



On constrained optimization with nonconvex regularization

E. G. Birgin¹ · J. M. Martínez² · A. Ramos³

Received: 16 September 2019 / Accepted: 1 April 2020 / Published online: 19 April 2020
© Springer Science+Business Media, LLC, part of Springer Nature 2020

Abstract

In many engineering applications, it is necessary to minimize smooth functions plus penalty (or regularization) terms that violate smoothness and convexity. Specific algorithms for this type of problems are available in recent literature. Here, a smooth reformulation is analyzed and equivalence with the original problem is proved both from the points of view of global and local optimization. Moreover, for the cases in which the objective function is much more expensive than the constraints, model-intensive algorithms, accompanied by their convergence and complexity theories, are introduced. Finally, numerical experiments are presented.

Keywords Constrained non-Lipschitz nonsmooth optimization · Complexity analysis · Optimality conditions

✉ E. G. Birgin
egbirgin@ime.usp.br

J. M. Martínez
martinez@ime.unicamp.br

A. Ramos
albertoramos@ufpr.br

¹ Department of Computer Science, Institute of Mathematics and Statistics,
University of São Paulo, Rua do Matão, 1010, Cidade Universitária, 05508-090,
São Paulo, SP, Brazil

² Department of Applied Mathematics, Institute of Mathematics, Statistics, Scientific Computing,
State University of Campinas, R. Sérgio Buarque de Holanda, 651, Cidade Universitária,
13083-859, Campinas, SP, Brazil

³ Department of Mathematics, Federal University of Paraná, Centro Politécnico, Jd. das Américas
CP 19081, 81531-980, Curitiba, PR, Brazil

1 Introduction

We are concerned with mathematical problems in which the objective is to find a point $x \in X$ such that:

$$F(D(x), E(x)) = 0. \quad (1)$$

In general, X is a subset of a finite-dimensional space. Constrained and unconstrained minimization problems may be formulated in this way, with $D(x)$ and $E(x)$ representing objective functions and constraints. Problems of the form (1) are almost always solved by means of iterative methods. This means that, given an approximation x^k to a solution, the approximation x^{k+1} is obtained by solving:

$$M_{F,x^k}(M_{D,x^k}(x), M_{E,x^k}(x)) = 0, \quad (2)$$

where M_{F,x^k} , M_{D,x^k} , and M_{E,x^k} are models of F , D , and E built using knowledge available at x^k . For example, if we deal with the constrained optimization problem:

$$\text{Minimize } f(x) \text{ subject to } g(x) \leq 0, \quad (3)$$

we may identify, say, f with D , g with E and F with the difference between $f(x)$ and the minimum for each feasible point x , in such a way that problem (2) may consist on the minimization of the quadratic Taylor approximation of f (perhaps plus a Lagrangian term) subject to the linearization of the constraints. This is, essentially, the idea of sequential quadratic programming methods [29].

In many cases, the arguments of F involve a function that is difficult to evaluate ($D(x)$) and a function that is easy to evaluate ($E(x)$). In these cases, the computer work associated with the solution of (2) may be overwhelmingly dominated by the evaluation of $D(x^k)$ (including, possibly, its derivatives). Moreover, even

$$M_{F,x^k}(M_{D,x^k}(x), E(x)) = 0 \quad (4)$$

and

$$F(M_{D,x^k}(x), E(x)) = 0 \quad (5)$$

may be solvable with a computational cost essentially equal to the cost of computing $D(x^k)$. For example, in the constrained optimization problem, the evaluation of the objective function could involve a huge collection of data whereas the constraints may be easy to evaluate. In this case, it is probably convenient to employ an iterative scheme based on (5) by means of which we solve, at each iteration, a powerful model of the objective function subject to the true constraints. This is the point of view of [26], where the complexity of a general scheme for constrained optimization based on high-order models of the objective function is analyzed.

The objective of this paper is to exploit the modeling idea (5) with respect to the minimization of nonconvexly regularized problems. The equivalence of this problem with a constraint optimization problem in which the objective function is generally much more expensive than the constraints will be exploited and it will be proved that, for some Taylor-like models with mild assumptions, the resulting method enjoys theoretical convergence and complexity properties that suggest a good computational behavior. Some illustrative examples will be shown that tend to corroborate this hypothesis.

