$$\mathcal{O}(x) := \arg\min_s \varphi(x,s) = \arg\min_s \psi(x,s) + h(s) = \mathcal{S}_\mu(x,y). \tag{4.3}$$

Accordingly, and owing to the separable problem structure, the reduced cost functions $\varphi_{\mathrm{red}}\colon \mathbb{R}^n \to \overline{\mathbb{R}}$ and $\psi_{\mathrm{red}}\colon \mathbb{R}^n \to \mathbb{R}$ can be defined by

$$\varphi_{\mathrm{red}}(x) := \inf_s \varphi(x,s) \qquad \text{and} \qquad \psi_{\mathrm{red}}(x) := \inf_s \psi(x,s) + h(s). \tag{4.4}$$

Notice that $\varphi_{\mathrm{red}} = \psi_{\mathrm{red}} + g = \mathcal{L}_\mu(\cdot,y)$, in light of (3.3), and, for all $x \in \operatorname{dom} g$, it is $\mathcal{O}(x) \subseteq \operatorname{dom} h$ and thus $\varphi_{\mathrm{red}}(x) < \infty$.

We proceed by detailing an algorithm for solving (4.2), namely minimizing the augmented Lagrangian $\mathcal{L}_\mu^{\mathrm{S}}(\cdot,\cdot,y)$ including the slack variables. In order to exploit the underlying problem structure, an oracle $\mathcal{O}$ is introduced that performs the explicit minimization over the slack variables, on the vein of “magic steps” [11, 17]. Then, we prove well-definedness of the algorithm and investigate its convergence properties. Finally, we show that the iterates generated by such an algorithm coincide with those of the original PANOC$^+$ for minimizing $\mathcal{L}_\mu(\cdot,y)$, that is, for solving the augmented Lagrangian subproblem without slack variables. In particular, it is demonstrated that one can invoke PANOC$^+$ and, at all $x \in \mathbb{R}^n$, employ the term $\nabla_x \mathcal{L}_\mu^{\mathrm{S}}(x,s,y)$ with any $s \in \mathcal{S}_\mu(x,y)$ as a proxy of the gradient oracle.

It should be stressed that we use PANOC$^+$ with an oracle only as a formal tool, a *virtual* algorithm for solving (4.2), as it is in fact transparent to the user that could as well be using PANOC$^+$ for minimizing $\mathcal{L}_\mu(\cdot,y)$.

Algorithm 2 PANOC$^+$ with an oracle

```
REQUIRE  $x^0 \in \mathbb{R}^n$, $\gamma_0 \in (0, \gamma_g)$, $\mathfrak{D} \geq 0$, $\alpha, \beta \in (0,1)$
INITIALIZE  $k \leftarrow 0$, and start from step 2.4
2.1: $\gamma_k \leftarrow \gamma_{k-1}$
2.2: Select an update direction $d^k \in \mathbb{R}^n$ with $\|d^k\| \leq \mathfrak{D}\|\bar{x}^{k-1} - x^{k-1}\|$ and set $\tau_k = 1$
2.3: Set $x^k = (1 - \tau_k)\bar{x}^{k-1} + \tau_k(x^{k-1} + d^k)$
2.4: Select $s^k \in \mathcal{O}(x^k)$ and use it to evaluate $\psi_{\text{red}}(x^k)$        // oracle
2.5: Compute $\bar{x}^k \in \mathrm{T}_{\gamma_k}(x^k, s^k)$ and set $\Phi_k$ as in (4.5)
2.6: Select $\bar{s}^k \in \mathcal{O}(\bar{x}^k)$ and use it to evaluate $\psi_{\text{red}}(\bar{x}^k)$        // oracle
2.7: IF $\psi_{\text{red}}(\bar{x}^k) > \psi_{\text{red}}(x^k) + \left\langle \nabla_x \psi(x^k, s^k), \bar{x}^k - x^k \right\rangle + \frac{\alpha}{2\gamma_k}\|\bar{x}^k - x^k\|^2$ THEN
         $\gamma_k \leftarrow \gamma_k/2$, and go back to step 2.2 if $k > 0$, or step 2.5 if $k = 0$
2.8: IF $k > 0$ AND $\Phi_k > \Phi_{k-1} - \beta\frac{1-\alpha}{2\gamma_{k-1}}\|\bar{x}^{k-1} - x^{k-1}\|^2$ THEN
         $\tau_k \leftarrow \tau_k/2$ and go back to step 2.3
2.9: $k \leftarrow k + 1$ and start the next iteration at step 2.1
```

Before discussing Algorithm 2, we define some tools and characterize the problem. Given any $\gamma > 0$, let $\mathrm{T}_\gamma \colon \mathbb{R}^n \times \mathbb{R}^m \rightrightarrows \mathbb{R}^n$ denote the proximal gradient mapping defined by

$$\mathrm{T}_\gamma(x, s) := \operatorname{prox}_{\gamma g}(x - \gamma \nabla_x \psi(x, s)).$$

Proposition 4.1. *Consider Eqs. (4.1) to (4.4) and let Assumption I hold. Then,*

(i) $\psi \colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ *is continuously differentiable with locally Lipschitz continuous derivatives;*

(ii) $\varphi \colon \mathbb{R}^n \times \mathbb{R}^m \to \overline{\mathbb{R}}$ *is proper, lsc, and* $\varphi(x, s)$ *is level-bounded in* s *locally uniformly in* x*;*

(iii) $\psi_{red} \colon \mathbb{R}^n \to \mathbb{R}$ *is continuous;*

(iv) $\mathcal{O} \colon \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ *is locally bounded, nonempty- and compact-valued.*

Proof. Assertions 4.1*(i)*-4.1*(ii)* directly follow from the definitions in (4.1), (4.2) and Assumptions I*(i)* and I*(ii)*. Then, owing to 4.1*(ii)* and continuity of ψ by 4.1*(i)*, [45, Thm 1.17(c)] yields assertion 4.1*(iii)*. Finally, by the properties of Π_D [45, Thm 1.25], assertion 4.1*(iv)* is a direct consequence of Assumption I*(i)*. □

Existence of a solution for (4.2) is guaranteed by Assumption I together with the following.

Assumption II. *In* (4.2), φ *is bounded from below, i.e.,* $\inf \varphi > -\infty$.

The aforementioned proximal gradient method equipped with an oracle is stated in Algorithm 2, that coincides with PANOC$^+$ [22] when variable s is neglected. Some comments are in order.

First, the method requires a starting point x^0 for x only, not for s. Then, by invoking the oracle at steps 2.4 and 2.6, it is $s^k \in \mathcal{O}(x^k)$ and $\bar{s}^k \in \mathcal{O}(\bar{x}^k)$ for all $k \in \mathbb{N}$, so that the cost function ψ is effectively evaluated only on the set described by the explicit minimization over s.

