



# A proximal augmented Lagrangian method for nonconvex optimization with equality and inequality constraints

Adeyemi D. Adeoye<sup>1</sup> · Puya Latafat<sup>1</sup> · Alberto Bemporad<sup>1</sup>

Received: 3 September 2025 / Accepted: 30 September 2026  
© The Author(s) 2026

## Abstract

We propose an inexact proximal augmented Lagrangian method (P-ALM) for solving nonconvex structured optimization problems. The proposed method features an easily implementable rule not only for updating the penalty parameters, but also for adaptively tuning the proximal term. It allows the penalty parameter to grow rapidly in the early stages to speed up progress, while ameliorating the issue of ill-conditioning in later iterations, a well-known drawback of the traditional approach of linearly increasing the penalty parameters. A key element in our analysis lies in the observation that the augmented Lagrangian can be controlled effectively along the iterates, provided an initial feasible point is available. Our analysis, while simple, provides a new theoretical perspective about P-ALM and, as a by-product, results in similar convergence properties for its non-proximal variant, the classical augmented Lagrangian method. Numerical experiments, including convex and nonconvex problem instances, demonstrate the effectiveness of our approach.

**Keywords** Augmented Lagrangian method · Nonlinear programming · Inexact proximal point algorithm · Constrained optimization · KKT conditions

**Mathematics Subject Classification** 65K05 · 93-08 · 49M37 · 90C06 · 90C53

---

P. Latafat is a member of the Gruppo Nazionale per l'Analisi Matematica, la Probabilità e le loro Applicazioni (GNAMPA - National Group for Mathematical Analysis, Probability and their Applications) of the Istituto Nazionale di Alta Matematica (INdAM - National Institute of Higher Mathematics).

---

✉ Adeyemi D. Adeoye  
adeyemi.adeoye@imtlucca.it

Puya Latafat  
puya.latafat@imtlucca.it

Alberto Bemporad  
alberto.bemporad@imtlucca.it

<sup>1</sup> IMT School for Advanced Studies Lucca, Lucca, Italy

# 1 Introduction

We consider the constrained nonlinear programming problem

$$\begin{aligned} & \underset{x \in \mathbb{R}^n}{\text{minimize}} && f(x) := f_1(x) + f_2(x) \\ & \text{subject to} && h(x) = 0 \\ & && g(x) \leq 0, \end{aligned} \quad (1)$$

where  $f_1 : \mathbb{R}^n \rightarrow \mathbb{R}$ ,  $h : \mathbb{R}^n \rightarrow \mathbb{R}^p$  and  $g : \mathbb{R}^n \rightarrow \mathbb{R}^m$  are continuously differentiable and possibly nonconvex functions, and  $f_2 : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$  is a proper closed convex function whose proximal mapping is easy to evaluate. Problem (1) covers many optimization problems in engineering, finance, and machine learning. Typically,  $f_1$  represents an objective function,  $h$  and  $g$  equality and inequality constraints on decision variables and  $f_2$  can capture regularization terms, such as the 1-norm, or encodes simple constraints, like box constraints. Because of this flexible structure, formulations of this type routinely arise across diverse domains, including optimal control [1], statistical machine learning [2], and operations research [3].

The augmented Lagrangian method (ALM), also known as the method of multipliers, first introduced by Hestenes [4] and Powell [5], is a widely used approach for solving constrained optimization problems of the form (1). In [6], Rockafellar proposed an elegant convergence analysis for ALM by viewing it as an instance of the proximal point algorithm (PPA) [7] applied to an operator derived from the dual problem. Specifically, standard ALM updates take the form [6, 7]:

$$\begin{aligned} x^{k+1} &\in \arg \min_{x \in \mathbb{R}^n} \left\{ f(x) + \langle \lambda^k, h(x) \rangle + \frac{\rho}{2} \|h(x)\|^2 + \frac{1}{2\nu_k} \|\max\{0, \nu_k g(x) + \mu^k\}\|^2 \right\} \\ \lambda^{k+1} &= \lambda^k + \rho_k h(x^{k+1}) \\ \mu^{k+1} &= \max\{0, \mu^k + \nu_k g(x^{k+1})\}, \end{aligned} \quad (2)$$

where  $\rho, \nu \in \mathbb{R}_{++}$  are penalty parameters, and  $\lambda \in \mathbb{R}^p$ ,  $\mu \in \mathbb{R}_+^m$  are the Lagrange multipliers for equality and inequality constraints, respectively.

The primary focus of this paper is on the proximal variant of ALM, known as the proximal augmented Lagrangian method (P-ALM). Rockafellar proposed P-ALM as an instance of PPA applied to the operator related to a primal-dual formulation, see (7). The iterates of proximal ALM differs from the ALM formulation in the update of the primal variable. In particular, (2) is replaced with

$$x^{k+1} \in \arg \min_{x \in \mathbb{R}^n} \left\{ \mathcal{L}_{\rho_k, \nu_k, \gamma_k}(x, \lambda^k, \mu^k; x^k) + f_2(x) \right\}, \quad (3)$$

where

$$\mathcal{L}_{\rho, \nu, \gamma}(x, \lambda, \mu; v) := f_1(x) + \langle \lambda, h(x) \rangle + \frac{\rho}{2} \|h(x)\|^2 + \frac{1}{2\nu} \|\max\{0, \nu g(x) + \mu\}\|^2$$

$$-\frac{1}{2\nu}\|\mu\|^2 + \frac{1}{2\gamma}\|x - v\|^2, \quad (4)$$

is the proximal augmented Lagrangian function associated with problem (1). Note that we intentionally isolate  $f_2$  from the definition of proximal augmented Lagrangian as it will be handled without introducing additional multipliers, provided its proximal mapping can be evaluated efficiently. Two distinct uses of “proximal” are at play here. The regularization term  $\frac{1}{2\gamma}\|x - v\|^2$  in (4) is what distinguishes P-ALM from ALM and is what the prefix “P-” refers to throughout, whereas the proximal mapping of  $f_2$  just mentioned is how the nonsmooth term is handled by the inner solver and is present in both variants; setting  $\gamma_k \equiv \infty$  removes the former and recovers ALM, leaving the latter unaffected.

## 1.1 Motivation and contributions

Recent years have witnessed a renewed interest in P-ALM and its non-proximal variant, ALM. Initially popular because they could efficiently solve a wide range of optimization problems, they were eclipsed by interior-point methods; however, their superior scalability on very large-scale, highly structured problems has spurred a resurgence among researchers seeking alternatives to conventional interior-point and sequential quadratic programming methods [8, 9]. In this work, we focus our convergence analysis primarily on P-ALM, with the results for ALM derived as a by-product.

A major advantage of the P-ALM, over its non-proximal variant, is that it results in Lagrangian subproblems that are better behaved in general. For instance, depending on the proximal parameter, a nonconvex subproblem may become (strongly) convex, or a lower unbounded subproblem may be guaranteed to admit a solution. Moreover, the proximal variant is closely related to many popular numerical methods such as the alternating direction method of multipliers [10–13], predictor-corrector proximal multiplier [14], and primal-dual hybrid gradient method [15]. We refer the reader to [8, 16] for a complete account of developments and variants of augmented Lagrangian methods.

The classical convergence analysis for augmented Lagrangian type methods crucially hinges on the maximal monotonicity assumption for the associated operator [13, 17]. Such analysis readily accommodates inexact variants under summability of the error sequence. The inexact case is particularly relevant for large-scale or structured problems, where solving subproblems exactly is impractical. In the monotone setting, relatively inexact variants of ALM have also been studied eliminating the need for a predefined sequence of summable inexactness tolerance while ensuring convergence (see [18]).

However, these powerful ideas do not extend to the nonconvex setting studied here. Instead, we adopt an inexactness framework as in [19, 20], which involves finding an approximate stationary point of the (proximal) augmented Lagrangian subproblem up to a fixed tolerance. Consequently, our analysis establishes termination at an approximate KKT point (see Definition 1) rather than proving sequential convergence.

Our convergence analysis is largely inspired by recent results for ALM in [21, 22]. A common feature of these works, shared also with ours, is the assumption

that an initial feasible point is known, which is used to ensure the boundedness of the augmented Lagrangian function (see [23, Ex. 4.12]; see also Fig. 1 for a practical illustration). While [21] considers both equality and inequality constraints, it implicitly requires the Lagrange multipliers to converge to zero for the penalty parameters to remain bounded. This limitation arises from the specific way the penalty parameters are updated, involving a scaled norm of the multiplier vectors [21, Eqs. (2.8) and (2.9)]. This issue was resolved in [22] for problems with equality constraints within an ALM framework. This mechanism is different from the multiplier safeguarding used in [24, 25] and, classically, in [23]. Rather than projecting or clipping the multipliers to a prescribed bounded set, we control their growth as a by-product of the initial feasibility condition. In fact,  $\lambda^k, \mu^k$  need not stay bounded, but their growth remains tied to that of the penalty parameters (Theorem 2(i)), which is what the finite-time termination argument uses. Building on these works, we provide a new analysis for P-ALM addressing (1) in its full generality, with ALM recovered as a special case. Unlike [22], which treats equality constraints only, we also handle inequalities. The analysis is carried out for the proximal variant directly, with ALM recovered as the special case  $\gamma_k \equiv \infty$ , so the growth conditions on  $\gamma_k$  in Assumption II must interact correctly with the penalty schedule. Moreover, the nonsmooth term  $f_2$  enters our notion of  $\varepsilon$ -KKT point (Definition 1) through a proximal residual rather than a gradient, so the stationarity measure the inner solver must certify is correspondingly different.

### 1.1.1 Adaptive parameter updates

A crucial practical issue in augmented Lagrangian methods lies in the update of the penalty parameters  $\rho_k, \nu_k$ . The standard approach is to increase the penalty parameters linearly, e.g.,  $\rho_{k+1} = \xi \rho_k$ , where  $\xi > 1$ , whenever the quantity  $\|h(x^{k+1})\|_\infty$  is not improved compared to the prior iterate (and similarly for the penalty parameter  $\nu_k$  associated with the inequality constraints). However, it is well-known that a linear rate of increase for the penalty parameter can lead to ill-conditioning [26, 27]. Recent nonconvex augmented Lagrangian methods with a global convergence guarantee [24, 25, 28], as well as the classical safeguarded framework of [23], all update the penalty this same multiplicative way. In contrast, whenever the feasibility measure does not improve, our analysis allows for parameter updates of the form

$$\rho_{k+1} = \max \left\{ \rho_k, \rho_0(k+1)^\alpha \right\},$$

where the penalty parameter grows rapidly in the initial phases but asymptotically grows at a much lower rate than linear growth, thereby ameliorating the issue of ill-conditioning. The penalty parameter  $\nu_k$  is updated similarly (see our **FS-P-ALM** algorithm below for details). Throughout, the prefix FS stands for *feasible start* as the methods assume that a feasible  $x^0$  is available. This is crucial for the control of the augmented Lagrangian and of the multiplier sequence (see Lemma 1 and Theorem 2).

Moreover, our analysis suggests an update rule for the proximal parameter  $\gamma_k$  presented in Step 7 of **FS-P-ALM** (see also Assumption II). This results in stronger regularization in early iterations while gradually transitioning the method toward a standard ALM variant.

## 1.2 Other related work

While our primary focus is the P-ALM, we highlight that several notable works have studied the classical ALM in the convex setting beyond the traditional PPA framework, particularly with respect to iteration-complexity estimates, see [20, 29–34]. For nonconvex problems with equality constraints [35–37] employed a regularity condition to study the global convergence and worst-case iteration-complexity bounds for ALM. The recent work [38] derived lower complexity bounds for certain classes of first-order methods that solve linearly-constrained nonconvex problems, including ALM. ALM variants in which simple constraints such as boxes or polytopes are handled separately from the equality and inequality constraints were studied in [20, 39–41]. For P-ALM, this formulation was studied in [42, 43], where iteration-complexity bounds were derived for nonconvex problems with linear constraints.

Despite its extensive practical use, P-ALM has not yet been fully understood for nonconvex problems. Recent studies such as [43, 44], derive outer-complexity estimates for special cases of (1), such as linearly-constrained and/or equality-constrained problems. Because of its close connection to the alternating direction method of multipliers, other seemingly-tangential but important results were developed in [42, 45, 46] for problems with linear equality constraints, and in [47, 48] where the objective may be nonsmooth and the penalty parameter is kept constant. The authors in [49, 50] focused on variants of the P-ALM where an accelerated gradient method solves nonconvex subproblems and demonstrates improved computational performance for linearly-constrained problems. Finally, [51] developed an inexact P-ALM for quadratic programming (QP) problems with linear constraints.

In contrast to most of these works, we present a general P-ALM framework that handles nonlinear equality and inequality constraints, as well as nonsmooth terms. Despite its simplicity, our analysis covers both the more stabilized Proximal variant (P-ALM) and the standard ALM settings while suggesting practical rules for updating the penalty parameters and adaptively tuning the proximal term. Additionally, our approach guarantees explicit control of the augmented Lagrangian and the multipliers under an initial feasibility condition (see Lemma 1). This feature is central to our main convergence results presented in Theorem 1. Overall, our results provide new theoretical guarantees and practical guidelines for designing efficient P-ALM algorithms applicable to a wide class of constrained optimization problems.

## 1.3 Organization

The rest of this paper is organized as follows. In Sect. 1.4, we introduce the notation conventions used throughout the paper. Then, in Sect. 2, we introduce the main definitions and problem assumptions adopted throughout the paper. Section 3 describes the proposed P-ALM (**FS-P-ALM**) and presents its main convergence properties. In Sect. 4, we establish several key theoretical properties of **FS-P-ALM** and present the proof of the main convergence results. We present an ALM (**FS-ALM**) as a special case of **FS-P-ALM** in Sect. 5 and establish its key theoretical properties from those of **FS-P-ALM**. Finally, Sect. 6 reports results from numerical simulations to illustrate

the performance of our methods on both convex and nonconvex problems, including instances from the Maros-Mészáros collection [52] and the basis pursuit problem [53] on synthetic datasets.

## 1.4 Notations

The set of real and extended-real numbers are  $\mathbb{R} := (-\infty, \infty)$  and  $\overline{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$ , while the positive and strictly positive reals are  $\mathbb{R}_+ := [0, \infty)$  and  $\mathbb{R}_{++} := (0, \infty)$ . We use the notation  $[x]_+ = \max\{0, x\}$ . With  $\text{id}$  we indicate the identity function  $x \mapsto x$  defined on a suitable space. We denote by  $\mathbb{R}^n$  the standard  $n$ -dimensional Euclidean space with inner product  $\langle \cdot, \cdot \rangle$  and induced norm  $\|\cdot\|$ . For a fixed integer  $n \geq 1$  we write  $\mathbb{R}_+^n := [0, \infty)^n = \{x = (x_1, \dots, x_n) \in \mathbb{R}^n \mid x_i \in [0, \infty) \text{ for all } i = 1, \dots, n\}$ . For a vector  $w = (w_1, \dots, w_N) \in \mathbb{R}^n$ ,  $w_i \in \mathbb{R}^{n_i}$  is used to denote its  $i$ -th (block) coordinate. The infinity norm, and 1-norm are denoted by  $\|\cdot\|_\infty$ , and  $\|\cdot\|_1$ , respectively. We use the notation  $(z^k)_{k \in I}$  to denote a sequence with indices in the set  $I \subseteq \mathbb{N}$ . When dealing with scalar sequences we use the subscript notation  $(\gamma_k)_{k \in I}$ . The set of indices  $\{1, 2, \dots, m\}$  is denoted by  $[m]$ .