Given $\mathcal{C} \subset \mathbb{R}^n$, continuously differentiable functions $f : \mathbb{R}^n \rightarrow \mathbb{R}$ and $g : \mathbb{R}^n \rightarrow \mathbb{R}^d$, a penalty function $\pi : \mathbb{R} \rightarrow \mathbb{R}_+$ such that $\pi(t) = 0$ if $t \leq 0$, and a regularization parameter $\sigma > 0$, we consider the constrained optimization problem by:

$$\begin{aligned} \text{Minimize} \quad & f(x) + \sigma \sum_{j=1}^d \pi(g_j(x)) \\ \text{subject to} \quad & x \in \mathcal{C}. \end{aligned} \quad (6)$$

Problem (6) may arise in the process of solving problem:

$$\text{Minimize } f(x) \text{ subject to } g(x) \leq 0, \quad x \in \mathcal{C} \quad (7)$$

by means of penalization with respect to the constraints $g(x) \leq 0$. In this case, we expect that a solution to (7) should emerge when σ is sufficiently large or, at least, when $\sigma \rightarrow \infty$ (see, for example, [29]). Many times, the objective function of (6) is smooth and $\pi'(0) = 0$. However, we are especially interested in the case in which $\pi'(0) > 0$ and, above all, in the case in which $\pi'(0)$ does not exist and $\lim_{t \rightarrow 0+} \pi'(t) = \infty$.

For several reasons, solutions to (6) for different values of σ may be meaningful independently of its connection with (7). In fact, many times, the solution to (6) when $\sigma \rightarrow \infty$ is not relevant and the fulfillment of all the constraints $g_j(x) \leq 0$ is not desirable or, perhaps, impossible. Instead, the ideal asymptotic problem that one wishes to solve is, roughly speaking:

$$\text{Minimize } f(x) \text{ subject to } x \in \mathcal{C} \quad (8)$$

with the compromise that the number of indices j such that $g_j(x) > 0$ should be small. For example, a typical portfolio problem could require the minimization of the expected loss $f(x)$ subject to standard constraints $x \in \mathcal{C}$ and the additional requirement that the number of scenarios under which shortfall occurs should be as small as possible [12]. We expect to achieve this objective employing penalty functions that are concave for $t \geq 0$ and such that $\lim_{t \rightarrow 0+} \pi'(t) > 0$.

The case in which \mathcal{C} is polyhedral and $\pi(t) = t^q$ if $t \geq 0$, with $q \in (0, 1)$, has been considered in [5, 24]. The problem:

$$\text{Minimize } f(x) + \sigma \sum_{j=1}^n |x_j|^q \text{ subject to } x \in \mathcal{C}, \quad (9)$$

called L_q -regularized optimization problem, has been the object of many studies starting from Tikhonov [30] and the enormous literature on inverse problems. The case $q = 2$ corresponds to the most classical regularization, intended to handle problems in which the mere minimization of $f(x)$ onto \mathcal{C} is very ill-conditioned and, so, its exact solution is, frequently, meaningless. The case $q = 1$ preserves the possible convexity of $f(x)$ and has been shown to induce sparse regularized minimizers of $f(x)$ (see [13]). With $q < 1$ possible convexity of the objective function is lost but the tendency to find sparse solutions x is even more emphatic. Writing $g_j(x) = x_j$ and $g_{n+j}(x) = -x_j$ for $j = 1, \dots, n$ and defining $\pi(t) = t^q$ for all $t \geq 0$ problem (9) takes the form (6); so (9) is a particular case of the

problem studied in this paper. Practical applications of problems of the form (6) include machine learning [18], information theory [20], image restoration [5], signal processing [4], variable selection [19], and others. Interesting examples were given in [24, Appx.1].

The rest of this paper is organized as follows. In Section 2, we introduce a smooth reformulation of (6) and we prove its equivalence with the original problem. In Section 3, we derive optimality conditions using the reformulation. In Section 4, we introduce a model-intensive algorithm for solving the problem and we prove its convergence and complexity. In Section 5, we introduce a specific algorithm for the case in which the constraints are linear. In Section 6, we present numerical experiments. Finally, in Section 7, we state conclusions and lines for future research.