The proximal gradient update at step 2.5 and the backtracking condition at step 2.7 for adapting the stepsize γ_k are concerned with variable x only. Analogously, the merit Φ_k evaluated at step 2.5 is defined by

$$\Phi_k = \psi_{\text{red}}(x^k) + \left\langle \nabla_x \psi(x^k, s^k), \bar{x}^k - x^k \right\rangle + \frac{1}{2\gamma_k} \|\bar{x}^k - x^k\|^2 + g(\bar{x}^k). \tag{4.5}$$

Invoked in the linesearch backtracking condition at step 2.8, Φ_k plays the role of a (partial) forward-backward envelope (FBE) of φ at (x^k, s^k) with stepsize parameter γ_k [22, 53]. We say it is partial because it relies on a quadratic model of ψ with respect to x only.

Overall, these observations show that Algorithm 2 is equivalent to PANOC$^+$ applied to the minimization problem $\min \psi_{\text{red}} + g$. The arbitrariness of the element chosen by the oracle $\mathcal{O}$ allows to keep variable s implicit, making the virtual algorithm transparent to the user and thus a formal tool only.

4.2 Convergence Analysis

In this section we investigate the convergence properties of Algorithm 2. The results and proofs stated in the following closely pattern those presented in [22], with minor adjustments to account for the differences with respect to the original PANOC$^+$. To avoid trivialities, it is assumed that $\bar{x}^k \neq x^k$ always holds. This is consistent with stopping criteria based on the proximal gradient residual, as discussed in Section 4.3.

Furthermore, we will consider the case $x^0 \in \operatorname{dom} g$, possibly discarding the first iteration but without affecting the convergence analysis. In fact, even if $x^0 \notin \operatorname{dom} g$, it is $x^1 \in \operatorname{dom} g$ by step 2.5, $s^1 \in \operatorname{dom} h$ by step 2.6, accepted at the first attempt by step 2.7, by the local boundedness of the proximal operator.

We are going to show first that Algorithm 2 is well defined, namely that each and every iteration terminates in finite time. This is an essential theoretical and practical condition that an implementable algorithm must satisfy. A detailed proof has been included in the Additional Material (p. 34).

Lemma 4.2 (Well definedness)**.** *Let Assumptions I and II hold and consider the iterates generated by Algorithm 2. The following hold:*

(i) At every iteration, the number of backtrackings at steps 2.7 and 2.8 is finite.

(ii) At the end of the k-th iteration ($k \geq 1$), one has

$$\varphi(\bar{x}^k, \bar{s}^k) + \delta_k \leq \Phi_k \leq \Phi_{k-1} - \beta\delta_{k-1} \quad \text{where} \quad \delta_k := \frac{1-\alpha}{2\gamma_k}\|\bar{x}^k - x^k\|^2.$$

(iii) Every iterate $(\bar{x}^k, \bar{s}^k)$ remains within $\operatorname{lev}_{\leq \Phi_0} \varphi$, where $\Phi_0 < \infty$.

Proof. Patterning the proof of [22, Lem. 4.2], the assertion is demonstrated by induction on k and exploiting two more facts. First, the oracle $\mathcal{O}$ defined in (4.3) gives the upper bound $\psi_{\mathrm{red}}(\bar{x}^k) \leq \psi(\bar{x}^k, s^k) + h(s^k)$, adopted to show that the condition at step 2.7 is violated after finitely many attempts. Then, local boundedness of $\mathcal{O}$ by Proposition 4.1*(iv)* is sufficient for the linesearch backtracking at step 2.8 to successfully terminate. □

We next consider an asymptotic analysis of the algorithm. The proof of the following assertions adheres to that of [22, Thm 4.3] and uses Proposition 4.1*(iv)*. A full proof is provided in the Additional Material (p. 36).

Theorem 4.3 (Asymptotic analysis)**.** *Let Assumptions I and II hold and consider a sequence of iterates generated by Algorithm 2. The following hold:*

(i) $\{\Phi_k\}$ *converges to a finite value* $\varphi_\star \geq \inf \varphi$ *from above.*

(ii) $\sum_{k\in\mathbb{N}} \frac{1}{\gamma_k}\|\bar{x}^k - x^k\|^2 < \infty$.

(iii) $\lim_{k\to\infty}\|x^k - \bar{x}^k\| = \lim_{k\to\infty}\|x^k - x^{k-1}\| = \lim_{k\to\infty}\|\bar{x}^k - \bar{x}^{k-1}\| = 0$, *and in particular the set of accumulation points of* $\{x^k\}$ *is closed and connected, and coincides with that of* $\{\bar{x}^k\}$.

(iv) $\sum_{k\in\mathbb{N}} \gamma_k = \infty$.

(v) $\liminf_{k\to\infty} \frac{1}{\gamma_k}\|\bar{x}^k - x^k\| = 0$.

(vi) Consider the following assertions:

(1) φ *is level bounded;*

(2) $\{\bar{x}^k\}$ *and* $\{\bar{s}^k\}$ *are bounded;*

(3) $\{x^k\}$ *and* $\{s^k\}$ *are bounded;*

(4) $\{\gamma_k\}$ *is asymptotically constant, i.e., there exist* $\kappa \in \mathbb{N}$ *and* $\gamma_\kappa > 0$ *such that* $\gamma_k = \gamma_\kappa$ *for every* $k \geq \kappa$.

One has (1) ⇒ (2) ⇔ (3) ⇒ (4).

If the iterates $\{x^k\}$ remain bounded, as is the case when φ is level bounded, the proximal stepsize γ^k is asymptotically constant, owing to Theorem 4.3*(vi)*. The following global convergence result is inspired by those in [22, 50, 53].

Theorem 4.4 (Subsequential convergence)**.** *Let Assumptions I and II hold and consider a sequence $\{(x^k, s^k)\}$ of iterates generated by Algorithm 2. Suppose that $\{x^k\}$ remains bounded and that the set ω of accumulation points of $\{x^k\}$ is nonempty. Let $x^\star \in \omega$ and $s^\star \in \mathcal{O}(x^\star)$. Then $(x^\star, s^\star)$ is a stationary point of φ.*