An operator or set-valued mapping  $A : \mathbb{R}^n \rightrightarrows \mathbb{R}^d$  maps each point  $x \in \mathbb{R}^n$  to a subset  $A(x)$  of  $\mathbb{R}^d$ . We denote the domain of  $A$  by  $\text{dom } A := \{x \in \mathbb{R}^n \mid A(x) \neq \emptyset\}$ . For an extended real-valued function  $f : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$ , the domain is defined as  $\text{dom } f := \{x \in \mathbb{R}^n \mid [f(x) < \infty]\}$ . We say that  $f$  is proper if  $\text{dom } f \neq \emptyset$  and that  $f$  is closed or lower semicontinuous if its epigraph  $\text{epi } f := \{(x, \alpha) \in \mathbb{R}^n \times \mathbb{R} \mid f(x) \leq \alpha\}$  is a closed subset of  $\mathbb{R}^{n+1}$ . For a differentiable function  $f$ ,  $\nabla_x f$  denotes its gradient with respect to  $x$ . For a convex function  $f$ ,  $\partial_x f$  is used to denote the subdifferential of  $f$  with respect to  $x$ . When the differentiation is taken with respect to the full variable (i.e., all components), we omit the subscript and simply write  $\partial f$ . For a vector-valued function  $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ , we denote the Jacobian of  $f$  at  $x$  by  $J_f(x)$ . The indicator function of a set  $C \subseteq \mathbb{R}^n$  is denoted by  $\delta_C$ , namely  $\delta_C(x) = 0$  if  $x \in C$  and  $\delta_C(x) = \infty$  if  $x \notin C$ . We denote the normal cone of  $C$  by  $N_C$ . The proximal operator of a function  $f$  at  $x$  is defined as

$$\text{prox}_{\gamma f}(x) := \arg \min_{y \in \mathbb{R}^n} \left\{ f(y) + \frac{1}{2\gamma} \|y - x\|^2 \right\}, \quad \forall x \in \mathbb{R}^n, \gamma \in \mathbb{R}_{++}.$$

## 2 Assumptions and preliminaries

In the following, we present the key assumptions we use throughout the paper.

**Assumption 1** (*Problem assumptions*). The following conditions hold for problem (1):

- A<sub>1</sub> The functions  $f_1 : \mathbb{R}^n \rightarrow \mathbb{R}$ ,  $h : \mathbb{R}^n \rightarrow \mathbb{R}^p$ , and  $g : \mathbb{R}^n \rightarrow \mathbb{R}^m$  are continuously differentiable.
- A<sub>2</sub> The function  $f_2 : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$  is proper, closed and convex, and its proximal mapping is easy to compute.
- A<sub>3</sub> There exists  $\bar{x} \in \text{dom } f_2$  such that  $h(\bar{x}) = 0$  and  $g(\bar{x}) \leq 0$ .

A<sub>4</sub> The objective function  $f$  is lower-bounded by a finite number, i.e.,  $-\infty < f^* := \inf f$  for some  $f^* \in \mathbb{R}$ .

The lower boundedness of the objective function required in Assumption I.A<sub>4</sub> ensures the well-definedness of the Lagrangian subproblems, as it guarantees the existence of  $\varepsilon$ -stationary points required in (9).

A tuple  $(x, \lambda, \mu)$  of primal and dual variables satisfies the Karush-Kuhn-Tucker (KKT) conditions for problem (1) if:

$$\begin{aligned} \|x - \text{prox}_{f_2}(x - \nabla_x \mathcal{L}(x, \lambda, \mu))\| &= 0, \\ h(x) &= 0, \quad g(x) \leq 0, \\ \mu_i &\geq 0, \quad \mu_i g_i(x) = 0 \quad \text{for all } i \in [m], \end{aligned} \quad (5)$$

where the Lagrangian function  $\mathcal{L}$  is defined by

$$\mathcal{L}(x, \lambda, \mu) := f_1(x) + \langle \lambda, h(x) \rangle + \langle \mu, g(x) \rangle. \quad (6)$$

Here, the primal stationarity condition is expressed through the natural residual mapping  $x - \text{prox}_{f_2}(x - \nabla_x \mathcal{L}(x, \lambda, \mu))$  that vanishes if and only if  $-\nabla_x \mathcal{L}(x, \lambda, \mu) \in \partial f_2(x)$ .

Finding a point that satisfies the KKT conditions is equivalent to finding a zero of the primal-dual (KKT) operator  $T: \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^m \rightrightarrows \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^m$  given by

$$T(x, \lambda, \mu) := \begin{bmatrix} \nabla f_1(x) + \partial f_2(x) + J_h(x)^\top \lambda + J_g(x)^\top \mu \\ -h(x) \\ -g(x) + \mathcal{N}_{\mathbb{R}_+^m}(\mu) \end{bmatrix}.$$

That is, the first order optimality conditions of P-ALM updates in (3) can be written compactly as

$$(\frac{1}{\gamma_k}(x^k - x^{k+1}), \frac{1}{\rho_k}(\lambda^k - \lambda^{k+1}), \frac{1}{v_k}(\mu^k - \mu^{k+1})) \in T(x^{k+1}, \lambda^{k+1}, \mu^{k+1}). \quad (7)$$

In practice, one aims to solve these fixed-point equations approximately. To this end, we consider the following notion of  $\varepsilon$ -KKT optimality for P-ALM subproblems.

**Definition 1** ( $\varepsilon$ -KKT optimality). A point  $(x, \lambda, \mu)$  is called an  $\varepsilon$ -KKT point of (1) if it satisfies the following conditions:

$$\begin{aligned} \|x - \text{prox}_{f_2}(x - \nabla_x \mathcal{L}(x, \lambda, \mu))\|_\infty &\leq \varepsilon, \quad \|h(x)\|_\infty \leq \varepsilon, \quad \|\max\{0, g(x)\}\|_\infty \leq \varepsilon, \\ \mu_i &\geq 0, \quad \mu_i = 0 \quad \text{whenever } g_i(x) < -\varepsilon, \quad \text{for all } i \in [m]. \end{aligned}$$

The  $\varepsilon$ -KKT conditions are a relaxation of the standard KKT conditions, allowing for approximate satisfaction of the optimality conditions. They are akin to those of [20] which studies the specific case where  $f_2$  is the indicator function of a box.

**Algorithm 1** **FS-P-ALM** ((Inexact) proximal feasible-start ALM)

**REQUIRE** Initial feasible point  $x^0 \in \text{dom } f_2$  such that  $h(x^0) = 0$ ,  $g(x^0) \leq 0$   
 Multipliers  $\lambda^0 \in \mathbb{R}^p$  and  $\mu^0 \in \mathbb{R}_+^m$  and penalties  $\rho_0, v_0, \gamma_0 \in \mathbb{R}_{++}$   
 Parameters  $\beta \in (0, 1)$ ,  $\xi_1, \xi_2 \geq 1$ , and  $\delta, \hat{\rho}, \hat{v}, \hat{\gamma} \in \mathbb{R}_{++}$   
 Sequence of tolerances  $(\tau_k)_{k \in \mathbb{N}} \subset \mathbb{R}_{++}$  with  $\tau_k \rightarrow \tau$  for some  $\tau \in \mathbb{R}_+$   
 Functions  $\phi, \phi_\gamma$  satisfying Assumption II

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

- 1: Set  $\hat{x}^k = x^k$  if  $\mathcal{L}_{\rho_k, v_k, \gamma_k}(x^k, \lambda^k, \mu^k; x^k) + f_2(x^k) \leq f(x^0) + \frac{1}{2\gamma_k} \|x^0 - x^k\|^2$ ,  
 or  $\hat{x}^k = x^0$  otherwise
- 2: Starting at  $\hat{x}^k$ , find a point  $x^{k+1} \in \text{dom } f_2$  approximately solving

$$\text{minimize } \left\{ \mathcal{L}_{\rho_k, v_k, \gamma_k}(\cdot, \lambda^k, \mu^k; x^k) + f_2(\cdot) \right\},$$

such that

$$\|x^{k+1} - \text{prox}_{f_2}(x^{k+1} - \nabla_x \mathcal{L}_{\rho_k, v_k, \gamma_k}(x^{k+1}, \lambda^k, \mu^k; x^k))\|_\infty \leq \tau_k \quad (9)$$

- 3: Set  $\lambda^{k+1} = \lambda^k + \rho_k h(x^{k+1})$
- 4: Set  $\mu^{k+1} = \max\{0, \mu^k + v_k g(x^{k+1})\}$
- 5: Set  $\rho_{k+1} = \rho_k$  if  $\|h(x^{k+1})\|_\infty \leq \beta \|h(x^k)\|_\infty$ ,  
 or  $\rho_{k+1} = \max\{\xi_1 \rho_k, \hat{\rho} \phi(k+1)\}$  otherwise
- 6: Set  $v_{k+1} = v_k$  if  $\|E^{k+1}\|_\infty \leq \beta \|E^k\|_\infty$  where  $E_k$  is as in (8),  
 or  $v_{k+1} = \max\{\xi_2 v_k, \hat{v} \phi(k+1)\}$  otherwise
- 7: Set  $\gamma_{k+1} = \max\{\delta \|x^0 - x^{k+1}\|^2, \hat{\gamma} \phi_\gamma(k+1)\}$

### 3 An inexact proximal ALM

Consider the inexact proximal augmented Lagrangian method **FS-P-ALM** (defined in Algorithm 1). Given an initial feasible point  $x^0$ , the algorithm sets the penalty parameters  $\rho_0$  and  $v_0$ , and the associated Lagrange multipliers which are iteratively updated in Steps 3 to 6. As is common in the literature, **FS-P-ALM** probes the changes in the quantity  $\|h(x^k)\|_\infty$  along iterations to decide if the penalty parameter corresponding to the equality constraints should be increased or not. For inequality constraints, it employs the quantity  $E^k$  defined by

$$E^k := \min \left\{ -g(x^k), \frac{1}{v_{k-1}} \mu^{k-1} \right\}, \quad (8)$$

which serves as a surrogate of approximate complimentary slackness (see Lemma 4). In particular, we use this quantity to determine when the penalty parameter  $v_k$  should be increased.

The main difference from the standard Augmented Lagrangian literature lies in the way the penalty parameters are updated. **FS-P-ALM** introduces a general update rule using a function  $\phi$  that, while encompassing the standard linear growth, also allows for alternatives such as  $\phi(k) = (k+1)^\alpha$  for  $\alpha > 1$ , which grow more aggressively in the early stages but increases more moderately in the long runs. This approach helps mitigating the numerical issues known to arise when a purely linear rate of increase is used. Our proposed update rule for the proximal parameter  $\gamma_k$  uses a separate function

$\phi_\gamma$ :

$$\gamma_k = \max \left\{ \delta \|x^0 - x^k\|^2, \gamma_0 \phi_\gamma(k) \right\},$$

where  $\delta \in \mathbb{R}_{++}$  is a constant. Our standing assumption for  $\phi$  and, separately, for  $\phi_\gamma$  is as follows.

**Assumption II** (*Growth function*). A function  $\psi: \mathbb{N} \rightarrow \mathbb{R}_{++}$  satisfies the growth condition if

$$\limsup_{k \rightarrow \infty} \frac{\psi(k+1)}{\psi(k)} < M, \quad \lim_{k \rightarrow \infty} \frac{k}{\psi(k)} \rightarrow 0,$$

for some  $M \in \mathbb{R}_{++}$ . We require that both the function  $\phi$ , appearing in the update rule for  $\rho_k$  and  $v_k$ , and the function  $\phi_\gamma$ , appearing in the update rule for  $\gamma_k$ , satisfy this growth condition.

In practice, we suggest to use a monotonically increasing  $\phi(k) = (k+1)^\alpha$ , for some  $\alpha > 1$ , as it allows for a quick recovery from a small initialization of the penalty parameters, while it ameliorates ill-conditioning typically observed with standard linear increments. For the proximal parameter, we recommend choosing  $\phi_\gamma(k) = (k+1)^{\alpha_\gamma}$  with  $\alpha_\gamma > 1$  close to 1. We remark that  $\rho_k$  and  $v_k$  could in principle also use different growth functions from one another (each still satisfying Assumption II). For simplicity of presentation we keep them tied to the same function  $\phi$ , distinct from the function  $\phi_\gamma$  used for the proximal parameter  $\gamma_k$ . We refer the reader to Sect. 6.1 for detailed discussion on the choice of parameters.

At each iteration of **FS-P-ALM**, a minimization problem involving the proximal augmented Lagrangian is solved. To ensure convergence, it is required that the algorithm provides a stationary point of the sum of the proximal augmented Lagrangian and  $f_2$  with a maximum tolerance of  $\tau_k \in \mathbb{R}_{++}$  at iteration  $k$ , where  $(\tau_k)_{k \in \mathbb{N}}$  is a sequence of positive tolerances converging to a constant  $\tau \in \mathbb{R}_+$ .

Because **FS-P-ALM** only requires  $x^{k+1}$  to satisfy (9) approximately, rather than to exactly solve (3), the inclusion (7) no longer holds verbatim for its iterates. Its second and third blocks remain exact regardless of how  $x^{k+1}$  is computed, since  $\frac{1}{\rho_k}(\lambda^k - \lambda^{k+1}) = -h(x^{k+1})$  and  $\frac{1}{v_k}(\mu^k - \mu^{k+1}) \in -g(x^{k+1}) + \mathcal{N}_{\mathbb{R}_+^m}(\mu^{k+1})$  follow directly from Steps 3 and 4. The first (primal stationarity) block, in contrast, picks up an additional error of size  $\tau_k$ , as made precise in Lemma 2(ii).

The main convergence properties of **FS-P-ALM** are stated in Theorem 1.

**Theorem 1** ( $\varepsilon$ -KKT termination for **FS-P-ALM**). Suppose that Assumptions I and II hold. Consider the sequence  $(x^k, \lambda^k, \mu^k)_{k \in \mathbb{N}}$  generated by **FS-P-ALM**. Then, the following hold:

- (i) The sequence  $(\frac{1}{\gamma_k} \|x^{k+1} - x^k\|)_{k \in \mathbb{N}}$  converges to zero;
- (ii) There exists an infinite subsequence  $\mathcal{K} \subseteq \mathbb{N}$  along which  $(\frac{1}{\rho_k} \|\lambda^{k+1} - \lambda^k\|^2 + \frac{1}{v_k^2} \|\mu^{k+1} - \mu^k\|^2)_{k \in \mathcal{K}}$  converges to zero;

(iii) An  $\varepsilon$ -KKT point is obtained in a finite number of iterations, for any  $\varepsilon > \tau \in \mathbb{R}_+$ .

The proof of Theorem 1 is deferred to Sect. 4, where several properties of **FS-P-ALM** are derived. Theorem 1(i) follows directly from Theorem 2(ii). Observe that, as a direct implication of how  $\lambda^k$  and  $\mu^k$  are updated in **FS-P-ALM**,

$$\frac{1}{\rho_k^2} \|\lambda^{k+1} - \lambda^k\|^2 + \frac{1}{v_k^2} \|\mu^{k+1} - \mu^k\|^2 = \|h(x^{k+1})\|^2 + \|E^{k+1}\|^2. \quad (10)$$

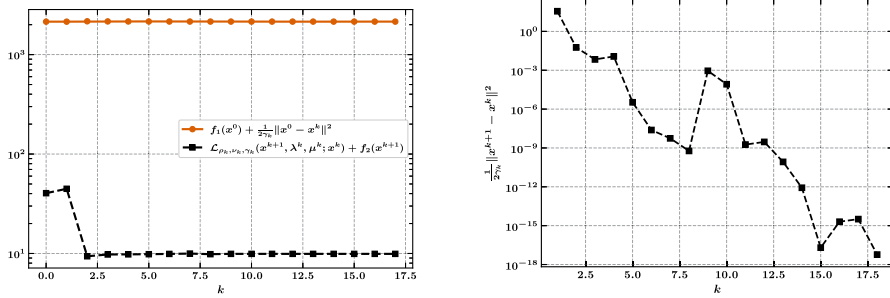
Hence, in Sect. 4, we prove Theorem 1(ii) by showing that  $\|h(x^{k+1})\|^2$  and  $\|E^{k+1}\|^2$  tend to zero along an infinite subsequence  $\mathcal{K} \subseteq \mathbb{N}$ , i.e., Theorem 1(ii) follows from (10) together with Theorem 3. In view of (7), and using the characterization of an  $\varepsilon$ -KKT point that follows from Lemma 4, we consequentially prove Theorem 1(iii).