Notation $\|\cdot\|$ denotes an arbitrary norm. \mathbb{R}_+ denotes the set of nonnegative elements of \mathbb{R} . The i th component of a vector v is denoted v_i or $[v]_i$. If $v \in \mathbb{R}^n$ is a vector with components v_i , v_+ is the vector with components $\max\{0, v_i\}$, $i = 1, \dots, n$.

2 Reformulation

A smooth reformulation of problem (6) will be introduced in this section. The penalty functions employed in this work will be required to satisfy the assumptions below.

Assumption A1 π is continuous, concave, strictly increasing for $t \geq 0$. Moreover, $\pi'(t)$ exists and is continuous for all $t > 0$.

Assumption A2 The restriction of π to \mathbb{R}_+ admits an inverse $\pi^{-1} : \mathbb{R}_+ \rightarrow \mathbb{R}_+$, the derivative $(\pi^{-1})'(t)$ exists for all $t > 0$, and the right derivative $(\pi^{-1})'(0)_+$ also exists.

Assumptions A1 and A2 are satisfied by penalty functions in practical applications such as ridge regression [21], smoothly clipped absolute deviation penalization (SCAD) [19], minimax concave penalties (MCP) [31], fraction penalty optimization [28], and log-penalty minimization [13]. Note that assumptions A1 and A2 do not require $\lim_{t \rightarrow 0+} \pi'(t) = \infty$, although this property will be interesting for the purpose of minimizing the number of indices j such that $g_j(x) > 0$.

The possible lack of smoothness of (6) is due to the nonsmoothness of $\pi(t)$ when $t = 0$. Therefore, standard methods based on gradient information for solving (6) may be inappropriate. In order to avoid this inconvenience, several approaches were proposed. Among them, we can mention smoothing approximation methods [5, 16, 24], iterative reweighted algorithms [17, 23, 25], and interior point methods [6, 22]. Our proposal, based on a smooth reformulation, has the advantage that standard well-established algorithms for constrained optimization based on first-order or higher-order derivative information on the derivatives may be used. This is possible thanks to the fact that the function $\pi^{-1} : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ is continuously differentiable.

Problem (6) is obviously equivalent to:

$$\text{Minimize } f(x) + \sigma \sum_{j=1}^d y_j \text{ subject to } y_j = \pi(g_j(x)) \text{ for } j = 1, \dots, d \text{ and } x \in \mathcal{C}. \quad (10)$$

If $y_j > \pi(g_j(x))$, taking $\bar{y}_j = \pi(g_j(x))$, we obtain that $f(x) + \sigma \sum_{j=1}^d \bar{y}_j \leq f(x) + \sigma \sum_{j=1}^d y_j$. Therefore, (10) is equivalent to:

$$\text{Minimize } f(x) + \sigma \sum_{j=1}^d y_j \text{ subject to } y_j \geq \pi(g_j(x)) \text{ for } j = 1, \dots, d \text{ and } x \in \mathcal{C}. \quad (11)$$

As $\pi(t) = 0$ if $t < 0$, (11) is equivalent to:

$$\text{Minimize } f(x) + \sigma \sum_{j=1}^d y_j \text{ subject to } y_j \geq 0, \quad y_j \geq \pi(g_j(x)) \text{ for } j = 1, \dots, d \quad (12)$$

and $x \in \mathcal{C}$.

If $y_j \geq 0$ and $y_j \geq \pi(g_j(x))$, since both y_j and $\pi(g_j(x))$ belong to \mathbb{R}_+ and the inverse of π onto \mathbb{R}_+ is increasing, we obtain $\pi^{-1}(y_j) \geq g_j(x)$. Reciprocally, if $\pi^{-1}(y_j) \geq g_j(x)$ and $y_j \geq 0$, we obtain $y_j \geq \pi(g_j(x))$ or $g_j(x) \leq 0$ in which case $y_j \geq \pi(g_j(x))$ also holds. Therefore, (12) is equivalent to:

$$\begin{aligned} \text{Minimize } \Phi(x, y) &:= f(x) + \sigma \sum_{j=1}^d y_j \\ \text{subject to } &x \in \mathcal{C}, \\ &g_j(x) - \pi^{-1}(y_j) \leq 0 \text{ for } j = 1, \dots, d, \\ &y \geq 0. \end{aligned} \quad (13)$$

The theorem below shows that a solution to the original problem (6) can be found by solving its smooth reformulation (13). (The reciprocal is also true, as shown in the following theorem.)