Proof. Up to possibly discarding early iterates, in light of the boundedness of the sequences, γ_k is asymptotically constant by Theorem 4.3*(vi)*, and we may assume that $\gamma_k \equiv \gamma > 0$ holds for all k. Let $x^\star \in \omega$ and $s^\star \in \mathcal{O}(x^\star)$ be arbitrary but fixed, and $\{\bar{x}^k\}_K$ a subsequence such that $\bar{x}^k \to_K x^\star$, so that $x^k \to_K x^\star$ too as it follows from Theorem 4.3*(iii)*. Letting $\varphi_\star$ be as in Theorem 4.3*(i)* and invoking Lemma 4.2*(ii)*, lower semicontinuity of φ yields $\varphi(x^\star, s^\star) \leq \varphi_\star$. Conversely, we have

$$\begin{aligned}
\varphi_\star &= \lim_{k\in K} \Phi_k \\
&= \lim_{k\in K} \psi_{\mathrm{red}}(x^k) + \left\langle \nabla_x\psi(x^k, s^k), \bar{x}^k - x^k \right\rangle + \frac{1}{2\gamma}\|\bar{x}^k - x^k\| + g(\bar{x}^k) \\
&\leq \limsup_{k\in K} \psi_{\mathrm{red}}(x^k) + \left\langle \nabla_x\psi(x^k, s^k), x^\star - x^k \right\rangle + \frac{1}{2\gamma}\|x^\star - x^k\| + g(x^\star) \\
&= \psi_{\mathrm{red}}(x^\star) + g(x^\star) = \varphi(x^\star, s^\star) \leq \varphi_\star,
\end{aligned}$$

owing to continuity of ψ_{red} and $\nabla\psi$ by Proposition 4.1, and the fact that $\|x^k - \bar{x}^k\|$ vanishes. By step 2.5, the minimizing property of $\bar{x}^k$ and Fermat's rule imply that

$$0 \in \nabla_x\psi(x^k, s^k) + \partial g(\bar{x}^k) + \frac{1}{\gamma}(\bar{x}^k - x^k)$$

for all $k \in \mathbb{N}$. By lower semicontinuity of g and since $\bar{x}^k \to_K x^\star$, necessarily $g(\bar{x}^k) \to_K g(x^\star)$ and taking the limit for $k \to_K \infty$ leads to the inclusion $0 \in \partial_x\varphi(x^\star, s^\star)$, since $s^k \in \mathcal{O}(x^k)$ for all $k \in \mathbb{N}$. Furthermore, it is necessarily $0 \in \partial_s\varphi(x^\star, s^\star)$, whence the claimed stationarity of $(x^\star, s^\star)$. The arbitrarity of $(x^\star, s^\star)$ proves the assertion. □

4.3 Termination Criteria

Algorithm 2 runs indefinitely, generating an infinite sequence of iterates. Suitable termination criteria should be inserted for stopping and returning an iterate that, in some sense, satisfactorily solves (4.2). Since $\bar{s}^k \in \mathcal{O}(\bar{x}^k)$ for all $k \in \mathbb{N}$, it suffices to check the (approximate) stationarity of $\bar{x}^k$. The ε-stationarity of a solution (x^*, s^*) requires some $\eta \in \mathbb{R}^n$, $\|\eta\| \leq \varepsilon$, such that $\eta \in \partial_x\varphi(x^*, s^*)$; cf. (3.4). Equivalently, it must be that $\operatorname{dist}_{\{0\}}(\partial_x\varphi(x^*, s^*)) \leq \varepsilon$. Following [22, §4.2], we retort to the necessary optimality condition in the minimization problem defining the proximal mapping at step 2.5, that reads

$$0 \in \partial g(\bar{x}^k) + \nabla_x\psi(x^k, s^k) + \frac{1}{\gamma_k}(\bar{x}^k - x^k)$$

and, combined with $\partial_x \varphi(\bar{x}^k, \bar{s}^k) = \nabla_x \psi(\bar{x}^k, \bar{s}^k) + \partial g(\bar{x}^k)$, yields the upper bound

$$\operatorname{dist}_{\{0\}}(\partial_x \varphi(\bar{x}^k, \bar{s}^k)) \leq \left\| \frac{1}{\gamma_k}(x^k - \bar{x}^k) - \nabla_x \psi(x^k, s^k) + \nabla_x \psi(\bar{x}^k, \bar{s}^k) \right\|.$$

Therefore, owing to the convergence guarantees presented in Section 4.2, it is reasonable to equip Algorithm 2 with the termination condition

$$\left\| \frac{1}{\gamma_k}(x^k - \bar{x}^k) - \nabla_x \psi(x^k, s^k) + \nabla_x \psi(\bar{x}^k, \bar{s}^k) \right\| \leq \varepsilon, \tag{4.6}$$

given some tolerance $\varepsilon > 0$.

5 Numerical Examples

This section presents a numerical implementation of Algorithm 1 and discusses its behaviour on some illustrative examples, showcasing the flexibility offered by the constrained structured 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 Section 5.2 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, Section 5.3 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 5.4 deals with a test collection of portfolio optimization problems from [26] which are equipped with a nonconvex sparsity-promoting term in the objective function. Finally, in Section 5.5 we address a class of matrix recovery problems discussed e.g. in [47] where the rank of the unknown matrix has to be minimized.

5.1 Implementation

We have implemented the proposed approach in the “Augmented Lagrangian Proximal Solver” (ALPS), an open-source software package in the Julia language [9]. ALPS can solve constrained structured 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. ALPS can be used to solve a wide spectrum of optimization problems, requiring only first-order primitives, i.e., gradient, proximal mapping and projections. The augmented Lagrangian subproblems at step 1.2 of Algorithm 1 are solved by default using the implementation of PANOC$^+$ [22] offered by ProximalAlgorithms.jl [49]. 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 Section 1. Further, the initialization requires a primal-dual starting point $(x^0, y^0) \in \mathbb{R}^n \times \mathbb{R}^m$. 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. User override of this oracle allows for tailored choices of Y, possibly exploiting the structure of D [48].

ALPS initializes Algorithm 1 by replacing x^0 with an arbitrary element of $\operatorname{prox}_{\gamma g}(x^0) \subset \operatorname{dom} q$, 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

$$\varepsilon_{k+1} = \max\{\kappa_\varepsilon \varepsilon_k, \varepsilon^{\text{dual}}\}, \tag{5.1}$$

starting from $\varepsilon_0 := \sqrt{\varepsilon^{\text{dual}}}$ and given some $\kappa_\varepsilon \in (0,1)$ [10]. The initial penalty parameter μ_0 is automatically chosen by default, similarly to [11, Eq. 12.1]. Given $x^0 \in \operatorname{dom} q$, we evaluate the constraints $c^0 := c(x^0)$, select an arbitrary element $p^0 \in \Pi_D(c^0)$ and compute the vector $d^0 := c^0 - p^0$. Then, the vector $\mu_0 \in \mathbb{R}^m$ of penalty parameters is selected componentwise as follows:

$$(\mu_0)_i := \max\left\{10^{-8}, \min\left\{\frac{1}{10}\frac{\max\{1, (d_i^0)^2/2\}}{\max\{1, q(x^0)\}}, 10^8\right\}\right\}, \tag{5.2}$$

effectively scaling the contribution of each constraint [16, 11]. Then, according to the overall feasibility-complementarity of the iterate, the penalty parameters are updated in unison at step 1.8, since using a different penalty parameter for each constraint is theoretically worse than using a common parameter [2, §3.4]; we set $\mu_{k+1} := \kappa\mu_k$, for some fixed $\kappa \in (0,1)$.

The default parameters in ALPS are $\theta = 0.8$, $\kappa = 0.5$ 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.

5.2 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

$$\operatorname*{minimize}_{x} \; 10(x_2 + 1 - (x_1+1)^2)^2 + |x_1| \quad \text{subject to} \;\; x_2 \leq -x_1 \;\vee\; x_2 \geq x_1 \tag{5.3}$$

and admits a unique (global) minimizer $x^\star = (0,0)$. The feasible set is nonconvex and connected; see Figure 2. We cast (5.3) into the form of (P) by defining the data functions 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 $D := D_{\text{EO}}$, where the (nonconvex) set

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

describes the either-or constraint.