As shown in Theorem 1(i), the quantity  $\frac{1}{\gamma_k} \|x^{k+1} - x^k\|$  vanishes independently of the tolerances  $\tau_k$ . This follows from  $\gamma_k$  growing faster than linearly (Assumption II) while  $\frac{1}{2\rho_k} \|\lambda^k\|^2 + \frac{1}{2v_k} \|\mu^k\|^2$  grows at most linearly (Theorem 2(i)), as shown in the proof of Theorem 2. As noted above, this is precisely the correction that the first block of (7) acquires once  $x^{k+1}$  is only computed inexactly. By Lemma 2(ii), the genuine stationarity gap of  $(x^{k+1}, \lambda^{k+1}, \mu^{k+1})$  is bounded by  $\tau_k$  plus this vanishing quantity, and therefore only gets arbitrarily close to  $\tau := \lim_k \tau_k$ . Hence, Theorem 1(iii) guarantees an  $\varepsilon$ -KKT point for any  $\varepsilon > \tau$ , rather than as  $\varepsilon \rightarrow 0$ .

**Remark 1** (AKKT/AM-stationarity). Combined with Theorem 1(ii), the above shows that a selection of the KKT operator  $T$  at  $(x^{k+1}, \lambda^{k+1}, \mu^{k+1})$ , as in (5), vanishes along  $\mathcal{K}$  up to the tolerance  $\tau_k$ . This is the sequential residual condition defining an approximate KKT (AKKT) point [54, 55], extended to composite problems of the form (1) as AM-stationarity in [24], except that the residual is driven to  $\tau$  rather than to zero, so that the two coincide exactly when  $\tau = 0$ . Accordingly, if  $\tau = 0$  and  $x^k \rightarrow x^*$  along a subsequence of  $\mathcal{K}$ , then  $x^*$  is AM-stationary, and a sequential constraint qualification at  $x^*$  upgrades this to a genuine KKT point. Our focus is instead the finite-time attainment of an  $\varepsilon$ -KKT point for a fixed tolerance  $\varepsilon > \tau$ .

**Remark 2** (Penalty parameters and initial feasibility).

- (i) The use of scalar penalty parameters  $\rho_k, v_k$  for equality and inequality constraints in **FS-P-ALM** is only for clarity of exposition, and our theoretical results extend trivially to the case of (block) diagonal penalty matrices. Specifically, given any decomposition  $h(x) = (h_1(x), h_2(x), \dots, h_{n_h}(x))$ , we may assign  $\rho_{i,k}$  as the penalty parameter corresponding to  $h_i$  with constraint violation measured in terms of  $\|h_i(x^k)\|_\infty$ .
- (ii) In many applications, such as those arising in optimal control, an initial feasible point can be obtained at little to no cost by simply unrolling the dynamics through a forward pass. More generally, however, a feasible point may be computed by solving a so-called *phase I* problem, as is common in the literature on barrier-type methods [56], see also [3, Section 11.4]. A brief discussion of this approach, in which **FS-P-ALM** itself may be used to solve the phase I problem is given in Appendix A.



**Fig. 1** Illustration of two defining properties of **FS-P-ALM** central to the proof of Theorem 1, on the basis-pursuit instance of Sect. 6.4 ( $p = 400$ ,  $n = 1024$ ), with  $\alpha = 1.01$  under PANOC. **Left:** the augmented Lagrangian value at  $x^{k+1}$  stays below the reference bound at every iteration. **Right:** the proximal term vanishes

- (iii) **FS-P-ALM** uses the same initial point  $x^0$  in Step 1 throughout iterations. However, in practice, should the algorithm stumble upon a new feasible point along the iterates, the new feasible point may be used to replace  $x^0$ .

## 4 Convergence analysis

We begin by presenting a series of results on the properties of the sequence generated by **FS-P-ALM**. These results are instrumental for the proof of the main result in Theorem 1 at the end of this section.

First, we present a simple observation in Lemma 1 that the augmented Lagrangian is upper bounded by a quantity that depends solely on the variable at the previous iterate and the initial feasible point. The lemma is a trivial consequence of the initialization for the Lagrangian subproblem (9) in Step 1 of **FS-P-ALM**, provided that Step 2 is solved by a descent method.

**Lemma 1** (Lagrangian upper bound). *Let Assumption I.A<sub>3</sub> hold, and let  $x^0$  be an initial primal point of **FS-P-ALM** such that  $h(x^0) = 0$  and  $g(x^0) \leq 0$ . Suppose further that  $x^{k+1}$  in Step 2 of **FS-P-ALM** is computed by a descent method, so that*

$$\mathcal{L}_{\rho_k, v_k, \gamma_k}(x^{k+1}, \lambda^k, \mu^k; x^k) + f_2(x^{k+1}) \leq \mathcal{L}_{\rho_k, v_k, \gamma_k}(\hat{x}^k, \lambda^k, \mu^k; x^k) + f_2(\hat{x}^k).$$

Then, for any  $k \in \mathbb{N}$ ,

$$\mathcal{L}_{\rho_k, v_k, \gamma_k}(x^{k+1}, \lambda^k, \mu^k; x^k) + f_2(x^{k+1}) \leq f(x^0) + \frac{1}{2\gamma_k} \|x^0 - x^k\|^2. \quad (11)$$

We next show that the scaled sum of the multipliers,  $\frac{1}{2\rho_{k+1}} \|\lambda^{k+1}\|^2 + \frac{1}{2v_{k+1}} \|\mu^{k+1}\|^2$ , can increase by at most a fixed constant from one iteration to the next. This key result implies, in particular, that the scaled primal residual  $\frac{1}{\gamma_k} \|x^{k+1} - x^k\|$  vanishes asymptotically.

**Theorem 2** (Multipliers control and primal residual). *Suppose that Assumptions I and II hold, and consider the iterates generated by **FS-P-ALM**. Then, the following hold:*

(i) For any  $k \in \mathbb{N}$

$$\frac{1}{2\rho_{k+1}} \|\lambda^{k+1}\|^2 + \frac{1}{2v_{k+1}} \|\mu^{k+1}\|^2 \leq \frac{1}{2\rho_k} \|\lambda^k\|^2 + \frac{1}{2v_k} \|\mu^k\|^2 + c_1 - \frac{1}{2\gamma_k} \|x^{k+1} - x^k\|^2,$$

and in particular,

$$\frac{1}{2\rho_k} \|\lambda^k\|^2 + \frac{1}{2v_k} \|\mu^k\|^2 \leq \frac{1}{2\rho_0} \|\lambda^0\|^2 + \frac{1}{2v_0} \|\mu^0\|^2 + kc_1, \quad (12)$$

where  $c_1 := f(x^0) - f^* + \frac{1}{2\delta}$ ;

(ii)  $\frac{1}{\gamma_k} \|x^{k+1} - x^k\| \rightarrow 0$  as  $k \rightarrow \infty$ .

**Proof** We can rewrite (4) as

$$\begin{aligned} \mathcal{L}_{\rho,v,\gamma}(x, \lambda, \mu; v) &= f_1(x) + \frac{1}{2\rho} \|\rho h(x) + \lambda\|^2 + \frac{1}{2v} \|\max\{0, v g(x) + \mu\}\|^2 \\ &\quad - \frac{1}{2\rho} \|\lambda\|^2 - \frac{1}{2v} \|\mu\|^2 \\ &\quad + \frac{1}{2\gamma} \|x - v\|^2. \end{aligned} \quad (13)$$

Using Steps 3 and 4 in (13) gives

$$\begin{aligned} \frac{1}{2\rho_k} \|\lambda^{k+1}\|^2 + \frac{1}{2v_k} \|\mu^{k+1}\|^2 &= \mathcal{L}_{\rho_k, v_k, \gamma_k}(x^{k+1}, \lambda^k, \mu^k; x^k) - f_1(x^{k+1}) - \frac{1}{2\gamma_k} \|x^{k+1} - x^k\|^2 \\ &\quad + \frac{1}{2\rho_k} \|\lambda^k\|^2 + \frac{1}{2v_k} \|\mu^k\|^2 \\ &\leq f(x^0) - f^* + \frac{1}{2\gamma_k} \|x^0 - x^k\|^2 - \frac{1}{2\gamma_k} \|x^{k+1} - x^k\|^2 \\ &\quad + \frac{1}{2\rho_k} \|\lambda^k\|^2 + \frac{1}{2v_k} \|\mu^k\|^2, \end{aligned}$$

where the inequality follows from Lemma 1 and Assumption I.A<sub>4</sub>. Now, using the fact that  $\gamma_k \geq \delta \|x^0 - x^k\|^2$  as ensured by Step 7 of **FS-P-ALM**, we have

$$\frac{1}{2\rho_k} \|\lambda^{k+1}\|^2 + \frac{1}{2v_k} \|\mu^{k+1}\|^2 \leq c_1 - \frac{1}{2\gamma_k} \|x^{k+1} - x^k\|^2 + \frac{1}{2\rho_k} \|\lambda^k\|^2 + \frac{1}{2v_k} \|\mu^k\|^2.$$

Since  $\rho_k$  and  $v_k$  are monotonically nondecreasing, this implies that

$$\begin{aligned} \frac{1}{2\rho_{k+1}} \|\lambda^{k+1}\|^2 + \frac{1}{2v_{k+1}} \|\mu^{k+1}\|^2 &\leq \frac{1}{2\rho_k} \|\lambda^{k+1}\|^2 + \frac{1}{2v_k} \|\mu^{k+1}\|^2 \\ &\leq c_1 - \frac{1}{2\gamma_k} \|x^{k+1} - x^k\|^2 + \frac{1}{2\rho_k} \|\lambda^k\|^2 + \frac{1}{2v_k} \|\mu^k\|^2 \\ &\leq \frac{1}{2\rho_k} \|\lambda^k\|^2 + \frac{1}{2v_k} \|\mu^k\|^2 + c_1. \end{aligned} \quad (14)$$

A telescoping argument then yields the bound stated in (12). Now, rearranging (14) gives

$$\frac{1}{2\gamma_k^2} \|x^{k+1} - x^k\|^2 \leq \frac{1}{\gamma_k} \left( c_1 + \frac{1}{2\rho_k} \|\lambda^k\|^2 + \frac{1}{2v_k} \|\mu^k\|^2 \right)$$

$$\begin{aligned} &\leq \frac{1}{\gamma_k} \left( (k+1)c_1 + \frac{1}{2\rho_0} \|\lambda^0\|^2 + \frac{1}{2v_0} \|\mu^0\|^2 \right) \\ &\leq \frac{1}{\hat{\gamma}\phi_\gamma(k)} \left( (k+1)c_1 + \frac{1}{2\rho_0} \|\lambda^0\|^2 + \frac{1}{2v_0} \|\mu^0\|^2 \right), \end{aligned}$$

where the second inequality is due to (12) and the third inequality uses the fact that  $\gamma_k \geq \hat{\gamma}\phi_\gamma(k)$  (see Step 7 of **FS-P-ALM**). Consequently, using Assumption II applied to  $\phi_\gamma$ , we find  $\frac{1}{\gamma_k} \|x^{k+1} - x^k\| \rightarrow 0$  as  $k \rightarrow \infty$ .  $\square$

**Lemma 2** (Fixed-point residual). *Suppose that Assumptions I and II hold, and let  $(x^k, \lambda^k, \mu^k)_{k \in \mathbb{N}}$  be the sequence of iterates generated by **FS-P-ALM**. Then, the following properties hold:*

- (i)  $\nabla_x \mathcal{L}(x^{k+1}, \lambda^{k+1}, \mu^{k+1}) = \nabla_x \mathcal{L}_{\rho_k, v_k, \gamma_k}(x^{k+1}, \lambda^k, \mu^k; x^k) - \frac{1}{\gamma_k} (x^{k+1} - x^k);$
- (ii)  $\|x^{k+1} - \text{prox}_{f_2}(x^{k+1} - \nabla_x \mathcal{L}(x^{k+1}, \lambda^{k+1}, \mu^{k+1}))\|_\infty \leq \varepsilon$ , for  $k$  large enough and any  $\varepsilon > \tau$ .

**Proof** Using Steps 3 and 4 in (6), we get that

$$\begin{aligned} \mathcal{L}(x, \lambda^{k+1}, \mu^{k+1}) &= \mathcal{L}_{\rho_k, v_k, \gamma_k}(x, \lambda^k, \mu^k; x^k) - \frac{\rho_k}{2} \|h(x) - h(x^{k+1})\|^2 \\ &\quad + \frac{\rho_k}{2} \|h(x^{k+1})\|^2 + \frac{1}{2v_k} \|\mu^k\|^2 \\ &\quad + \langle \max\{0, \mu^k + v_k g(x^{k+1})\}, g(x) \rangle - \frac{1}{2v_k} \|\max\{0, \mu^k + v_k g(x)\}\|^2 \\ &\quad - \frac{1}{2\gamma_k} \|x - x^k\|^2, \end{aligned}$$

and

$$\begin{aligned} \nabla_x \mathcal{L}(x, \lambda^{k+1}, \mu^{k+1}) &= \nabla_x \mathcal{L}_{\rho_k, v_k, \gamma_k}(x, \lambda^k, \mu^k; x^k) - \rho_k J_h(x)^\top (h(x) - h(x^{k+1})) \\ &\quad + J_g(x)^\top (\max\{0, \mu^k + v_k g(x^{k+1})\}) \\ &\quad - J_g(x)^\top (\max\{0, \mu^k + v_k g(x)\}) \\ &\quad - \frac{1}{\gamma_k} (x - x^k). \end{aligned}$$

Hence, we have

$$\nabla_x \mathcal{L}(x^{k+1}, \lambda^{k+1}, \mu^{k+1}) = \nabla_x \mathcal{L}_{\rho_k, v_k, \gamma_k}(x^{k+1}, \lambda^k, \mu^k; x^k) - \frac{1}{\gamma_k} (x^{k+1} - x^k). \quad (15)$$

By the triangle inequality, we have

$$\begin{aligned} &\|x^{k+1} - \text{prox}_{f_2}(x^{k+1} - \nabla_x \mathcal{L}(x^{k+1}, \lambda^{k+1}, \mu^{k+1}))\|_\infty \\ &\leq \|x^{k+1} - \text{prox}_{f_2}(x^{k+1} - \nabla_x \mathcal{L}(x^{k+1}, \lambda^{k+1}, \mu^{k+1}) - \frac{1}{\gamma_k} (x^{k+1} - x^k))\|_\infty \\ &\quad + \|\text{prox}_{f_2}(x^{k+1} - \nabla_x \mathcal{L}(x^{k+1}, \lambda^{k+1}, \mu^{k+1}) - \frac{1}{\gamma_k} (x^{k+1} - x^k)) \\ &\quad - \text{prox}_{f_2}(x^{k+1} - \nabla_x \mathcal{L}(x^{k+1}, \lambda^{k+1}, \mu^{k+1}))\|_\infty \\ &\leq \|x^{k+1} - \text{prox}_{f_2}(x^{k+1} - \nabla_x \mathcal{L}_{\rho_k, v_k, \gamma_k}(x^{k+1}, \lambda^k, \mu^k; x^k))\|_\infty + \frac{1}{\gamma_k} \|x^{k+1} - x^k\|, \end{aligned}$$

where the second inequality results from (15), the nonexpansiveness of the proximal operator and  $\|\cdot\|_\infty \leq \|\cdot\|$ . Now, by (9) of Step 2 in **FS-P-ALM**, we get

$$\|x^{k+1} - \text{prox}_{f_2}(x^{k+1} - \nabla_x \mathcal{L}(x^{k+1}, \lambda^{k+1}, \mu^{k+1}))\|_\infty \leq \tau_k + \frac{1}{\gamma_k} \|x^{k+1} - x^k\|.$$

Since  $\tau_k \rightarrow \tau$  and, by Theorem 2(ii),  $\frac{1}{\gamma_k} \|x^{k+1} - x^k\| \rightarrow 0$ , for  $k$  large enough the right-hand side is smaller than any  $\varepsilon > \tau$ .  $\square$

For every iteration  $k$  at which the penalty parameters increase in Steps 5 and 6 of **FS-P-ALM**, we derive upper bounds on the feasibility and complementarity of  $(x^k, \lambda^k, \mu^k)$  in the following lemma.