Theorem 2.1 *Suppose that (x^*, y^*) is a local (global) minimizer of (13). Then, $y_j^* = \pi(g_j(x^*))$ for all $j = 1, \dots, d$ and x^* is a local (global) minimizer of (6). Reciprocally, if x^* is a local (global) minimizer of (6) and $y_j^* = \pi(g_j(x^*))$ for all $i = 1, \dots, d$, then (x^*, y^*) is a local (global) minimizer of (13).*

Proof Suppose that (x^*, y^*) is a local minimizer of (13). If $y_j^* > \pi(g_j(x^*))$ and $y_j^* > y_j \geq \pi(g_j(x^*))$, we have $\pi^{-1}(y_j^*) > \pi^{-1}(y_j) \geq g_j(x^*)$. Therefore, if $y_i = y_i^*$ for all $i \neq j$, we have that (x^*, y) is a feasible point of (13) and $\Phi(x^*, y) < \Phi(x^*, y^*)$. Therefore, (x^*, y^*) would not be a local minimizer of (13). This implies that, necessarily, $y_j^* = \pi(g_j(x^*))$ for all $j = 1, \dots, d$.

Let $\varepsilon > 0$ be such that for all (x, y) such that $x \in \mathcal{C}$, $y \geq 0$, $g_j(x) \leq \pi^{-1}(y_j)$, $j = 1, \dots, d$, $\|x - x^*\| \leq \varepsilon$ and $\|y - y^*\|_\infty \leq \varepsilon$, we have that $f(x) + \sigma \sum_{j=1}^d y_j \geq f(x^*) + \sigma \sum_{j=1}^d y_j^*$. By the first part of the proof, $y_j^* = \pi(g_j(x^*))$; therefore, $f(x) + \sigma \sum_{j=1}^d y_j \geq f(x^*) + \sigma \sum_{j=1}^d \pi(g_j(x^*))$. Let $\varepsilon_1 \in (0, \varepsilon]$ be such that $\|x - x^*\| \leq \varepsilon_1$ implies that $|\pi(g_j(x)) - \pi(g_j(x^*))| \leq \varepsilon$ for all $j = 1, \dots, d$. Then, if $\|x - x^*\| \leq \varepsilon_1$, $f(x) + \sigma \sum_{j=1}^d \pi(g_j(x)) \geq f(x^*) + \sigma \sum_{j=1}^d \pi(g_j(x^*))$. This implies that x^* is a local minimizer of (6).

Reciprocally, suppose that x^* is a local minimizer of (6). Therefore, there exists $\varepsilon > 0$ such that for all $x \in \mathcal{C}$ such that $\|x - x^*\| \leq \varepsilon$, $f(x^*) + \sigma \sum_{j=1}^d \pi(g_j(x^*)) \leq f(x) + \sigma \sum_{j=1}^d \pi(g_j(x))$. Therefore, $f(x^*) + \sigma \sum_{j=1}^d y_j^* \leq f(x) + \sigma \sum_{j=1}^d \pi(g_j(x))$ for all $x \in \mathcal{C}$ such that $\|x - x^*\| \leq \varepsilon$. Thus, $f(x^*) + \sigma \sum_{j=1}^d y_j^* \leq f(x) + \sigma \sum_{j=1}^d y_j$ for all $x \in \mathcal{C}$ such that $\|x - x^*\| \leq \varepsilon$ if $y_j \geq 0$ and $y_j \geq \pi(g_j(x))$ for all $j = 1, \dots, d$. But, due to the increasing property of π^{-1} onto \mathbb{R}_+ , $y_j \geq 0$ and $y_j \geq \pi(g_j(x))$ is equivalent to $y_j \geq 0$ and $\pi^{-1}(y_j) \geq g_j(x)$. Therefore, $f(x^*) + \sigma \sum_{j=1}^d y_j^*$ is less than or equal to $f(x) + \sigma \sum_{j=1}^d y_j$ for all (x, y) satisfying $x \in \mathcal{C}$, $\|x - x^*\| \leq \varepsilon$, and the constraints of (13).