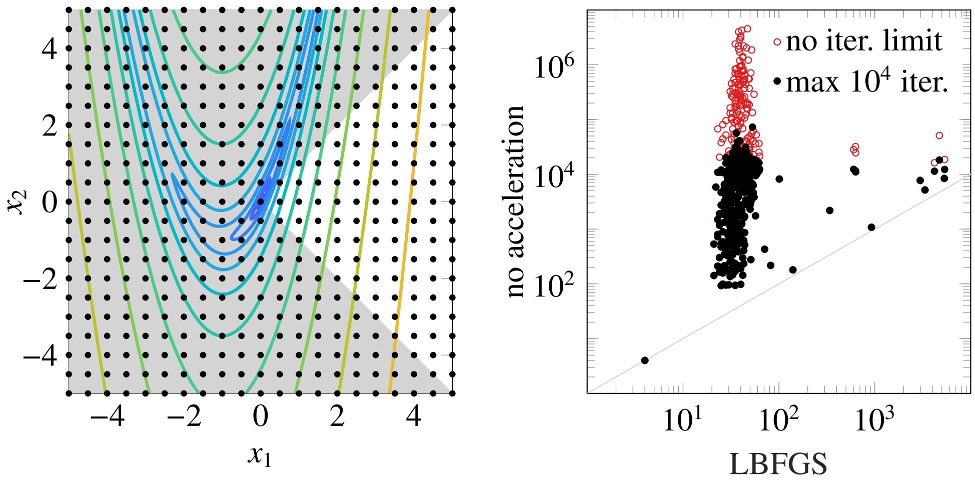


Figure 2: *Setup and results for the illustrative problem* (5.3)*. Left: Feasible region (gray background), objective contour lines and grid of starting points. The global minimizer* $x^\star = (0, 0)$ *is found in all cases. Right: Comparison of inner iterations needed without acceleration against LBFGS acceleration; each mark corresponds to a starting point and the gray line has unitary slope.*

We consider a uniform grid of $21^2 = 441$ starting points x^0 in $[-5, 5]^2$ and let the initial dual estimate be $y^0 = 0$. Also, we compare the performance of ALPS by solving the subproblems using PANOC$^+$ without or with acceleration. In the latter case, we use the default acceleration in ProximalAlgorithms.jl, namely LBFGS directions with memory 5. Furthermore, we also observe the effect of setting a limit to the number of iterations the inner solver is allowed to take.

ALPS 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 (5.3), some contour lines of its objective function and the grid of starting points x^0. Over all problems, ALPS with no acceleration takes at most 4 548 892 (cumulative) inner iterations to find a solution (median 7 864), whereas with LBFGS directions only 5 345 inner iterations are needed at most (median 38). By limiting the subsolver to 10^4 iterations, ALPS still manages to solve all problem instances; in this configuration it requires at most 72 762 inner iterations (same median). A more detailed comparison in Figure 2 demonstrates that not only the accelerated method usually requires far less iterations, but also that its behaviour 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 [50, 52].

5.3 Sparse Switching Time Optimization

Constrained structured 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 [46, §6.4], it is typically stated as

$$
\begin{aligned}
&\underset{x,u}{\text{minimize}} && \int_0^T \|x(t) - 1\|^2 \mathrm{d}t && && (5.4)\\
&\text{subject to} && \dot{x}_1(t) = x_1(t)[-c_1 u(t) - x_2(t) + 1] && \text{for a.e. } t \in [0, T],\\
& && \dot{x}_2(t) = x_2(t)[-c_2 u(t) + x_1(t) - 1] && \text{for a.e. } t \in [0, T],\\
& && x(0) = x_0,\\
& && u(t) \in \{0, 1\} && \text{for } t \in [0, T],
\end{aligned}
$$

where final time $T = 12$, initial state $x_0 := (0.5, 0.7)$ and parameters $c = (0.4, 0.2)$ 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 [46, §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 [46, §6.5], we introduce switch costs to penalize solutions that show frequent switching of the binary control trajectory, yielding more practical results. Following [19], [20, Chapter 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

$$\underset{\tau}{\text{minimize}} \quad f(\tau) + \delta_{\mathbb{R}_+}(\tau) + \sigma\|\tau\|_0 \qquad \text{subject to} \quad 1^\top \tau \in D. \tag{5.5}$$

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 $\delta_{\mathbb{R}_+}$ 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 [19, §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 $T := 1^\top\tau$. Hence, the constraint set $D \subseteq \mathbb{R}_+$ 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 [51]. We solve (5.5) 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 (5.4). 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 $D := [0, 12]$ and ALPS returns solutions whose final time reaches values around $T \approx 8.5$. Then, we consider a second case with the disconnected constraint set $D := [5, 7] \cup [10, 12]$, so to impact on the solution.

ALPS is able to find reasonable solutions that satisfy the constraints, despite the nonconvexity of the switching time approach [46, Appendix B.4], the discrete nature of the

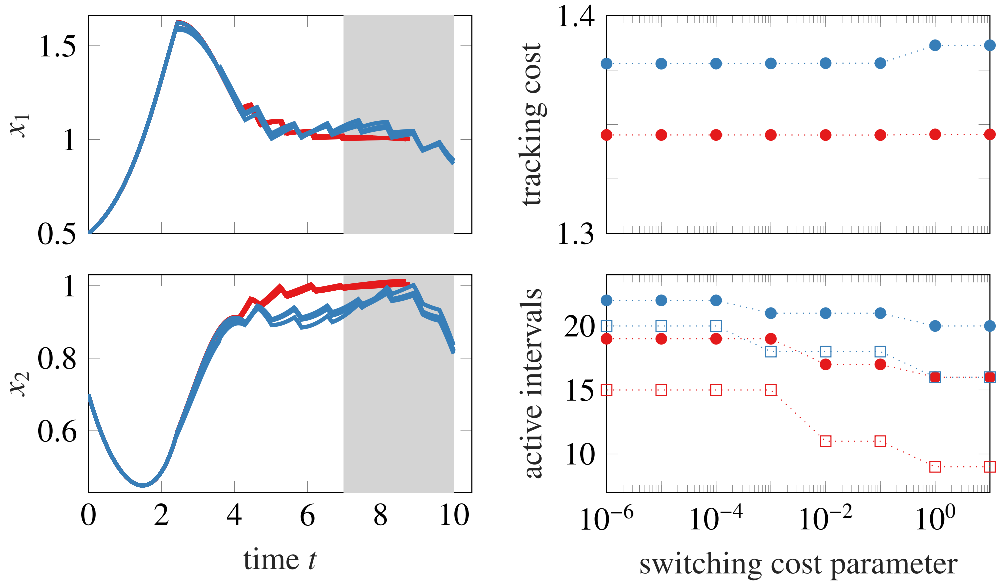


Figure 3: *Results for the illustrative problem (5.5) using switching time optimization with a sequence of* 24 *binary controls and several values for the switching cost parameter σ. Left: Prohibited region for the final time (gray background) and state trajectories with (blue) or without (red) constraint. 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.*

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 ALPS are poor in terms of objective value, as we are about to show.