**Lemma 3** (Feasibility and complementarity control). *Suppose that Assumptions I and II hold, and consider the sequence  $(x^k, \lambda^k, \mu^k)_{k \in \mathbb{N}}$  of iterates generated by **FS-P-ALM**. Define the index sets*

$$\mathcal{K}_\rho := \{k \in \mathbb{N} \mid \|h(x^k)\|_\infty > \beta \|h(x^{k-1})\|_\infty\}, \quad \mathcal{K}_v := \{k \in \mathbb{N} \mid \|E^k\|_\infty > \beta \|E^{k-1}\|_\infty\}, \quad (16)$$

and let

$$A_k := 2k(f(x^0) - f^* + \frac{1}{2\delta}) + \frac{1}{\rho_0} \|\lambda^0\|^2 + \frac{1}{v_0} \|\mu^0\|^2.$$

Then, the following hold:

(i) For any  $k \in \mathcal{K}_v$ ,

$$\frac{1}{2} \|E^{k+1}\|^2 \leq \frac{1}{v} \max \left\{ \frac{\xi_2}{\phi(k)}, \frac{\phi(k+1)}{\phi(k)^2} \right\} A_{k+1} + \frac{1}{\hat{v}\phi(k)} A_k; \quad (17)$$

(ii) For any  $k \in \mathcal{K}_\rho$ ,

$$\frac{1}{2} \|h(x^{k+1})\|^2 \leq \frac{1}{\rho} \max \left\{ \frac{\xi_1}{\phi(k)}, \frac{\phi(k+1)}{\phi(k)^2} \right\} A_{k+1} + \frac{1}{\hat{\rho}\phi(k)} A_k. \quad (18)$$

**Proof** Observe that

$$\begin{aligned} \|E^{k+1}\|^2 &= \|\min\{-g(x^{k+1}), \frac{1}{v_k} \mu^k\}\|^2 \\ &\leq 2 \|\max\{g(x^{k+1}), -\frac{1}{v_k} \mu^k\} + \frac{1}{v_k} \mu^k\|^2 + 2 \|\frac{1}{v_k} \mu^k\|^2 \\ &= \frac{2}{v_k^2} \|\mu^{k+1}\|^2 + \frac{2}{v_k^2} \|\mu^k\|^2. \end{aligned} \quad (19)$$

It follows from (19) and Theorem 2(i) that

$$\begin{aligned} \frac{1}{2} \|E^{k+1}\|^2 &\leq \frac{1}{v_k^2} (\|\mu^{k+1}\|^2 + \|\mu^k\|^2) \\ &\leq \frac{v_{k+1}}{v_k^2} A_{k+1} + \frac{1}{v_k} A_k \end{aligned}$$

$$\leq \max \left\{ \frac{\xi_2}{\nu_k}, \frac{\hat{\nu}\phi(k+1)}{\nu_k^2} \right\} A_{k+1} + \frac{1}{\nu_k} A_k,$$

where the last inequality follows from the fact that, regardless of the condition in Step 6 of **FS-P-ALM**,  $\nu_{k+1} \leq \max\{\xi_2 \nu_k, \hat{\nu}\phi(k+1)\}$  since  $\xi_2 \geq 1$ . For any  $k \in \mathcal{K}_\nu$  it holds that  $\nu_k \geq \hat{\nu}\phi(k)$ . The claimed inequality (17) follows immediately by using this lower bound for  $\nu_k$  in the above inequality.

The second claim follows similarly, this time noting that

$$\begin{aligned} \|h(x^{k+1})\|^2 &\leq 2\|h(x^{k+1}) + \frac{1}{\rho_k} \lambda^k\|^2 + 2\|\frac{1}{\rho_k} \lambda^k\|^2 \\ &= \frac{2}{\rho_k^2} \|\lambda^{k+1}\|^2 + \frac{2}{\rho_k^2} \|\lambda^k\|^2 \end{aligned}$$

and using the bound  $\rho_k \geq \hat{\rho}\phi(k)$ .  $\square$

We next obtain the following result which is used to establish that the sequence of iterates of **FS-P-ALM** is such that the feasibility and complementarity residuals tend to zero along an infinite subsequence.

**Theorem 3** (Feasibility and complementarity residuals). *Suppose that Assumptions I and II hold, and let  $(x^k, \lambda^k, \mu^k)_{k \in \mathbb{N}}$  be the sequence of iterates generated by **FS-P-ALM**. Then, for any  $\varepsilon > 0$ , there exists an infinite subsequence  $\mathcal{K} \subseteq \mathbb{N}$  such that, for all  $k \in \mathcal{K}$ , it holds that  $\|h(x^k)\|_\infty \leq \varepsilon$  and  $\|E^k\|_\infty \leq \varepsilon$ .*

**Proof** For  $k \in \mathbb{N}$  define

$$t_\rho(k) := \max \{i \leq k \mid i \in \mathcal{K}_\rho\}, \quad t_\nu(k) := \max \{i \leq k \mid i \in \mathcal{K}_\nu\},$$

with the convention  $\max \emptyset := 0$ , and let

$$\mathcal{S}_\rho := \{j \in \mathcal{K}_\rho \mid [j-1] \notin \mathcal{K}_\rho\}, \quad \mathcal{S}_\nu := \{j \in \mathcal{K}_\nu \mid [j-1] \notin \mathcal{K}_\nu\}$$

denote the indices at which a run of consecutive increases begins. Since  $\rho_k$  is nondecreasing and equals  $\max\{\xi_1 \rho_{k-1}, \hat{\rho}\phi(k)\} \geq \hat{\rho}\phi(k)$  at every index at which it increases, and remains constant otherwise,

$$\rho_k \geq \hat{\rho}\phi(t_\rho(k)) \quad \text{whenever } \mathcal{K}_\rho \cap [0, k] \neq \emptyset, \quad (20)$$

and analogously  $\nu_k \geq \hat{\nu}\phi(t_\nu(k))$  whenever  $\mathcal{K}_\nu \cap [0, k] \neq \emptyset$ . We proceed with intermediate claims.

*Claim 1:*  $\|h(x^j)\| \rightarrow 0$  as  $\mathbb{N} \setminus \mathcal{S}_\rho \ni j \rightarrow \infty$ , and  $\|E^j\| \rightarrow 0$  as  $\mathbb{N} \setminus \mathcal{S}_\nu \ni j \rightarrow \infty$ . Let  $j \notin \mathcal{S}_\rho$ . If  $j-1 \in \mathcal{K}_\rho$ , then Lemma 3(ii) applied at  $j-1$  together with Assumption II gives  $\|h(x^j)\| \rightarrow 0$ . Otherwise  $j-1 \notin \mathcal{K}_\rho$ , and since  $j \notin \mathcal{S}_\rho$  this forces  $j \notin \mathcal{K}_\rho$ ; setting  $t := t_\rho(j) < j$ , the indices  $t+1, \dots, j$  all lie outside  $\mathcal{K}_\rho$ , so that  $j-t-1$  applications of (16) yield

$$\|h(x^j)\|_\infty \leq \beta \|h(x^{j-1})\|_\infty \leq \dots \leq \beta^{j-t-1} \|h(x^{t+1})\|_\infty.$$

Two cases arise. If  $\mathcal{K}_\rho$  is infinite, then  $t = t_\rho(j) \rightarrow \infty$  as  $j \rightarrow \infty$ , so that  $\|h(x^{t+1})\| \rightarrow 0$  by Lemma 3(ii) applied at  $t \in \mathcal{K}_\rho$ , while the factor  $\beta^{j-t-1}$  is bounded by 1 since  $\beta \in (0, 1)$  and  $j \geq t + 1$ . If instead  $\mathcal{K}_\rho$  is finite, then  $t_\rho(j)$  is constant and equal to  $\bar{t} := \max \mathcal{K}_\rho$  for every  $j > \bar{t}$ , so that  $\|h(x^{\bar{t}+1})\|_\infty$  is a fixed constant while the factor  $\beta^{j-\bar{t}-1}$  tends to zero as  $j \rightarrow \infty$ . In either case the right-hand side tends to zero. (If  $\mathcal{K}_\rho = \emptyset$ , the same argument applies with  $\bar{t} = 0$ .) Thus,  $\|h(x^j)\| \rightarrow 0$ . The argument for  $E^j$  is identical.

*Claim 2: if  $\mathbb{N} \setminus (\mathcal{S}_\rho \cup \mathcal{S}_v)$  is infinite, the assertion of the theorem holds.*

This set is contained in both  $\mathbb{N} \setminus \mathcal{S}_\rho$  and  $\mathbb{N} \setminus \mathcal{S}_v$ , so by Claim 1 both  $\|h(x^j)\|$  and  $\|E^j\|$  tend to zero along it; any  $\varepsilon > 0$  is thus attained on an infinite subset of it.

*Claim 3: if  $\mathbb{N} \setminus (\mathcal{S}_\rho \cup \mathcal{S}_v)$  is finite, the assertion of the theorem holds.*

Suppose first that  $\mathcal{S}_\rho \cap \mathcal{S}_v$  is infinite. Then Lemmas 3(i) and 3(ii) both apply at every  $j \in \mathcal{S}_\rho \cap \mathcal{S}_v \subseteq \mathcal{K}_\rho \cap \mathcal{K}_v$ , and by Assumption II, the assertion holds along the infinite set  $\{j + 1\} [j \in \mathcal{S}_\rho \cap \mathcal{S}_v]$ .

Suppose instead that  $\mathcal{S}_\rho \cap \mathcal{S}_v$  is finite. By definition neither  $\mathcal{S}_\rho$  nor  $\mathcal{S}_v$  contains two consecutive integers. Let  $J \in \mathbb{N}$  be such that  $\mathcal{S}_\rho \cup \mathcal{S}_v \supseteq \{j \in \mathbb{N}\} [j \geq J]$ . Enlarging  $J$  if necessary, every  $j \geq J$  then belongs to exactly one of  $\mathcal{S}_\rho$  and  $\mathcal{S}_v$ , and since neither set contains two consecutive integers, membership alternates. Consequently, for every  $j \geq J + 1$ , one of the two consecutive indices  $j - 1, j$  belongs to  $\mathcal{K}_\rho$  and the other to  $\mathcal{K}_v$ , so that

$$t_\rho(j) \geq j - 1 \quad \text{and} \quad t_v(j) \geq j - 1.$$

By Assumption II there exists  $M > 1$  with  $\phi(k + 1) < M\phi(k)$  for all  $k$  large. Hence  $\phi(j - 1) > \phi(j)/M$ , and since  $t_\rho(j) \in \{j - 1, j\}$  we obtain  $\phi(t_\rho(j)) \geq \phi(j)/M$  in either case. The same holds for  $t_v(j)$ . Combining with (20),

$$\rho_j \geq \frac{\hat{\rho}}{M}\phi(j), \quad v_j \geq \frac{\hat{v}}{M}\phi(j),$$

repeating the derivation of (18) with  $\rho_j \geq \frac{\hat{\rho}}{M}\phi(j)$  in place of  $\rho_j \geq \hat{\rho}\phi(j)$ , together with  $\phi(j + 1) < M\phi(j)$ , gives

$$\frac{1}{2}\|h(x^{j+1})\|^2 \leq \frac{M}{\hat{\rho}} \max \left\{ \frac{\xi_1}{\phi(j)}, \frac{M^2}{\phi(j)} \right\} A_{j+1} + \frac{M}{\hat{\rho}\phi(j)} A_j = \frac{M \max\{\xi_1, M^2\}}{\hat{\rho}} \frac{A_{j+1}}{\phi(j)} + \frac{M}{\hat{\rho}} \frac{A_j}{\phi(j)},$$

and analogously for  $\frac{1}{2}\|E^{j+1}\|^2$ , repeating the derivation of (17) with  $v_j \geq \frac{\hat{v}}{M}\phi(j)$  in place of  $v_j \geq \hat{v}\phi(j)$ . Since  $A_k = O(k)$  and  $k/\phi(k) \rightarrow 0$  by Assumption II, both right-hand sides tend to zero, so both residuals tend to zero at the common index  $j + 1$  for every large  $j$ .

Claims 2 and 3 are exhaustive, which concludes the proof.  $\square$

Combining the above results, we can conclude that **FS-P-ALM** finds an approximate KKT point after a finite number of iterations. To formalize the proof statement about this convergence property (i.e., as stated in Theorem 1), we state the following instrumental result about the quantity  $E^k$  defined in (8) for quantifying inequality constraint violations and complementarity. The proof is given in Appendix B for completeness.

---

**Algorithm 2** **FS-ALM** ((Inexact) feasible-start ALM)

---

**REQUIRE** Initial feasible point  $x^0 \in \text{dom } f_2$  such that  $h(x^0) = 0$ ,  $g(x^0) \leq 0$   
Multipliers  $\lambda^0 \in \mathbb{R}^p$  and  $\mu^0 \in \mathbb{R}_+^m$  and penalties  $\rho_0, v_0 \in \mathbb{R}_{++}$   
Parameters  $\beta \in (0, 1)$ ,  $\xi_1, \xi_2 \geq 1$ ,  $\hat{\rho}, \hat{v} \in \mathbb{R}_+$   
Sequence of tolerances  $(\tau_k)_{k \in \mathbb{N}} \subset \mathbb{R}_{++}$  with  $\tau_k \rightarrow \tau$  for some  $\tau \in \mathbb{R}_+$   
Function  $\phi$  satisfying Assumption II (e.g.,  $\phi(k) = (k+1)^\alpha$  for some  $\alpha > 1$ )

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

- 1: Set  $\hat{x}^k = x^k$  if  $\mathcal{L}_{\rho_k, v_k}(x^k, \lambda^k, \mu^k) + f_2(x^k) \leq f(x^0)$ , or  $\hat{x}^k = x^0$  otherwise
- 2: Starting at  $\hat{x}^k$ , find a point  $x^{k+1} \in \text{dom } f_2$  approximately solving

$$\text{minimize } \left\{ \mathcal{L}_{\rho_k, v_k}(\cdot, \lambda^k, \mu^k) + f_2(\cdot) \right\},$$

such that

$$\|x^{k+1} - \text{prox}_{f_2}(x^{k+1} - \nabla_x \mathcal{L}_{\rho_k, v_k}(x^{k+1}, \lambda^k, \mu^k))\|_\infty \leq \tau_k$$

- 3: Set  $\lambda^{k+1} = \lambda^k + \rho_k h(x^{k+1})$
  - 4: Set  $\mu^{k+1} = \max\{0, \mu^k + v_k g(x^{k+1})\}$
  - 5: Set  $\rho_{k+1} = \rho_k$  if  $\|h(x^{k+1})\|_\infty \leq \beta \|h(x^k)\|_\infty$ ,  
or  $\rho_{k+1} = \max\{\xi_1 \rho_k, \hat{\rho} \phi(k+1)\}$  otherwise
  - 6: Set  $v_{k+1} = v_k$  if  $\|E^{k+1}\|_\infty \leq \beta \|E^k\|_\infty$ , where  $E^k = \min\left\{-g(x^k), \frac{1}{v_{k-1}} \mu^{k-1}\right\}$   
or  $v_{k+1} = \max\{\xi_2 v_k, \hat{v} \phi(k+1)\}$  otherwise
- 

**Lemma 4** ([20, Lemma 3.1]). *Suppose that Assumption 1.A<sub>3</sub> holds, and let  $E^k$  be defined as in (8). Then, for all  $k \in \mathbb{N}$  and any  $\varepsilon \in \mathbb{R}_+$ ,  $\|E^{k+1}\|_\infty \leq \varepsilon$  implies that  $\|\max\{0, g(x^{k+1})\}\|_\infty \leq \varepsilon$  and  $\mu_i^{k+1} = 0$  for all  $i \in [m]$  whenever  $g_i(x^{k+1}) < -\varepsilon$ .*

We are now ready to combine all the above results to establish that the algorithm finds an approximate KKT point in finite number of iterations.