The equivalence proof with respect global minimizers is similar. \square

3 Optimality conditions

Assume that:

$$\mathcal{C} = \{x \in \mathbb{R}^n \mid H(x) = 0 \text{ and } G(x) \leq 0\}, \quad (14)$$

where $H : \mathbb{R}^n \rightarrow \mathbb{R}^{n_H}$ and $G : \mathbb{R}^n \rightarrow \mathbb{R}^{n_G}$ are continuously differentiable. If a local minimizer (x^*, y^*) of (13) satisfies a constraint qualification then the KKT conditions hold at (x^*, y^*) . By Theorem 2.1, if x^* is a local minimizer of (6), $y^* = \pi(g_j(x^*))$ for all $j = 1, \dots, d$, and a constraint qualification for (13) is fulfilled, the KKT conditions are satisfied too. The KKT conditions for (13) obviously involve both x^* and y^* . Since y is merely an auxiliary variable, it is interesting to derive KKT-like conditions in which only the primal variables x^* are involved. This is the objective of Theorem 3.1 below, which, in turn, generalizes [15, Corollary 2.2] and [24] using straightforward optimization arguments.

Theorem 3.1 *Let x^* be a local minimizer of (6) and assume that π satisfies Assumptions A1 and A2. Define:*

$$\begin{aligned} I(x^*) &:= \{j \mid g_j(x^*) < 0\}, \quad J(x^*) := \{j \mid g_j(x^*) = 0\}, \text{ and} \\ K(x^*) &:= \{j \mid g_j(x^*) > 0\}. \end{aligned} \quad (15)$$

Then, x^ is a local minimizer of the problem:*

$$\text{Minimize } f(x) + \sigma \sum_{j \in K(x^*)} \pi(g_j(x)) \text{ subject to } x \in \mathcal{C} \text{ and } g_j(x) \leq 0 \quad \forall j \in J(x^*). \quad (16)$$

Moreover, if \mathcal{C} is defined by $H(x) = 0$ and $G(x) \leq 0$, where $H : \mathbb{R}^n \rightarrow \mathbb{R}^{n_H}$ and $G : \mathbb{R}^n \rightarrow \mathbb{R}^{n_G}$ are continuously differentiable, and the system of equalities and inequalities:

$$H(x) = 0, \quad G(x) \leq 0, \quad g_j(x) \leq 0, \quad \forall j \in J(x^*) \quad (17)$$

satisfies a constraint qualification at x^* then we have that x^* satisfies the KKT conditions of problem (16).

Proof Let $y^* \geq 0$ be such that (x^*, y^*) is a local minimizer of (13). By Theorem 2.1, $y_j^* = \pi(g_j(x^*))$ for all $j = 1, \dots, d$. If $j \in I(x^*)$, we have that $g_j(x) < 0$ for all x in a neighborhood of x^* . Moreover, $y_j^* = 0$ in this case. Therefore, (x^*, y^*) is a local minimizer of:

$$\text{Minimize } f(x) + \sigma \sum_{j \notin I(x^*)} y_j \text{ subject to } x \in \mathcal{C}, \quad y_j \geq \pi(g_j(x)) \text{ and } y_j \geq 0 \quad \forall j \notin I(x^*).$$

So, (x^*, y^*) is a local minimizer of

$$\text{Minimize } f(x) + \sigma \sum_{j \notin I(x^*)} y_j \text{ subject to } x \in \mathcal{C}, \quad y_j = \pi(g_j(x)) \text{ and } y_j \geq 0 \quad \forall j \notin I(x^*).$$

If $j \in K(x^*)$, we have that $g_j(x) > 0$ in a neighborhood of x^* and $\pi(g_j(x))$ is continuous and differentiable on that neighborhood. Therefore, x^* is a local minimizer of the locally smooth problem given by:

$$\text{Minimize } f(x) + \sigma \left[\sum_{j \in J(x^*)} y_j + \sum_{j \in K(x^*)} \pi(g_j(x)) \right]$$

subject to $x \in \mathcal{C}$, $y_j = \pi(g_j(x))$ and $y_j \geq 0 \quad \forall j \in J(x^*)$.