The state trajectories are depicted in Figure 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 $(\alpha+\beta, 0, 0, \ldots)$ and $(0, 0, \alpha+\beta, \ldots)$ should be preferred over $(\alpha, 0, \beta, \ldots)$, for they yield the same control trajectory whilst having fewer nonzero elements. The solutions returned by ALPS are compared against equivalent although sparser ones in Figure 3. Clearly, and not surprisingly, the solutions obtained are far from being globally optimal.

5.4 Sparse Portfolio Optimization

Let us consider portfolio optimization problems in the form

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & \frac{1}{2} x^\top Q x + \alpha \|x\|_0 \\ \text{subject to} \quad & \mu^\top x \geq \varrho, \quad 1_n^\top x = 1, \quad 0 \leq x \leq u. \end{aligned} \tag{5.6}$$

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. [32], 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 (5.6), which can be obtained via the theory provided in [25], we compute a solution using CPLEX [18], for comparison. Based on our experiences from Section 5.3, we also solve (5.6) 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 (5.6) 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 (5.6) by the nonnegativity and equality constraints.

The data Q, μ, ϱ and u is taken from the test problem collection [26], which has been created randomly and is available from the webpage `https://commalab.di.unipi.it/datasets/MV/`. Here, we used all 30 test instances of dimension $n := 200$ and the two different values $\alpha \in \{10, 100\}$ for each problem.

Let us mention that ALPS solved all problem instances. Below, we comment on some median values for our experiments with parameters $\alpha = 10/100$: a direct use of ℓ_0 minimization resulted in 15/10 outer and 138/257 inner iterations, while warm-starting with the continuous ℓ_p^p function required 13/9 outer and 240/781 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, whereas the nonconvex ℓ_p^p penalty already leads to very sparse solutions.

5.5 Matrix Completion with Minimum Rank

For some $\ell \in \mathbb{N}$, $\ell \geq 2$, let us consider $N \in \mathbb{N}$ points $x_1, \dots, x_N \in \mathbb{R}^\ell$ and define a block matrix $X \in \mathbb{R}^{N \times \ell}$ by means of $X := [x_1, x_2, \dots, x_N]^\top$. Let $D \in \mathbb{R}^{N \times N}$ denote the Euclidean distance matrix associated with these points, given by $D_{ij} := \|x_i - x_j\|^2 = (x_i - x_j)^\top (x_i - x_j)$ for all $i, j \in \mathcal{I} := \{1, \dots, N\}$. We aim at recovering X based on a partial knowledge of D. In

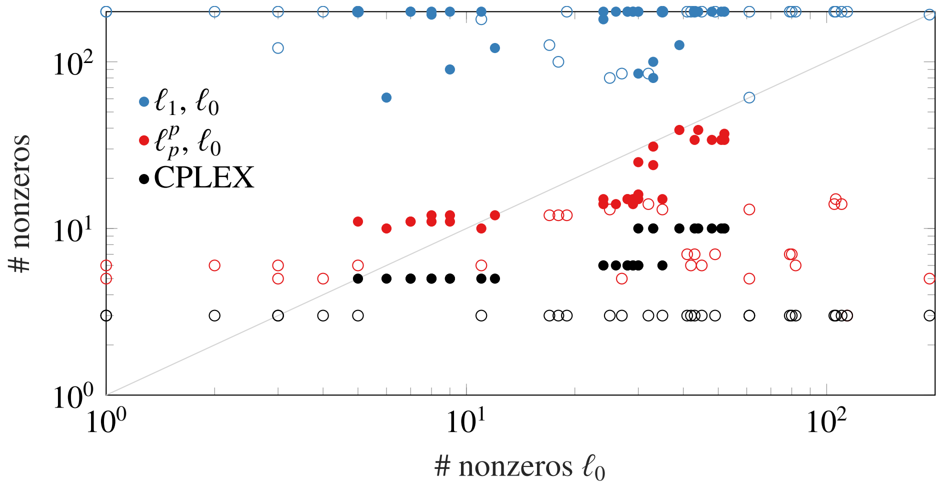


Figure 4: *Results for the portfolio problem* (5.6)*: Comparison of the solutions found with ℓ_0, CPLEX and ℓ_0 warm-started with ℓ_1 or ℓ_p^p. We depict the number of nonzero entries of the solutions found, for $\alpha = 10$ (dot) and $\alpha = 100$ (circle). The gray line has unitary slope.*

particular, we assume that $\Omega \subset \mathcal{I}^2$ is a set of pairs such that only the entries D_{ij}, $(i, j) \in \Omega$, of D are known.

Following [47], 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., $B_{ii} + B_{jj} - B_{ij} - B_{ji} = D_{ij}$ 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

$$\begin{aligned} \operatorname*{minimize}_{B} \quad & g(B) && \\ \text{subject to} \quad & B_{ii} + B_{jj} - B_{ij} - B_{ji} = D_{ij} && \forall (i, j) \in \Omega, \\ & B_{ij} = B_{ji} && \forall i, j \in \mathcal{I}, j < i \end{aligned} \tag{5.7}$$

where the function $g\colon \mathbb{R}^{N\times N} \to \mathbb{R}$ encodes a matrix regularization term. In the following, we consider $g := \operatorname{rank} := \|\sigma(\cdot)\|_0$, 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.

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 (5.7). We reformulate the model in the form of (P) by setting $f := 0$, $D := \{0_m\}$ 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 $N \in \{10, 15, 20\}$, $\ell = 5$, $m_o = \lfloor (n - m_s)/3 \rfloor$, $p = 0.5$ and consider 20 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 D. Finally, we sample observations by selecting m_o different entries of D with uniform probability.