**Proof of Theorem 1** ( $\varepsilon$ -KKT termination for **FS-P-ALM**). Theorem 1(i) follows directly from Theorem 2(ii), and Theorem 1(ii) follows from (10) together with Theorem 3. As shown in Lemma 2(ii), it holds that for  $k$  large enough  $\|x^k - \text{prox}_{f_2}(x^k - \nabla_x \mathcal{L}(x^k, \lambda^k, \mu^k))\|_\infty \leq \varepsilon$ . Moreover, by Theorem 3, there exists a subsequence along which both  $\|h(x^k)\|_\infty \leq \varepsilon$  and  $\|E^k\|_\infty \leq \varepsilon$  hold. Now, note that by Lemma 4 and Definition 1, these imply that  $(x^k, \lambda^k, \mu^k)$  is an  $\varepsilon$ -KKT point of problem (1). Consequently, we conclude that an  $\varepsilon$ -KKT point of (1) is obtained after a finite number of iterations.  $\square$

## 5 An inexact ALM as a limiting case

The results obtained for **FS-P-ALM** in the previous section extend to its ALM counterpart which does not involve a proximal term. In this section, we present **FS-ALM** (Algorithm 2), an ALM version of **FS-P-ALM**, and briefly discuss how some results on **FS-P-ALM**, particularly Theorem 1, hold also for **FS-ALM** as special

cases. For ease of our presentation, we define

$$\begin{aligned} \mathcal{L}_{\rho,v}(x, \lambda, \mu) := & f_1(x) + \langle \lambda, h(x) \rangle + \frac{\rho}{2} \|h(x)\|^2 \\ & + \frac{1}{2v} \|\max\{0, vg(x) + \mu\}\|^2 - \frac{1}{2v} \|\mu\|^2, \end{aligned} \quad (21)$$

which is the augmented Lagrangian for problem (1) without the proximal term.

As a first instance, the following is analogous, for **FS-ALM**, to the feasibility property (11) (via a similar argument):

$$\mathcal{L}_{\rho_k, v_k}(x^{k+1}, \lambda^k, \mu^k) + f_2(x^{k+1}) \leq f(x^0), \quad (22)$$

for every  $k \in \mathbb{N}$ . This property of **FS-ALM** helps to initialize subproblems properly, as well as establish the following result, whose analogue for **FS-P-ALM** appears in the proof of Theorem 2.

**Lemma 5** (Multipliers control for **FS-ALM**). *Let Assumption I hold and suppose that the sequence  $(x^k, \lambda^k, \mu^k)_{k \in \mathbb{N}}$  is generated by **FS-ALM**. Then, for any  $k \in \mathbb{N}$ :*

$$\frac{1}{2\rho_{k+1}} \|\lambda^{k+1}\|^2 + \frac{1}{2v_{k+1}} \|\mu^{k+1}\|^2 \leq c + \frac{1}{2\rho_k} \|\lambda^k\|^2 + \frac{1}{2v_k} \|\mu^k\|^2, \quad (23)$$

where  $c := f(x^0) - f^*$ . In particular, the following holds for any  $K \in \mathbb{N}$ :

$$\frac{1}{\rho_K} \|\lambda^K\|^2 + \frac{1}{v_K} \|\mu^K\|^2 \leq 2Kc + \frac{1}{\rho_0} \|\lambda^0\|^2 + \frac{1}{v_0} \|\mu^0\|^2.$$

**Proof** See Appendix B. □

Similar to **FS-P-ALM**, the inequalities (22) and (23) are central to the proof of the successful termination of **FS-ALM**. Observe that, in the absence of the proximal term (which is the precise case for **FS-ALM**), one gets from Lemma 2 that

$$\|x^{k+1} - \text{prox}_{f_2}(x^{k+1} - \nabla_x \mathcal{L}(x^{k+1}, \lambda^{k+1}, \mu^{k+1}))\|_\infty \leq \varepsilon$$

holds for all  $k$  large enough and any  $\varepsilon > \tau \in \mathbb{R}_+$ , where  $\tau := \lim_k \tau_k$ . The proof follows similarly as the proof of Theorem 2, with the difference that now the primal residual term  $\frac{1}{\gamma_k}(x^{k+1} - x^k)$  is absent. The same observation is made in the proof of Theorem 3 from which we can establish for **FS-ALM** that  $\|h(x^{k+1})\|^2 \rightarrow 0$  and  $\|E^{k+1}\|^2 \rightarrow 0$  along some infinite subsequence  $\mathcal{K} \subseteq \mathbb{N}$ . Bringing these together leads to the following corollary to Theorem 1 which states that **FS-ALM** terminates successfully at an  $\varepsilon$ -KKT point of problem (1).

**Corollary 1** ( $\varepsilon$ -KKT termination for **FS-ALM**). *Let  $(x^k, \lambda^k, \mu^k)_{k \in \mathbb{N}}$  be the sequence of iterates generated by **FS-ALM**, and let Assumptions I and II hold. Then, an  $\varepsilon$ -KKT point of (1) is obtained after finitely many iterations, for any  $\varepsilon \in \mathbb{R}_{++}$ .*

**Table 1** Algorithms used for comparison

| Algorithm                           | Description                                                                                                                                                                |
|-------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>FS-P-ALM-<math>\alpha</math></b> | <i>Algorithm 1</i> with $\xi_1 = \xi_2 = 1$ and $\phi(k) = (k+1)^\alpha$ , where $\alpha > 1$                                                                              |
| <b>FS-ALM-<math>\alpha</math></b>   | <i>Algorithm 2</i> with $\xi_1 = \xi_2 = 1$ and $\phi(k) = (k+1)^\alpha$ , where $\alpha > 1$                                                                              |
| ALM- $\xi$                          | Standard ALM corresponding to $\xi_1 = \xi_2 = \xi > 1$ and $\phi(k) = 0$ in <i>Algorithm 2</i> , without the initial feasibility condition and the check in <i>Step 1</i> |

*Note.* Unless stated otherwise, we use  $\alpha \in \{1.01, 2, 4\}$  throughout (Sect. 6.1). ALM- $\xi$  takes  $\phi \equiv 0$  and therefore falls outside Assumption II; it is included as a standard-practice baseline and is not covered by our analysis

## 6 Numerical experiments

In this section, we present numerical experiments to demonstrate the performance of our proposed schemes. A Python implementation is available online.<sup>1</sup>

### 6.1 Algorithmic setup

The algorithms considered for comparison are summarized in Table 1. Every algorithm is run from the same initial point and with the same inner tolerance sequence  $(\tau_k)_{k \in \mathbb{N}}$ .

#### Inner solvers

Each experiment is run with two inner solvers and is reported for both. The first is a proximal-gradient method, in which the stepsize is selected by a standard backtracking linesearch on the smooth part, warm started at 1.1 times the last accepted stepsize. This inner solver satisfies the hypothesis of Lemma 1 by construction. We allow it maximum 5000 inner iterations. The second is PANOC [57] from the software Alpaqa [58] with L-BFGS directions, setting a memory size of 20 and a maximum iteration number of 1000. Its linesearch is performed on the forward-backward envelope, so the descent property required by Lemma 1 is not formally guaranteed at every inner iteration. We did not observe any adverse effect from this in our experiments.

#### Choice of the initial parameters

In all our tests, we use  $\beta = 0.5$ ,  $\tau_k = \max\{\tau_0 \kappa^k, \varepsilon\}$  with  $\tau_0 = 0.1$ ,  $\kappa = 0.5$  and  $\varepsilon$  the requested accuracy, as in Algorithm 1, and  $\phi_\gamma(k) = (k+1)^{1.01}$  as suggested in Assumption II, unless otherwise stated. For simplicity, we set  $\hat{\rho} = \rho_0$ ,  $\hat{\nu} = \nu_0$ , and  $\hat{\gamma} = \gamma_0$ . The initial multipliers are set to  $\lambda^0 = \mu^0 = 0$ .

<sup>1</sup> <https://github.com/adeyemiadeoye/fs-palm>.

For the initial penalty parameters we adopt the following simple rule of thumb,

$$\rho_0 = \frac{\kappa_\rho}{\phi(K_r + 1)} \frac{\|\nabla^2 f_1(x^0)\|}{\|J_h(x^0)\|^2}, \quad v_0 = \frac{\kappa_\rho}{\phi(K_r + 1)} \frac{\|\nabla^2 f_1(x^0)\|}{\|J_g(x^0)\|^2}, \quad (24)$$

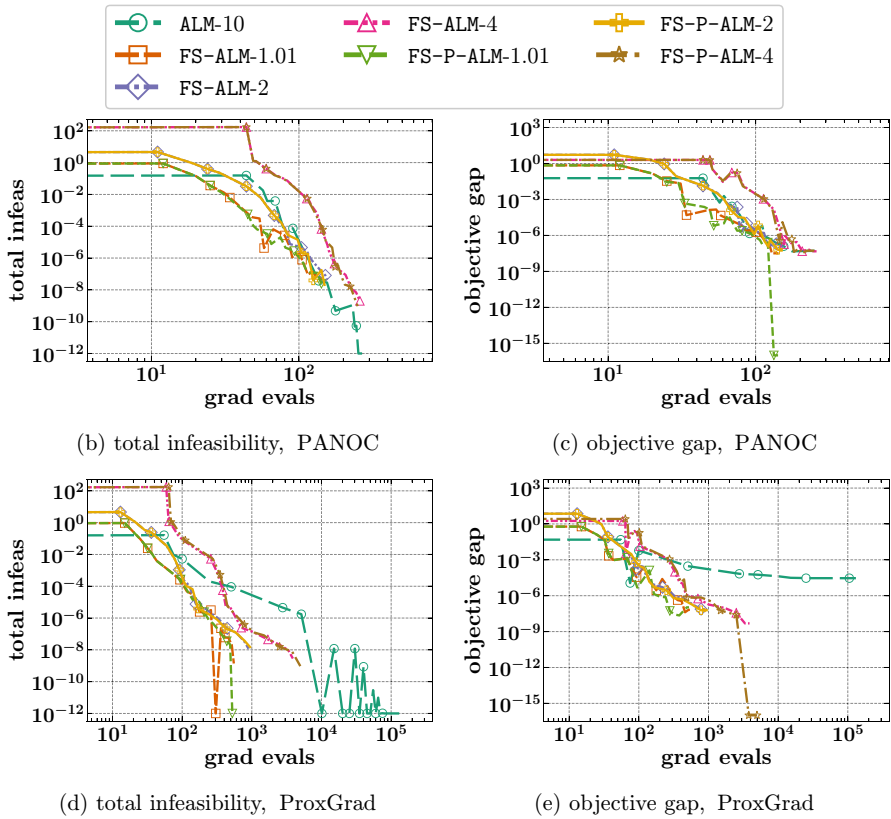
with  $\kappa_\rho = 10^2$  and  $K_r = 5$ . Considering the potential growth according to  $\phi(k + 1)$ , each penalty modulus reaches a reasonable reference level after  $K_r$  outer iterations for every  $\alpha$ . The dimensionless factor  $\kappa_\rho$  is unchanged throughout. All norms are obtained by power iteration on Hessian- and Jacobian-vector products, without forming either matrix. The motivation for this initialization is most direct for  $\rho_0$ . Since Step 1 takes  $x^0$  feasible, so that  $h(x^0) = 0$ , the term  $\frac{\rho}{2}\|h(x)\|^2$  contributes exactly  $\rho J_h^\top J_h$  to the Hessian of the augmented Lagrangian at  $x^0$ , and the corresponding ratio in (24) is the value of  $\rho$  at which the equality constraints are felt as strongly as the curvature of the objective. We use an expression of the same type for  $v_0$ , with  $J_g$  in place of  $J_h$ . The same reasoning does not carry over, nevertheless, taking the full Jacobian regardless errs towards a smaller initial penalty, which the growth schedule then recovers.

For the proximal parameters we set  $\gamma_0 = \kappa_\gamma / \hat{L}$  with  $\kappa_\gamma = 10^3$  and  $\hat{L}$  the spectral norm of the Hessian of the augmented Lagrangian at  $x^0$ , and  $\delta = 1 / \max\{1, |f_1(x^0)|\}$ . Since  $\frac{1}{2\gamma}\|x - \hat{x}\|^2$  adds  $1/\gamma$  to the curvature seen by the subproblem solver,  $\kappa_\gamma$  is the inverse fraction of the local curvature contributed by the proximal term.

## Initial feasibility and stopping criteria

The algorithms are stopped when the  $\varepsilon$ -KKT residual of Definition 1 falls below  $\varepsilon = 10^{-6}$ , that is, when  $\max\{\|R_\gamma(x^k)\|_\infty, \|h(x^k)\|_\infty, \|E^k\|_\infty\} \leq 10^{-6}$ , where  $R_\gamma$  denotes the fixed-point residual at unit stepsize. The first term is the primal stationarity measure and is included in the test, so that all three conditions of Definition 1 are certified rather than feasibility and complementarity alone. In our tests, we track the total infeasibility measure and the suboptimality gap given, respectively, by  $\|h(x^k)\|_\infty + \|\max\{0, g(x^k)\}\|_\infty$  and  $\frac{|f(x^k) - f_{\text{ref}}|}{|f(x^0) - f_{\text{ref}}|}$ , where  $f_{\text{ref}}$  is a reference value for the optimal value of the problem at hand, specified for each problem below. For problems where an initial feasible point is not readily available, we solve a phase I problem (see Appendix A). Note that, to ensure a fair comparison, we use this initial point for all algorithms (including ALM, even though the algorithm does not require a feasible initial point). In our experiments, this is only needed for the CUTEst instances of Sect. 6.3, since for all other problems a feasible point is known by construction. The phase I problem itself always has a trivially feasible point, so **FS-P-ALM** is used to solve it. We refer to Appendix A for the details and for the limitations of this approach.

In all figures, cost is measured by the cumulative number of evaluations of the gradient of the smooth part of the subproblem objective, including all evaluations performed by the inner solver.



**Fig. 2** Sparse nonconvex QCQP with  $n = 50$ ,  $m = 10$ , under both inner solvers, for  $\alpha \in \{1.01, 2, 4\}$  (and  $\xi = 10$  for the ALM variant). Under PANOC every run converges. Under ProxGrad **FS-P-ALM** and **FS-ALM** still converge at every  $\alpha$ , whereas ALM does not

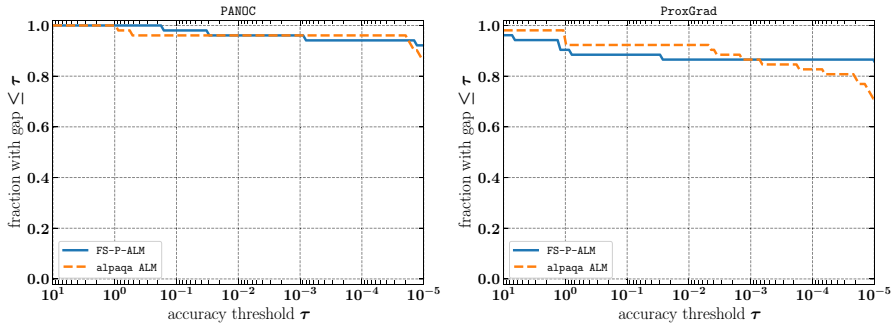
## 6.2 Sparse nonconvex QCQP

We next consider a sparse quadratically constrained quadratic program (QCQP) of the form

$$\begin{aligned} & \underset{x \in \mathbb{R}^n}{\text{minimize}} && \frac{1}{2}x^\top P_0 x + q_0^\top x + \eta \|x\|_1 \\ & \text{subject to} && \frac{1}{2}x^\top P_i x + q_i^\top x - r_i \leq 0, \quad i = 1, \dots, m, \end{aligned} \quad (25)$$

where  $P_i \in \mathbb{R}^{n \times n}$  are symmetric,  $q_i \in \mathbb{R}^n$ , and  $r_i \in \mathbb{R}$ . No definiteness is assumed on the  $P_i$ , so the problem is in general nonconvex. The  $\ell_1$  penalty parameter is set to  $\eta = \frac{1}{2}\|q_0\|_\infty$ . Problem (25) is an instance of (1) with  $f_1(x) = \frac{1}{2}x^\top P_0 x + q_0^\top x$ ,  $f_2(x) = \eta\|x\|_1$ ,  $h \equiv 0$ , and  $g(x) = (\frac{1}{2}x^\top P_i x + q_i^\top x - r_i)_{i=1}^m \in \mathbb{R}^m$ .

Following the random-instance conventions of [59], the entries of  $q_0, \dots, q_m$  are drawn i.i.d. from  $\mathcal{N}(0, 1)$ . For the objective we form  $S = \frac{1}{2\sqrt{n}}(B + B^\top)$  with  $B$  having i.i.d.  $\mathcal{N}(0, 1)$  entries and set  $P_0 = V \text{diag}(\sigma(w/\|w\|_\infty + \theta))V^\top$ , where  $S = V \text{diag}(w)V^\top$  is the spectral decomposition. The shift  $\theta$  and the scale  $\sigma$  are drawn per



**Fig. 3** Accuracy attained by the point each method returns, against the threshold  $\tau$ , on the 53 CUTEst QPs. Under PANOC, **FS-P-ALM** and alpaga satisfy their respective stopping tests on 46 and 47 instances; under ProxGrad, on 36 and 34

instance so that both the curvature of  $f_1$  and the proportion of negative eigenvalues of  $P_0$  vary across the test set,  $P_0$  remaining indefinite throughout. For the constraints we take  $P_i = \sigma_i \frac{1}{n} A_i^\top A_i \succeq 0$  with  $A_i \in \mathbb{R}^{n \times n}$  having i.i.d.  $\mathcal{N}(0, 1)$  entries and  $\sigma_i$  an instance-dependent scale, and  $r_i > 0$ . Each constraint then describes an ellipsoid, their intersection is bounded, and  $x^0 = 0$  is strictly feasible for every instance, which supplies the feasible initialization required by **FS-P-ALM** without recourse to the phase I problem. Since the problem is nonconvex, its global optimal value is not available. The reference value  $f_{\text{ref}}$  is therefore the smallest value of  $f$  attained at the final points of all runs with the same inner solver in Fig. 2 and at a local solution computed by IPOPT [60], applied to the standard smooth reformulation of the  $\ell_1$  term by means of auxiliary variables. A sample instance is shown in Fig. 2. Under PANOC the three methods converge comparably, while under ProxGrad only **FS-P-ALM** and **FS-ALM** reach the tolerance and ALM stalls.

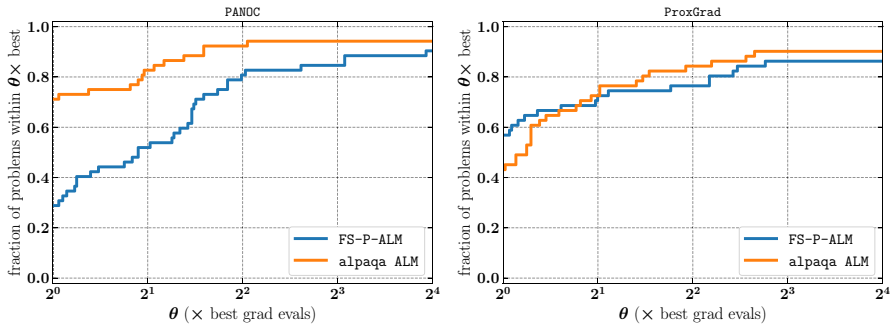
### 6.3 Quadratic programs

We illustrate the proposed algorithms on quadratic programming (QP) problems in the form (1) with  $f_1(x) = \frac{1}{2}x^\top Qx + q^\top x + c$ , where  $Q \in \mathbb{R}^{n \times n}$  is symmetric and not necessarily positive semidefinite,  $q \in \mathbb{R}^n$ ,  $c \in \mathbb{R}$ ;  $f_2(x) = \delta_\Omega(x)$ , where  $\Omega = \{x \in \mathbb{R}^n : x_l \leq x \leq x_u\}$ ; and  $g(x) = [Ax - u, l - Ax]$ , where  $A \in \mathbb{R}^{m \times n}$ .

We report two sets of experiments, one on instances taken from an established public collection and one on randomly generated instances whose conditioning we control.

#### CUTEst instances

The first set is drawn from the CUTEst collection [61] and comprises 53 instances. Of these, 44 are convex, from the Maros-Mészáros repository [52] and related CUTEst QP families, while the remaining nine are the NCVXQP instances, whose Hessians are indefinite. The dimensions range from  $n = 2$  to  $n = 745$  and from  $m = 1$  to  $m = 1001$ . The upper end of this range is a limitation of our implementation



**Fig. 4** Performance profile in gradient evaluations on the same set, in the sense of [62]. The ordering reverses with the inner solver. Under PANOC, whose quasi-Newton directions absorb subproblem ill-conditioning, alpaga is cheaper on 35 of the 48 instances both solve. Under ProxGrad, where that conditioning is felt directly, **FS-P-ALM** is cheaper on 26 of 42

rather than of the method itself, since it assembles Hessians and Jacobians densely for automatic differentiation and therefore cannot reach the largest sparse instances of the collection.

On this set we compare **FS-P-ALM** against the augmented Lagrangian solver of alpaga [58], which safeguards the multipliers by projection onto a bounded set and increases the penalty multiplicatively. Both methods drive the same PANOC inner solver with identical settings and run in the same process. We use  $\alpha = 4$  and  $\alpha_\gamma = 4$ , a common stopping tolerance of  $10^{-5}$  for both methods, and a limit of 500 outer iterations. The tolerance enters only through where each method chooses to stop, and hence which point it returns.

Figure 3 reports, against the accuracy threshold  $\tau$ , the fraction of instances on which the point a method *returns* attains  $\max\{\text{violation}, |f - f_{\text{ref}}|/\max\{1, |f_{\text{ref}}|\}\} \leq \tau$ , where  $f_{\text{ref}}$  is the best value either method reached at a feasible point. In this figure we report the attained accuracy rather than the cost of reaching a prescribed target because the two methods do not stop on the same test. Neither method dominates over the whole range, and the ordering depends both on the threshold and on the inner solver. Under PANOC, **FS-P-ALM** satisfies its stopping test on 46 instances and alpaga on 47; under ProxGrad, the corresponding counts are 36 and 34. Among the remaining runs, **FS-P-ALM** reaches the outer-iteration limit while returning a point feasible to  $10^{-5}$  on 4 instances under PANOC and 11 under ProxGrad. The corresponding counts for alpaga are 5 and 17. These runs are feasible according to the common post-processing measure, but do not satisfy the method's complete internal stopping test and hence are not certified as converged. **FS-P-ALM** returns a genuinely infeasible point on 1 instance under PANOC and 5 under ProxGrad; the corresponding counts for alpaga are 1 and 2. On QPCBOEI1, and under ProxGrad also on QPCBOEI2, phase I does not produce a feasible starting point, so **FS-P-ALM** is not run.

Figure 4 plots the performance profile [62] in gradient evaluations on the same dataset. Cost is the number of evaluations at which a method first reaches one common target, feasible to  $10^{-5}$  and within 10% of the best objective either method found.

Thus, a higher curve means that the method reaches the common target within a smaller multiple of the best recorded cost on more jointly successful problems. Under PANOC, `alpaqa` has the smaller cost on 35 of the 48 jointly successful instances; under ProxGrad, **FS-P-ALM** has the smaller cost on 26 of 42. This confirms the overall observation that Panoc directions absorb the ill-conditioning of the subproblems, while ProxGrad does not.

## Randomly generated instances

The second set consists of randomly generated convex QPs, which allows the conditioning of the subproblems to be varied in a controlled way. The instances are generated with a prescribed condition number  $\kappa(Q)$ . Concretely,  $Q = V \text{diag}(w) V^\top$  with  $V$  a random orthogonal matrix and  $w$  log-spaced over  $[1, \kappa(Q)]$ .  $q$  has i.i.d. standard Gaussian entries and  $c = 0$ . The linear constraints  $A$  are placed around a strictly feasible interior point, with a quarter of the rows tightened to near-activity there so the constraints are actually exercised rather than redundant, and the box is fixed at  $[-10, 10]^n$ . The same interior point serves as  $x^0$ , so no phase I step is needed. Since these instances are convex, the reference value  $f_{\text{ref}}$  is their optimal value, which we compute with the interior-point solver `Clarabel` [63] through `CVXPY` [64]. The results are displayed in Fig. 5, for  $\alpha \in \{1.01, 2, 4\}$  in accordance with Table 1, with the initial parameters set as in Sect. 6.1. Under PANOC all runs converge. Under ProxGrad, where the subproblem conditioning is felt directly, ALM does not reach the tolerance while **FS-P-ALM** and **FS-ALM** do.

## 6.4 Basis pursuit (nonconvex formulation)

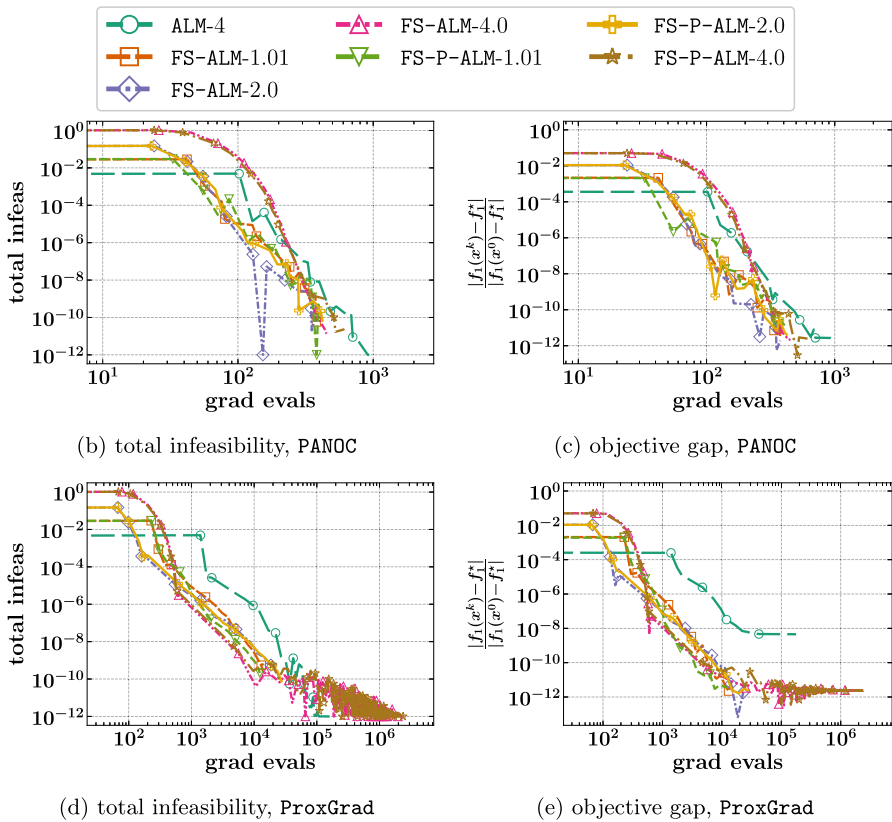
We consider the basis pursuit problem which involves finding the sparsest solution to an undetermined linear system of equations. That is,

$$\underset{z \in \mathbb{R}^n}{\text{minimize}} \quad \|z\|_1 \quad \text{subject to} \quad Bz = b, \quad (26)$$

where  $B \in \mathbb{R}^{p \times n}$ , with  $p < n$ , and  $b \in \mathbb{R}^p$ . We consider a nonconvex reformulation of (26) [35]:

$$\underset{x \in \mathbb{R}^{2n}}{\text{minimize}} \quad \|x\|^2 \quad \text{subject to} \quad \bar{B}x^{\circ 2} = b,$$

where  $x := (x_1, x_2) \in \mathbb{R}^{2n}$ ,  $\bar{B} := [B, -B] \in \mathbb{R}^{p \times 2n}$ , and  $\circ$  is used indicate element-wise power. Letting  $z^+, z^-$  denote the positive and negative parts of  $z$  satisfying  $z = z^+ - z^-$ , it holds that  $x_1^{\circ 2} = z^+$ ,  $x_2^{\circ 2} = z^-$ . The entries of  $B$  are independent draws from the standard normal distribution  $\mathcal{N}(0, 1)$ . We then set  $b = Bz^*$ , where  $z^*$  is also a random vector with i.i.d. entries from  $\mathcal{N}(0, 1)$  having  $k$  nonzero values fixed at 10. The proximal parameters are set as in Sect. 6.1. The reference value is  $f_{\text{ref}} = \|z^*\|_1$ , which is the optimal value of both (26) and its reformulation. That  $z^*$  solves (26) is verified by solving its linear programming reformulation with `HiGHS` [65].

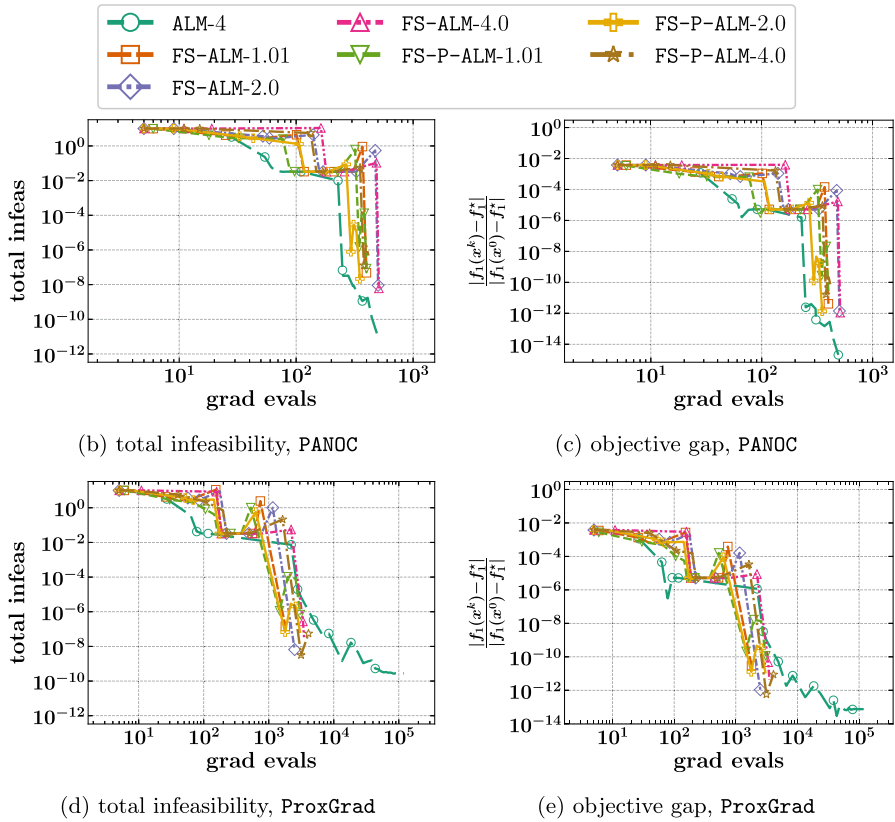


**Fig. 5** Convex QP with  $n = 60, m = 30$  and  $\kappa(Q) = 10^2$ , under both inner solvers, for  $\alpha \in \{1.01, 2, 4\}$  (and  $\xi = 4$  for the ALM variant). Under PANOC every run converges. Under ProxGrad, where the subproblem conditioning is felt directly, ALM breaks down while **FS-P-ALM** and **FS-ALM** still converge

Figure 6 compares the algorithms for  $\alpha \in \{1.01, 2, 4\}$  under both inner solvers. Under PANOC the three methods lie within a small factor of one another, while under ProxGrad **FS-P-ALM** and **FS-ALM** converge at every  $\alpha$  while ALM does not.

Both **FS-P-ALM** and **FS-ALM** provide advantages over ALM on these problems, avoiding the ill-conditioning issue. The size of the advantage depends on how the subproblems are solved and on how the initial penalty is chosen. Under a quasi-Newton inner solver, and with each method given a well-scaled initial penalty, the three lie within a small factor of one another. The separation appears when the subproblems are solved by a first-order method, or when the initial penalty is misspecified. See also Appendix C for the effect of  $\alpha$  given a fixed initialization of the penalty parameter.