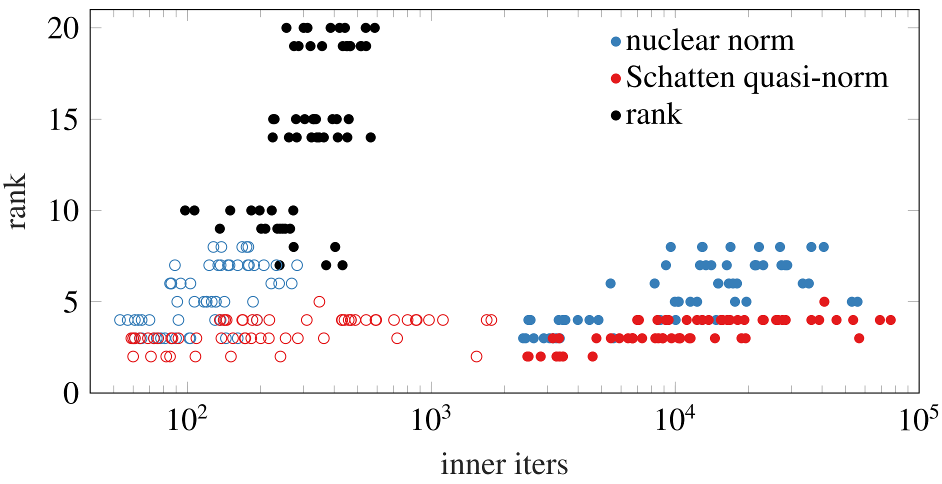


Figure 5: *Results for the matrix recovery problem* (5.7)*: Comparison of (accumulated) inner iteration numbers and rank of the solutions found with different formulations, including warm-started rank minimization (circle).*

We run our solver ALPS 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 based on $B^0_{ij} \sim \mathcal{N}(0,1)$, $(i,j) \in \mathcal{I}^2$, whereas the dual initial guess is fixed to $y^0 := 0_m$. We invoke ALPS directly for solving (5.7) 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 Figure 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 7 and 20. This behavior is not surprising since (5.7) possesses plenty of non-global minimizers in case where minimization of the discontinuous rank is considered, and ALPS 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 9 cases after minimization of the nuclear norm; in all other cases, no deterioration has been observed. In summary, ALPS manages to find feasible solutions of (5.7) 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 ALPS or tailored mixed-integer numerical methods.

6 Conclusions

We presented the class of constrained structured optimization problems and proposed a general-purpose solver based on augmented Lagrangian and proximal methods. The outer augmented Lagrangian loop generates a sequence of subproblems, each one being a dual proximal regularization of the original, that can be solved 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 ALPS solver, disentangled from modeling tools and subproblem solvers. Thanks to its low memory footprint and simple, yet fast and robust iterations, ALPS 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 structured programs, the benefits of implicit formulations and the impact of accelerated methods for solving the inner problems.

Acknowledgements

Alberto De Marchi wishes to thank Prof. Andreas Themelis (Kyushu University), for sharing his insight and rigour. The support and guidance of Prof. Matthias Gerdts (Universität der Bundeswehr München) is also gratefully acknowledged.
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.

References

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

[2] R. Andreani, G. Haeser, L. M. Mito, A. Ramos, and L. D. Secchin. On the best achievable quality of limit points of augmented Lagrangian schemes. *Numerical Algorithms*, 2021.

[3] H. Antil, D. P. Kouri, and D. Ridzal. ALESQP: An augmented Lagrangian equality-constrained SQP method for optimization with general constraints. URL http://www.optimization-online.org/DB_HTML/2021/01/8232.html, 2020.

[4] P. Armand and N. N. Tran. Rapid infeasibility detection in a mixed logarithmic barrier-augmented Lagrangian method for nonlinear optimization. *Optimization Methods and Software*, 34(5):991–1013, 2019.

[5] E. Balas. *Disjunctive Programming*. Springer, Cham, 2018.

[6] A. Beck and N. Hallak. Optimization problems involving group sparsity terms. *Mathematical Programming*, 2018.

[7] M. Benko and P. Mehlitz. On implicit variables in optimization theory. *Journal of Nonsmooth Analysis and Optimization*, 2:7215, 2021.

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

[9] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah. Julia: A fresh approach to numerical computing. *SIAM Review*, 59(1):65–98, 2017.

[10] E. G. Birgin and J. M. Martínez. Augmented Lagrangian method with nonmonotone penalty parameters for constrained optimization. *Computational Optimization and Applications*, 51(3):941–965, 2012.

[11] E. G. Birgin and J. M. Martínez. *Practical Augmented Lagrangian Methods for Constrained Optimization*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2014.

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

[13] J. V. Burke, F. E. Curtis, and H. Wang. A sequential quadratic optimization algorithm with rapid infeasibility detection. *SIAM Journal on Optimization*, 24(2):839–872, 2014.

[14] X. Chen, L. Guo, Z. Lu, and J. J. Ye. An augmented Lagrangian method for non-Lipschitz nonconvex programming. *SIAM Journal on Numerical Analysis*, 55:168–193, 2017.

[15] P. L. Combettes and J.-C. Pesquet. *Proximal splitting methods in signal processing*, pages 185–212. Springer, New York, 2011.

[16] A. R. Conn, N. I. M. Gould, and P. L. Toint. A globally convergent augmented Lagrangian algorithm for optimization with general constraints and simple bounds. *SIAM Journal on Numerical Analysis*, 28(2):545–572, 1991.

[17] A. R. Conn, N. I. M. Gould, and P. L. Toint. *Trust Region Methods*. Society for Industrial and Applied Mathematics, 2000.

[18] IBM ILOG Cplex. V12. 1: User's Manual for CPLEX. *International Business Machines Corporation*, 46(53):157, 2009.

[19] A. De Marchi. Constrained and sparse switching times optimization via augmented Lagrangian proximal methods. In *2020 American Control Conference (ACC)*, pages 3633–3638, 2020.

[20] A. De Marchi. *Augmented Lagrangian and Proximal Methods for Constrained Structured Optimization*. PhD thesis, Universität der Bundeswehr München, 2021.

[21] A. De Marchi. On a primal-dual Newton proximal method for convex quadratic programs. *Computational Optimization and Applications*, 2022.

[22] A. De Marchi and A. Themelis. Proximal gradient algorithms under local Lipschitz gradient continuity: A convergence and robustness analysis of PANOC. URL `https://arxiv.org/abs/2112.13000`, 2021.

[23] N. K. Dhingra, S. Z. Khong, and M. R. Jovanović. The proximal augmented Lagrangian method for nonsmooth composite optimization. *IEEE Transactions on Automatic Control*, 64(7):2861–2868, 2019.

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

[25] M. Feng, J. E. Mitchell, J.-S. Pang, X. Shen, and A. Wächter. Complementarity formulations of ℓ_0-norm optimization problems. *Pacific Journal of Optimization*, 14(2):273–305, 2018.

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

[27] P. E. Gill and D. P. Robinson. A primal-dual augmented Lagrangian. *Computational Optimization and Applications*, 51(1):1–25, 2012.

[28] G. N. Grapiglia and Y. Yuan. On the complexity of an augmented Lagrangian method for nonconvex optimization. *IMA Journal of Numerical Analysis*, 2020.

[29] L. Guo and Z. Deng. A new augmented Lagrangian method for MPCCs – Theoretical and numerical comparison with existing augmented Lagrangian methods. *Mathematics of Operations Research*, 2021.

[30] L. Guo and J. J. Ye. Necessary optimality conditions and exact penalization for non-Lipschitz nonlinear programs. *Mathematical Programming*, 168(1):571–598, 2018.

[31] M. R. Hestenes. Multiplier and gradient methods. *Journal of Optimization Theory and Applications*, 4(5):303–320, 1969.

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

[33] C. Kanzow and P. Mehlitz. Convergence properties of monotone and nonmonotone proximal gradient methods revisited. URL https://arxiv.org/abs/2112.01798, 2021.