Figure 7 shows the penalty trajectories on a basis-pursuit instance, labelled by the number of times the  $\beta$ -progress test fires. That test, rather than an inexact subproblem solve, is what triggers a penalty update, and it fires at comparable rates for all three methods. ALM multiplies by  $\xi$  on each firing, so firings compound geometrically,



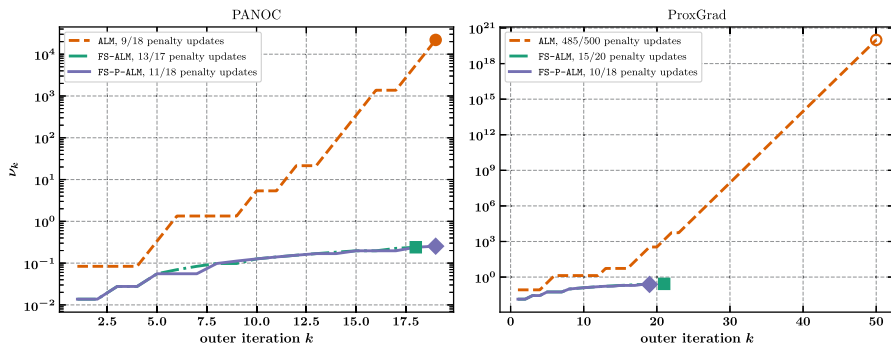
**Fig. 6** Basis pursuit with  $p = 400$ ,  $n = 1024$  and  $k = 10$ , under both inner solvers, for  $\alpha \in \{1.01, 2, 4\}$  and  $\xi = 4$ . Under PANOC all runs converge and lie within a small factor of one another. Under ProxGrad **FS-P-ALM** and **FS-ALM** still converge at every  $\alpha$ , whereas ALM does not

whereas our update reduces to  $\rho \leftarrow \max\{\rho, \hat{\rho}\phi(k+1)\}$  and is bounded by the schedule.

These results demonstrate the promising practical performance of our proposed approach on nonconvex problems, highlighting their potential to effectively solve challenging problems in many application domains.

## 7 Conclusions

We proposed an inexact proximal augmented Lagrangian method that enables aggressive and adaptive penalty updates for solving nonconvex structured optimization problems. Our theoretical analysis is based on properties of the iterates that stem from the initial feasibility assumption, establishing that the method attains an approximate Karush–Kuhn–Tucker point, up to a prescribed accuracy, in finitely many iterations. The analysis further extends to the limiting case without the proximal term, thereby



**Fig. 7** Penalty  $\rho_k$  against outer iteration on a basis-pursuit instance, under each inner solver, labelled by the number of  $\beta$ -test firings. Filled end markers denote convergence within the set runtime budget; otherwise, the end marker is unfilled (curves are drawn over the range in which the trajectory remains numerically meaningful.)

encompassing the classical augmented Lagrangian method as a special case. Numerical experiments on both convex and nonconvex problems of varying complexity highlight the efficiency of the proposed approach.

Future research directions include extending the method to incorporate positive definite matrices in place of scalar proximal stepsizes, together with appropriate update rules for such a matrix. Another important avenue is the development of iteration complexity estimates for the proposed framework.

## Appendix A: The phase I problem for initialization

When an initial *feasible* point is not readily available, a so-called phase I problem can be solved to find one for **FS-P-ALM**. We consider the following problem:

$$\underset{(x,s) \in \mathbb{R}^{n+1}}{\text{minimize}} \quad \frac{1}{2} \|h(x)\|^2 + s^2 + \delta_{\text{dom } f_2}(x) \quad \text{subject to} \quad g(x) \leq s\mathbf{1}. \quad (\text{A1})$$

Here,  $\mathbf{1} \in \mathbb{R}^m$  is the vector of all ones, and  $\text{dom } f_2$  is assumed to be closed. For any  $x^0 \in \text{dom } f_2$ , the pair  $(x^0, s^0)$  with  $s^0 = \max\{g_1(x^0), \dots, g_m(x^0)\}$  is feasible for (A1), so that **FS-P-ALM** can be applied to it starting from this point. Moreover, (1) is feasible if and only if (A1) admits a point  $(x, s)$  with zero objective value, in which case  $s = 0$  and  $x$  is feasible for (1). However, since (A1) is nonconvex in general, **FS-P-ALM** is only guaranteed to return an  $\varepsilon$ -KKT point of it, which need not be feasible for (1). The outcome of phase I thus has to be checked. In our experiments, a feasible initial point is available by construction for the QCQP of Sect. 6.2, the randomly generated quadratic programs of Sect. 6.3, and the basis pursuit problem of Sect. 6.4. Phase I is only needed for the CUTEst instances of Sect. 6.3, where it does not produce a feasible point for QPCBOEI1, and under ProxGrad also for QPCBOEI2, so that **FS-P-ALM** is not run on these instances.

## Appendix B: Proofs for Sects. 4 and 5

**Proof of Lemma 4** ([20, Lemma 3.1]). By the definition of  $E^k$  in (8) and  $\|E^{k+1}\|_\infty \leq \varepsilon$ , we have that if  $E_i^{k+1} = \frac{1}{v_k} \mu_i^k$ , then  $g_i(x^{k+1}) \leq 0 \leq \varepsilon$ , since  $\frac{1}{v_k} \mu_i^k \in \mathbb{R}_+$ . Likewise,  $|E_i^{k+1}| \leq \varepsilon$  implies that  $g_i(x^{k+1}) \leq \varepsilon$  whenever  $E_i^{k+1} = -g_i(x^{k+1})$ . Hence,  $g_i(x^{k+1}) \leq \varepsilon$  for all  $i \in [m]$  and  $\|\max\{0, g(x^{k+1})\}\|_\infty \leq \varepsilon$ . Now, if  $g_i(x^{k+1}) < -\varepsilon$ , then by  $|E_i^{k+1}| \leq \varepsilon$ , we must have  $E_i^{k+1} = \frac{1}{v_k} \mu_i^k \leq \varepsilon$ . These give  $g_i(x^{k+1}) + \frac{1}{v_k} \mu_i^k < -\varepsilon + \varepsilon = 0$ . By the definition of  $\mu_i^{k+1}$  in Step 4, we have  $\mu_i^{k+1} = 0$ .  $\square$

**Proof of Lemma 5** (Multipliers control for **FS-ALM**). Using Steps 3 and 4 of **FS-ALM** in the definition (21) of  $\mathcal{L}_{\rho, v}$ , we get

$$\begin{aligned} \frac{1}{2\rho_k} \|\lambda^{k+1}\|^2 + \frac{1}{2v_k} \|\mu^{k+1}\|^2 &= \frac{\rho_k}{2} \|h(x^{k+1}) + \frac{\lambda^k}{\rho_k}\|^2 + \frac{1}{2v_k} \|\max\{0, v_k g(x^{k+1}) + \mu^k\}\|^2 \\ &= \mathcal{L}_{\rho_k, v_k}(x^{k+1}, \lambda^k, \mu^k) - f_1(x^{k+1}) + \frac{1}{2\rho_k} \|\lambda^k\|^2 + \frac{1}{2v_k} \|\mu^k\|^2 \\ &\leq f(x^0) - f_2(x^{k+1}) - f_1(x^{k+1}) + \frac{1}{2\rho_k} \|\lambda^k\|^2 + \frac{1}{2v_k} \|\mu^k\|^2 \\ &\leq f(x^0) - f^* + \frac{1}{2\rho_k} \|\lambda^k\|^2 + \frac{1}{2v_k} \|\mu^k\|^2, \end{aligned}$$

where the first inequality follows from (22) and the second one from Assumption I.A<sub>4</sub>. Since  $\rho_{k+1} \geq \rho_k$  and  $v_{k+1} \geq v_k$ , the left-hand side of (23) is bounded above by the left-hand side of the above chain, which proves (23). Summing (23) over  $k = 0, \dots, K-1$  and multiplying by 2 yields the last claim.  $\square$

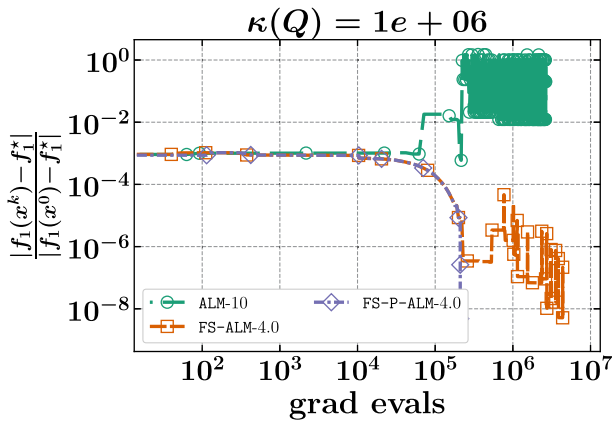
## Appendix C: Additional numerical results

### Effect of the proximal term

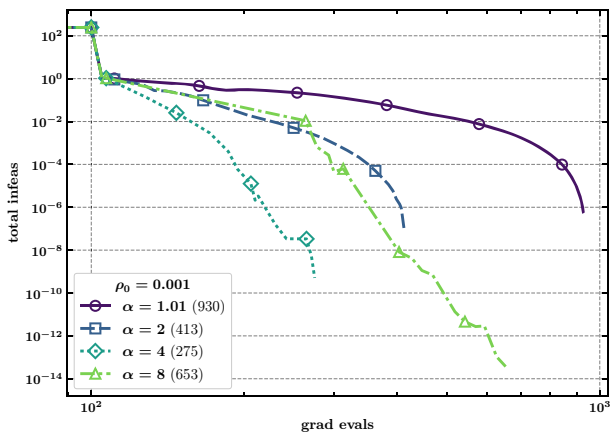
The proximal term acts on the conditioning of the subproblems, so its effect is visible on ill-conditioned instances rather than uniformly. Figure 8 isolates it on DUALC1, the worst conditioned instance of the set, with  $\kappa(Q) \approx 10^6$ , under PANOC and at a common  $\alpha = 4$ . ALM breaks down. **FS-ALM** descends but then oscillates without settling, and does not meet the tolerance within the budget. **FS-P-ALM** descends monotonically and terminates cleanly. This is the sense in which the proximal term contributes numerically, namely robustness on ill-conditioned subproblems rather than a uniform gain in cost. On the better-conditioned instances of Fig. 10 the two are essentially indistinguishable.

### Effect of $\alpha$ under a fixed initial penalty

Figure 9 holds  $\rho_0$  fixed at  $10^{-3}$  and sweeps  $\alpha$ . In this configuration the schedule is the only mechanism that can raise the penalty to a useful magnitude, so a larger exponent is needed and the cost is lowest at an intermediate  $\alpha$  rather than at the smallest one.



**Fig. 8** Relative suboptimality against gradient evaluations on DUALC1 under PANOC, for ALM, **FS-ALM** and **FS-P-ALM** at a common growth exponent. Only **FS-P-ALM** reaches the requested tolerance

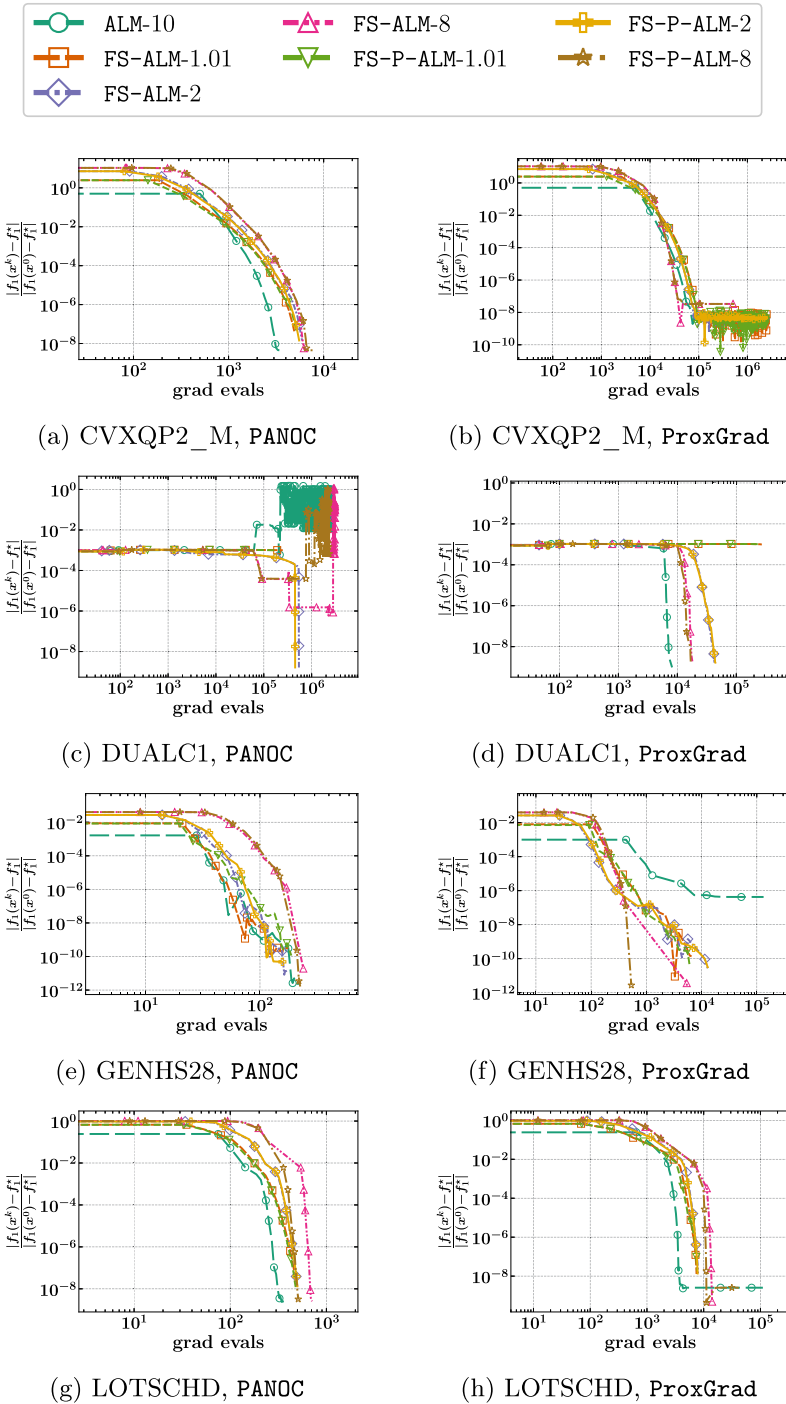


**Fig. 9** Sparse nonconvex QCQP under PANOC with  $\rho_0$  fixed at  $10^{-3}$  and  $\alpha$  swept, **FS-P-ALM** throughout; gradient evaluations to termination in the legend

This behavior with respect to the choice of  $\alpha$  is ultimately dictated by the initialization of the penalty parameters.

### Effect of $\alpha$ on CUTEst instances

The figures below report four CUTEst instances for a wider range of  $\alpha$ . We compare  $\alpha \in \{1.01, 2, 8\}$  against ALM with  $\xi = 10$ , with all initial parameters set as in Sect. 6.1, so that  $\rho_0$  follows (24) throughout. On the better conditioned instances every variant converges and the exponent has little effect, whereas on DUALC1 the runs separate and several fail to settle within the budget under either inner solver.



**Fig. 10** Four CUTEst QP instances, one per row, under both inner solvers, for  $\alpha \in \{1.01, 2, 8\}$  (and  $\xi = 10$  for the ALM variant). Each panel plots the relative suboptimality against gradient evaluations

**Acknowledgements** The authors thank Pieter Pas (KU Leuven) for informed discussions on the PANOC solver.

**Funding** Open access funding provided by Scuola IMT Alti Studi Lucca within the CRUI-CARE Agreement. This work was funded by the European Union (ERC Advanced Research Grant COMPACT, No. 101141351). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

**Data availability** Some numerical experiments were carried out on a subset of the Maros-Mészáros quadratic programming test set, which is publicly available at <https://www.doc.ic.ac.uk/~im/> and [https://old.sztaki.hu/~meszaros/public\\_ftp/qpdata/](https://old.sztaki.hu/~meszaros/public_ftp/qpdata/) (see also [52]). A Python implementation of our approach, together with all supporting scripts, is publicly available on GitHub at <https://github.com/adeyemiadeoye/fs-palm>.