[34] C. Kanzow, D. Steck, and D. Wachsmuth. An augmented Lagrangian method for optimization problems in Banach spaces. *SIAM Journal on Control and Optimization*, 56(1):272–291, 2018.

[35] A. Y. Kruger and P. Mehlitz. Optimality conditions, approximate stationarity, and applications – a story beyond Lipschitzness. *ESAIM: Control, Optimisation and Calculus of Variations*, 2022. accepted for publication.

[36] D. Ma, K. L. Judd, D. Orban, and M. A. Saunders. Stabilized optimization via an NCL algorithm. In M. Al-Baali, L. Grandinetti, and A. Purnama, editors, *Numerical Analysis and Optimization*, pages 173–191. Springer, 2018.

[37] P. Mehlitz. Asymptotic stationarity and regularity for nonsmooth optimization problems. *Journal of Nonsmooth Analysis and Optimization*, 1:6575, 2020.

[38] P. Mehlitz. A comparison of first-order methods for the numerical solution of or-constrained optimization problems. *Computational Optimization and Applications*, 76:233–275, 2020.

[39] B. S. Mordukhovich. *Variational Analysis and Generalized Differentiation, Part I: Basic Theory, Part II: Applications*. Springer, Berlin, 2006.

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

[41] N. Parikh and S. Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):127–239, 2014.

[42] A. Potschka and H. G. Bock. A sequential homotopy method for mathematical programming problems. *Mathematical Programming*, 187(1):459–486, 2021.

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

[44] R. T. Rockafellar. Augmented Lagrangians and applications of the proximal point algorithm in convex programming. *Mathematics of operations research*, 1(2):97–116, 1976.

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

[46] S. Sager. *Numerical methods for mixed-integer optimal control problems*. PhD thesis, University of Heidelberg, 2005. Interdisciplinary Center for Scientific Computing.

[47] X. Shen and J. E. Mitchell. A penalty method for rank minimization problems in symmetric matrices. *Computational Optimization and Applications*, 71(2):353–380, 2018.

[48] P. Sopasakis, E. Fresk, and P. Patrinos. OpEn: Code generation for embedded nonconvex optimization. *IFAC-PapersOnLine*, 53(2):6548–6554, 2020. 21st IFAC World Congress.

[49] L. Stella and contributors. ProximalAlgorithms.jl: Proximal algorithms for nonsmooth optimization in Julia. URL https://github.com/JuliaFirstOrder/ProximalAlgorithms.jl.

[50] L. Stella, A. Themelis, P. Sopasakis, and P. Patrinos. A simple and efficient algorithm for nonlinear model predictive control. In *56th IEEE Conference on Decision and Control (CDC)*, pages 1939–1944, 2017.

[51] B. Stellato, S. Ober-Blöbaum, and P. J. Goulart. Second-order switching time optimization for switched dynamical systems. *IEEE Transaction on Automatic Control*, 62(10):5407–5414, 2017.

[52] A. Themelis. *Proximal Algorithms for Structured Nonconvex Optimization*. PhD thesis, KU Leuven, Arenberg Doctoral School, 2018. Faculty of Engineering Science.

[53] A. Themelis, L. Stella, and P. Patrinos. Forward-backward envelope for the sum of two nonconvex functions: Further properties and nonmonotone linesearch algorithms. *SIAM Journal on Optimization*, 28(3):2274–2303, 2018.

Additional Material

Proof of Proposition 3.2

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 have $x^k \in \operatorname{dom} q$ for all $k \in \mathbb{N}$, by step 1.2 of Algorithm 1, so that lower semicontinuity of q yields $x^* \in \operatorname{dom} q$. It remains to consider the set-membership constraint $c(x) \in D$.

(i) Since $\{\mu_k\}$ is bounded away from zero, the conditions at steps 1.5 and 1.8 of Algorithm 1 imply that $\|c(x^k) - s^k\| \to 0$ for $k \to \infty$. By the upper bound $\|c(x^k) - s^k\| \geq \operatorname{dist}_D(c(x^k))$ for all $k \in \mathbb{N}$, taking the limit $k \to_K \infty$ and continuity yield $\operatorname{dist}_D(c(x^*)) = 0$, hence $c(x^*) \in D$.

(ii) By assumption, we have

$$\mathcal{L}_{\mu_k}(x^k, \hat{y}^k) = q(x^k) + \frac{1}{2\mu_k} \operatorname{dist}_D^2\left(c(x^k) + \mu_k \hat{y}^k\right) - \frac{\mu_k}{2}\|\hat{y}^k\|^2 \leq B$$

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

$$\operatorname{dist}_D^2\left(c(x^k) + \mu_k \hat{y}^k\right) \leq 2\mu_k[B - q(x^k)] + \|\mu_k \hat{y}^k\|^2$$

for all $k \in \mathbb{N}$. In view of Proposition 3.2*(i)*, it suffices to consider the case $\mu_k \to 0$. Taking the limit $k \to_K \infty$ and using the boundedness of $\{\hat{y}^k\} \subset Y$, we obtain

$$\operatorname{dist}_D^2(c(x^*)) = \lim_{k \to_K \infty} \operatorname{dist}_D^2\left(c(x^k) + \mu_k \hat{y}^k\right) = 0$$

by continuity, hence $c(x^*) \in D$.

Then, according to Definition 2.2, any accumulation point x^* is feasible. □

Proof of Lemma 4.2

Proof. Each iteration k defines or updates only variables indexed with a k sub- or superscript, while those defined in previous iterations remain untouched. Let us index by k, j the variables defined at the j-th attempt within iteration k.

• 4.2*(i)*: We proceed by induction on k. If $k = 0$, the oracle is invoked only once to compute s^0 and there is no backtracking on τ. From [22, Lem. 4.1] we conclude that all the trials $\bar{x}^{0,j}$ remain confined in a bounded set Ω_0, and therefore any stepsize $\gamma_{0,j} < \min\{1/L_{\psi,\Omega_0}, \gamma_g\}$ is accepted, since $\psi_{\rm red}(\bar{x}^k) \leq \psi(\bar{x}^k, s^k) + h(s^k)$.