## Declarations

**Conflict of interest** The authors have no conflict of interest to declare.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

1. Mayne, D.Q., Rawlings, J.B., Rao, C.V., Scokaert, P.O.: Constrained model predictive control: stability and optimality. *Automatica* **36**(6), 789–814 (2000). [https://doi.org/10.1016/S0005-1098\(99\)00214-9](https://doi.org/10.1016/S0005-1098(99)00214-9)
2. Hastie, T., Tibshirani, R., Friedman, J.H.: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd edn. Springer, New York (2009). <https://doi.org/10.1007/978-0-387-84858-7>
3. Boyd, S.P., Vandenberghe, L.: *Convex Optimization*. Cambridge University Press (2004). <https://doi.org/10.1017/CBO9780511804441>
4. Hestenes, M.R.: Multiplier and gradient methods. *J. Optim. Theory Appl.* **4**(5), 303–320 (1969). <https://doi.org/10.1007/BF00927673>
5. Powell, M.J.D.: A method for nonlinear constraints in minimization problems. In: Fletcher, R. (ed.) *Optimization*, pp. 283–298. Academic Press, London (1969)
6. Rockafellar, R.T.: Augmented Lagrangians and applications of the proximal point algorithm in convex programming. *Math. Oper. Res.* **1**(2), 97–116 (1976). <https://doi.org/10.1287/moor.1.2.97>
7. Rockafellar, R.T.: Monotone operators and the proximal point algorithm. *SIAM J. Control. Optim.* **14**(5), 877–898 (1976). <https://doi.org/10.1137/0314056>
8. Shefi, R., Teboulle, M.: Rate of convergence analysis of decomposition methods based on the proximal method of multipliers for convex minimization. *SIAM J. Optim.* **24**(1), 269–297 (2014). <https://doi.org/10.1137/130910774>
9. Curtis, F.E., Jiang, H., Robinson, D.P.: An adaptive augmented Lagrangian method for large-scale constrained optimization. *Math. Program.* **152**(1), 201–245 (2015). <https://doi.org/10.1007/s10107-014-0784-y>
10. Douglas, J., Rachford, H.H.: On the numerical solution of heat conduction problems in two and three space variables. *Trans. Am. Math. Soc.* **82**(2), 421–439 (1956). <https://doi.org/10.1090/S0002-9947-1956-0084194-4>
11. Glowinski, R., Le Tallec, P.: *Augmented Lagrangian and Operator-splitting Methods in Nonlinear Mechanics*. Society for Industrial and Applied Mathematics, Philadelphia (1989). <https://doi.org/10.1137/1.9781611970838>

12. Tseng, P.: Applications of a splitting algorithm to decomposition in convex programming and variational inequalities. *SIAM J. Control. Optim.* **29**(1), 119–138 (1991). <https://doi.org/10.1137/0329006>
13. Eckstein, J., Bertsekas, D.P.: On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Math. Program.* **55**(1–3), 293–318 (1992). <https://doi.org/10.1007/BF01581204>
14. Chen, G., Teboulle, M.: A proximal-based decomposition method for convex minimization problems. *Math. Program.* **64**(1), 81–101 (1994). <https://doi.org/10.1007/BF01582566>
15. Chambolle, A., Pock, T.: A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vision* **40**, 120–145 (2011). <https://doi.org/10.1007/s10851-010-0251-1>
16. Bertsekas, D.P.: *Constrained Optimization and Lagrange Multiplier Methods*. Athena Scientific Optimization and Computation Series. Athena Scientific, Belmont (1996)
17. Rockafellar, R.T.: Monotone operators associated with saddle-functions and minimax problems. In: Browder, F.E. (ed.) *Nonlinear Functional Analysis, Part I. Proceedings of Symposia in Pure Mathematics*, vol. 18, pp. 241–250. American Mathematical Society, Providence, RI (1970). <https://doi.org/10.1090/pspum/018.1/0285942>
18. Eckstein, J., Silva, P.J.: A practical relative error criterion for augmented Lagrangians. *Math. Program.* **141**(1), 319–348 (2013). <https://doi.org/10.1007/s10107-012-0528-9>
19. Birgin, E.G., Haeser, G., Ramos, A.: Augmented Lagrangians with constrained subproblems and convergence to second-order stationary points. *Comput. Optim. Appl.* **69**(1), 51–75 (2018). <https://doi.org/10.1007/s10589-017-9937-2>
20. Birgin, E.G., Martínez, J.M.: Complexity and performance of an augmented Lagrangian algorithm. *Optim. Methods Softw.* **35**(5), 885–920 (2020). <https://doi.org/10.1080/10556788.2020.1746962>
21. Grapiglia, G.N., Yuan, Y.-X.: On the complexity of an augmented Lagrangian method for nonconvex optimization. *IMA J. Numer. Anal.* **41**(2), 1546–1568 (2021). <https://doi.org/10.1093/imanum/draa021>
22. Evens, B., Latafat, P., Themelis, A., Suykens, J., Patrinos, P.: Neural network training as an optimal control problem: an augmented Lagrangian approach. In: 2021 60th IEEE Conference on Decision and Control (CDC), pp. 5136–5143 (2021). <https://doi.org/10.1109/CDC45484.2021.9682842>
23. Birgin, E.G., Martínez, J.M.: *Practical Augmented Lagrangian Methods for Constrained Optimization*. Society for Industrial and Applied Mathematics, Philadelphia (2014). <https://doi.org/10.1137/1.9781611973365>
24. De Marchi, A., Jia, X., Kanzow, C., Mehlitz, P.: Constrained composite optimization and augmented Lagrangian methods. *Math. Program.* **201**(1), 863–896 (2023). <https://doi.org/10.1007/s10107-022-01922-4>
25. Jia, X., Kanzow, C., Mehlitz, P., Wachsmuth, G.: An augmented Lagrangian method for optimization problems with structured geometric constraints. *Math. Program.* **199**(1), 1365–1415 (2023). <https://doi.org/10.1007/s10107-022-01870-z>
26. Bertsekas, D.P.: Combined primal-dual and penalty methods for constrained minimization. *SIAM J. Control* **13**(3), 521–544 (1975). <https://doi.org/10.1137/0313030>
27. Bertsekas, D.P.: Multiplier methods: a survey. *Automatica* **12**(2), 133–145 (1976). [https://doi.org/10.1016/0005-1098\(76\)90077-7](https://doi.org/10.1016/0005-1098(76)90077-7)
28. Chen, X., Guo, L., Lu, Z., Ye, J.J.: An augmented Lagrangian method for non-Lipschitz nonconvex programming. *SIAM J. Numer. Anal.* **55**(1), 168–193 (2017). <https://doi.org/10.1137/15M1052834>
29. Lan, G., Monteiro, R.D.: Iteration-complexity of first-order augmented Lagrangian methods for convex programming. *Math. Program.* **155**(1), 511–547 (2016). <https://doi.org/10.1007/s10107-015-0861-x>
30. Patrascu, A., Necoara, I., Tran-Dinh, Q.: Adaptive inexact fast augmented Lagrangian methods for constrained convex optimization. *Optim. Lett.* **11**(3), 609–626 (2017). <https://doi.org/10.1007/s11590-016-1024-6>
31. Necoara, I., Patrascu, A., Glineur, F.: Complexity of first-order inexact Lagrangian and penalty methods for conic convex programming. *Optim. Methods Softw.* **34**(2), 305–335 (2019). <https://doi.org/10.1080/10556788.2017.1380642>
32. Li, F., Qu, Z.: An inexact proximal augmented Lagrangian framework with arbitrary linearly convergent inner solver for composite convex optimization. *Math. Program. Comput.* **13**(3), 583–644 (2021). <https://doi.org/10.1007/s12532-021-00205-x>
33. Xu, Y.: Iteration complexity of inexact augmented Lagrangian methods for constrained convex programming. *Math. Program.* **185**(1), 199–244 (2021). <https://doi.org/10.1007/s10107-019-01425-9>
34. Lu, Z., Zhou, Z.: Iteration-complexity of first-order augmented Lagrangian methods for convex conic programming. *SIAM J. Optim.* **33**(2), 1159–1190 (2023). <https://doi.org/10.1137/21M1403837>

35. Sahin, M.F., Eftekhari, A., Alacaoglu, A., Latorre, F., Cevher, V.: An inexact augmented Lagrangian framework for nonconvex optimization with nonlinear constraints. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates Inc, Red Hook (2019)
36. Li, Z., Chen, P.-Y., Liu, S., Lu, S., Xu, Y.: Rate-improved inexact augmented Lagrangian method for constrained nonconvex optimization. In: Banerjee, A., Fukumizu, K. (eds.) *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research*, vol. 130, pp. 2170–2178 (2021)
37. Bodard, A., Oikonomidis, K., Laude, E., Patrinos, P.: The inexact power augmented Lagrangian method for constrained nonconvex optimization (2024). <https://doi.org/10.48550/arXiv.2410.20153>. arXiv preprint [arXiv:2410.20153](https://arxiv.org/abs/2410.20153)
38. Liu, W., Lin, Q., Xu, Y.: Lower complexity bounds of first-order methods for affinely constrained composite nonconvex problems. *Math. Oper. Res.* **51**(2), 1659–1682 (2026). <https://doi.org/10.1287/moor.2023.0377>
39. Conn, A.R., Gould, N.I., Toint, P.: A globally convergent augmented Lagrangian algorithm for optimization with general constraints and simple bounds. *SIAM J. Numer. Anal.* **28**(2), 545–572 (1991). <https://doi.org/10.1137/0728030>
40. Conn, A.R., Gould, N.I.M., Sartenar, A., Toint, P.L.: Convergence properties of an augmented Lagrangian algorithm for optimization with a combination of general equality and linear constraints. *SIAM J. Optim.* **6**(3), 674–703 (1996). <https://doi.org/10.1137/S1052623493251463>
41. Andreani, R., Birgin, E.G., Martínez, J.M., Schuverdt, M.L.: On augmented Lagrangian methods with general lower-level constraints. *SIAM J. Optim.* **18**(4), 1286–1309 (2008). <https://doi.org/10.1137/060654797>
42. Zhang, J., Luo, Z.-Q.: A proximal alternating direction method of multiplier for linearly constrained nonconvex minimization. *SIAM J. Optim.* **30**(3), 2272–2302 (2020). <https://doi.org/10.1137/19M1242276>
43. Zhang, J., Luo, Z.-Q.: A global dual error bound and its application to the analysis of linearly constrained nonconvex optimization. *SIAM J. Optim.* **32**(3), 2319–2346 (2022). <https://doi.org/10.1137/20M135474X>
44. Xie, Y., Wright, S.J.: Complexity of proximal augmented Lagrangian for nonconvex optimization with nonlinear equality constraints. *J. Sci. Comput.* **86**(3), 38 (2021). <https://doi.org/10.1007/s10915-021-01409-y>
45. Hong, M., Hajinezhad, D., Zhao, M.-M.: Prox-PDA: The proximal primal-dual algorithm for fast distributed nonconvex optimization and learning over networks. In: Precup, D., Teh, Y.W. (eds.) *Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research*, vol. 70, pp. 1529–1538 (2017)
46. Jiang, B., Lin, T., Ma, S., Zhang, S.: Structured nonconvex and nonsmooth optimization: algorithms and iteration complexity analysis. *Comput. Optim. Appl.* **72**(1), 115–157 (2019). <https://doi.org/10.1007/s10589-018-0034-y>
47. Hajinezhad, D., Hong, M.: Perturbed proximal primal-dual algorithm for nonconvex nonsmooth optimization. *Math. Program.* **176**(1), 207–245 (2019). <https://doi.org/10.1007/s10107-019-01365-4>
48. Boţ, R.I., Nguyen, D.-K.: The proximal alternating direction method of multipliers in the nonconvex setting: convergence analysis and rates. *Math. Oper. Res.* **45**(2), 682–712 (2020). <https://doi.org/10.1287/moor.2019.1008>
49. Sujanani, A., Monteiro, R.D.: An adaptive superfast inexact proximal augmented Lagrangian method for smooth nonconvex composite optimization problems. *J. Sci. Comput.* **97**(2), 34 (2023). <https://doi.org/10.1007/s10915-023-02350-y>
50. Kong, W., Melo, J.G., Monteiro, R.D.: Iteration complexity of an inner accelerated inexact proximal augmented Lagrangian method based on the classical Lagrangian function. *SIAM J. Optim.* **33**(1), 181–210 (2023). <https://doi.org/10.1137/20M136147X>
51. Hermans, B., Themelis, A., Patrinos, P.: QPALM: a proximal augmented Lagrangian method for nonconvex quadratic programs. *Math. Program. Comput.* **14**(3), 497–541 (2022). <https://doi.org/10.1007/s12532-022-00218-0>
52. Maros, I., Mészáros, C.: A repository of convex quadratic programming problems. *Optim. Methods Softw.* **11**(1–4), 671–681 (1999). <https://doi.org/10.1080/10556789908805768>
53. Chen, S.S., Donoho, D.L., Saunders, M.A.: Atomic decomposition by basis pursuit. *SIAM Rev.* **43**(1), 129–159 (2001). <https://doi.org/10.1137/S003614450037906X>

54. Andreani, R., Martínez, J.M., Svaiter, B.F.: A new sequential optimality condition for constrained optimization and algorithmic consequences. *SIAM J. Optim.* **20**(6), 3533–3554 (2010). <https://doi.org/10.1137/090777189>
55. Andreani, R., Martínez, J.M., Ramos, A., Silva, P.J.S.: A cone-continuity constraint qualification and algorithmic consequences. *SIAM J. Optim.* **26**(1), 96–110 (2016). <https://doi.org/10.1137/15M1008488>
56. Conn, A.R., Gould, N.I.M., Toint, P.L.: *Trust Region Methods*. Society for Industrial and Applied Mathematics, Philadelphia (2000). <https://doi.org/10.1137/1.9780898719857>
57. Stella, L., Themelis, A., Sotasakis, P., Patrinos, P.: A simple and efficient algorithm for nonlinear model predictive control. In: *IEEE 56th Annual Conference on Decision and Control (CDC)*, pp. 1939–1944 (2017). <https://doi.org/10.1109/CDC.2017.8263933>
58. Pas, P., Schuurmans, M., Patrinos, P.: Alpaqa: a matrix-free solver for nonlinear MPC and large-scale nonconvex optimization. In: *2022 European Control Conference (ECC)*, pp. 417–422 (2022). <https://doi.org/10.23919/ECC55457.2022.9838172>. IEEE
59. Park, J., Boyd, S.: General heuristics for nonconvex quadratically constrained quadratic programming (2017). <https://doi.org/10.48550/arXiv.1703.07870>. arXiv preprint [arXiv:1703.07870](https://arxiv.org/abs/1703.07870)
60. Wächter, A., Biegler, L.T.: On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Math. Program.* **106**(1), 25–57 (2006). <https://doi.org/10.1007/s10107-004-0559-y>
61. Gould, N.I.M., Orban, D., Toint, P.L.: CUTEst: a constrained and unconstrained testing environment with safe threads for mathematical optimization. *Comput. Optim. Appl.* **60**(3), 545–557 (2015). <https://doi.org/10.1007/s10589-014-9687-3>
62. Dolan, E.D., Moré, J.J.: Benchmarking optimization software with performance profiles. *Math. Program.* **91**(2), 201–213 (2002). <https://doi.org/10.1007/s101070100263>
63. Goulart, P.J., Chen, Y.: Clarabel: an interior-point solver for conic programs with quadratic objectives. *Program. Comput. Math.* (2026). <https://doi.org/10.1007/s12532-026-00320-7>
64. Diamond, S., Boyd, S.: CVXPY: a python-embedded modeling language for convex optimization. *J. Mach. Learn. Res.* **17**(83), 1–5 (2016)
65. Huangfu, Q., Hall, J.A.J.: Parallelizing the dual revised simplex method. *Math. Program. Comput.* **10**(1), 119–142 (2018). <https://doi.org/10.1007/s12532-017-0130-5>