Suppose now that $k > 0$ and observe that, by the definition of Φ_k and the failure of the condition at step 2.7, the inequality

$$\varphi(\bar{x}^{k-1}, \bar{s}^{k-1}) \leq \Phi_{k-1} - \frac{1-\alpha}{2\gamma_{k-1}}\|\bar{x}^{k-1} - x^{k-1}\|^2 \tag{6.1}$$

holds. Since $\|d^{k,j}\| \leq \mathfrak{D}\|\bar{x}^{k-1} - x^{k-1}\|$ and $\tau_{k,j} \in [0,1]$, any attempt $x^{k,j}$ defined at step 2.3 during the k-th iteration satisfies

$$\|x^{k,j} - \bar{x}^{k-1}\| = \tau_{k,j}\|x^{k-1} - \bar{x}^{k-1} + d^{k,j}\| \leq (1+\mathfrak{D})\|\bar{x}^{k-1} - x^{k-1}\|$$

and thus remains in a bounded set, be it Ω_k. To arrive to a contradiction, suppose that $\gamma_{k,j} \searrow 0$ as $j \to \infty$. Owing to the minimizing property of $\bar{x}^{k,j}$ at step 2.5, after choosing $x = \bar{x}^{k-1}$ and rearranging, we have that

$$\begin{aligned} g(\bar{x}^{k,j}) + \left\langle \nabla_x\psi(x^{k,j}, s^{k,j}), \bar{x}^{k,j} - \bar{x}^{k-1}\right\rangle + \frac{1}{2\gamma_{k,j}}\|\bar{x}^{k,j} - x^{k,j}\|^2 \\ \leq g(\bar{x}^{k-1}) + \frac{1}{2\gamma_{k,j}}\|\bar{x}^{k-1} - x^{k,j}\|^2. \end{aligned}$$

Since $(x^{k,j})_{j\in\mathbb{N}}$ is bounded, an application of [22, Lem. 4.1] reveals that $(\bar{x}^{k,j})_{k\in\mathbb{N}}$ too is bounded. Up to possibly enlarging the set, both sequences remain confined in the bounded set Ω_k, implying that the condition at step 2.7 should have terminated in finite time, whence the sought contradiction.

Hence, $\gamma_{k,j}$ is backtracked finitely many times within iteration k; up to discarding early attempts, we may denote $\gamma_{k,j} = \gamma_k$. By the minimizing property of $\bar{x}^{k,j}$, we have

$$\Phi_{k,j} \leq \psi_{\rm red}(x^{k,j}) + g(\bar{x}^{k-1}) + \left\langle \nabla_x\psi(x^{k,j}, s^{k,j}), \bar{x}^{k-1} - x^{k,j}\right\rangle + \frac{1}{2\gamma_k}\|\bar{x}^{k-1} - x^{k,j}\|^2.$$

As $\tau_{k,j} \searrow 0$, one has that $x^{k,j} \to \bar{x}^{k-1}$, by step 2.3. Proposition 4.1 guarantees continuity of $\psi_{\rm red}$ and $\nabla\psi$ and local boundedness of O. Then, the right-hand side of the inequality converges to $\psi_{\rm red}(\bar{x}^{k-1}) + g(\bar{x}^{k-1}) = \varphi_{\rm red}(\bar{x}^{k-1}) = \varphi(\bar{x}^{k-1}, \bar{s}^{k-1})$, overall resulting in

$$\limsup_{j\to\infty} \Phi_{k,j} \leq \varphi(\bar{x}^{k-1}, \bar{s}^{k-1}) \overset{(6.1)}{\leq} \Phi_{k-1} - \frac{1-\alpha}{2\gamma_{k-1}}\|\bar{x}^{k-1} - x^{k-1}\|^2.$$

Since $\|\bar{x}^{k-1} - x^{k-1}\| > 0$ and $\beta < 1$, for j large enough the condition at step 2.8 will be violated and therefore the k-th iteration successfully terminated.

• 4.2*(ii)*: Follows by combining (6.1) with the failure of the condition at step 2.8 at the end of the iteration.

• 4.2*(iii)*: Direct consequence of assertion 4.2*(ii)*. □

Proof of Theorem 4.3

Proof. First, the sequence $\{(x^k, s^k)\}$ of iterates is well-defined by Lemma 4.2.

- 4.3*(i)*: It follows from Lemma 4.2*(ii)* that $\{\Phi_k\}$ is monotonically decreasing. Lower boundedness of φ gives convergence of the sequence to some finite value from above.

- 4.3*(ii)*: A telescoping argument on the inequality in Lemma 4.2*(ii)* yields

$$\beta(1-\alpha)\sum_{k\in\mathbb{N}}\frac{1}{2\gamma_k}\|\bar{x}^k - x^k\|^2 \leq \Phi_0 - \inf\varphi < \infty, \tag{6.2}$$

whence the claimed finite sum.

- 4.3*(iii)*: By assertion 4.3*(ii)* it follows that $\frac{1}{\gamma_k}\|\bar{x}^k - x^k\|^2 \to 0$, and then it is $\|\bar{x}^k - x^k\| \to 0$ since γ_k is upper bounded. Next, by the conditions at steps 2.2 and 2.3, observe that

$$\begin{aligned}\|x^k - x^{k-1}\| &= \|(1-\tau_k)(\bar{x}^{k-1} - x^{k-1}) + \tau_k d^k\| \\ &\leq (1+\mathfrak{D})\|\bar{x}^{k-1} - x^{k-1}\|\end{aligned} \tag{6.3}$$

and thus $\|x^k - x^{k-1}\|$ vanishes, and in turn so does $\|\bar{x}^k - \bar{x}^{k-1}\|$ since

$$\|\bar{x}^k - \bar{x}^{k-1}\| \leq \|x^k - \bar{x}^k\| + \|\bar{x}^{k-1} - x^{k-1}\| + \|x^k - x^{k-1}\|.$$

- 4.3*(vi)*: The first implication follows from Lemma 4.2*(iii)*, and the second one from assertion 4.3*(ii)* and local boundedness of $\mathcal{O}$ by Proposition 4.1*(iv)*. Finally, if $\{(x^k, s^k)\}$ is bounded, and thus so is $\{(\bar{x}^k, \bar{s}^k)\}$, the set Ω_k in the proof of Lemma 4.2*(i)* can be taken independent of k, and asymptotic constancy of γ_k follows from the same arguments therein.

- 4.3*(iv)*: By iteratively applying inequality (6.3), we obtain that

$$\begin{aligned}\|x^k - x^0\| \leq (1+\mathfrak{D})\sum_{j=0}^{k-1}\|\bar{x}^j - x^j\| &= (1+\mathfrak{D})\sum_{j=0}^{k-1}\frac{\|\bar{x}^j - x^j\|}{\gamma_j^{1/2}}\gamma_j^{1/2} \\ &\leq (1+\mathfrak{D})\sqrt{\sum_{j=0}^{k-1}\frac{\|\bar{x}^j - x^j\|^2}{\gamma_j}}\sqrt{\sum_{j=0}^{k-1}\gamma_j} \\ &\overset{(6.2)}{\leq} (1+\mathfrak{D})\sqrt{2\frac{\Phi_0 - \inf\varphi}{\beta(1-\alpha)}}\sqrt{\sum_{j=0}^{k-1}\gamma_j}.\end{aligned}$$

Contrary to the claim, if $\sum_{k\in\mathbb{N}}\gamma_k < \infty$ holds, then $\{x^k\}$ is bounded. From assertion 4.3*(vi)* proven above we then infer that γ_k is asymptotically constant, thus contradicting the finiteness of $\sum_{k\in\mathbb{N}}\gamma_k$.

- 4.3*(v)*: Immediate consequence of assertions 4.3*(ii)* and 4.3*(iv)*. □