Since $g_j(x^*) \leq 0$ for all $j \in J(x^*)$, x^* is also a local minimizer of:

$$\text{Minimize } f(x) + \sigma \left[\sum_{j \in J(x^*)} y_j + \sum_{j \in K(x^*)} \pi(g_j(x)) \right]$$

subject to $x \in \mathcal{C}$, $g_j(x) \leq 0$, $y_j = \pi(g_j(x))$ and $y_j \geq 0 \quad \forall j \in J(x^*)$.

Moreover, since $y_j^* = 0$ for all $j \in J(x^*)$, x^* is a local minimizer of:

$$\text{Minimize } f(x) + \sigma \sum_{j \in K(x^*)} \pi(g_j(x)) \text{ subject to } x \in \mathcal{C}, \quad g_j(x) \leq 0,$$

$$y_j = \pi(g_j(x)) \text{ and } y_j \geq 0 \quad \forall j \in J(x^*).$$

Clearly, the variables y_j for $j \in J(x^*)$ play no role in this problem, so x^* is a local minimizer of:

$$\text{Minimize } f(x) + \sigma \sum_{j \in K(x^*)} \pi(g_j(x)) \text{ subject to } x \in \mathcal{C} \text{ and } g_j(x) \leq 0 \quad \forall j \in J(x^*).$$

Then, if the constraints that define \mathcal{C} together with the constraints $g_j(x) \leq 0$, $j \in J(x^*)$ satisfy a constraint qualification at x^* , the KKT conditions of (16) hold at x^* . This completes the proof. \square

Observe that (16) is a constrained optimization problem of the form

$$\text{Minimize } f_{\text{obj}}(w) \text{ subject to } h_E(w) = 0 \text{ and } h_I(w) \leq 0, \quad (18)$$

where $f_{\text{obj}} : \mathbb{R}^n \rightarrow \mathbb{R}$, $h_E : \mathbb{R}^n \rightarrow \mathbb{R}^{n_E}$, and $h_I : \mathbb{R}^n \rightarrow \mathbb{R}^{n_I}$ are continuously differentiable. It is well known that, independently of constraint qualifications, every local minimizer w^* of (18) satisfies the Approximate KKT (AKKT) optimality condition [2, 7], which means that, given $\varepsilon_{\text{sol}} > 0$, $\varepsilon_{\text{kkt}} > 0$, $\varepsilon_{\text{feas}} > 0$, and $\varepsilon_{\text{comp}} > 0$, there exist $w \in \mathbb{R}^n$, $\lambda \in \mathbb{R}^{n_E}$, and $\mu \in \mathbb{R}_+^{n_I}$ such that $\|w - w^*\| \leq \varepsilon_{\text{sol}}$,

$$\|h_E(w)\|_\infty \leq \varepsilon_{\text{feas}}, \quad \|h_I(w)_+\|_\infty \leq \varepsilon_{\text{feas}}, \quad (19)$$

$$\mu_j \leq \varepsilon_{\text{comp}} \text{ for all } j \text{ such that } [h_I(w)]_j < -\varepsilon_{\text{feas}}, \quad (20)$$

and

$$\|\nabla f_{\text{obj}}(w) + \nabla h_E(w)\lambda + \nabla h_I(w)\mu\| \leq \varepsilon_{\text{kkt}}. \quad (21)$$

If $w \in \mathbb{R}^n$ is such that there exist $\lambda \in \mathbb{R}^{n_E}$ and $\mu \in \mathbb{R}_+^{n_I}$ fulfilling (19), (20), and (21), we say that w is an $(\varepsilon_{\text{feas}}/\varepsilon_{\text{comp}}/\varepsilon_{\text{kkt}})$ -AKKT point of (18).

Applying the AKKT definition to (16), we obtain the following result, independently of constraint qualifications.

Theorem 3.2 *Let x^* be a local minimizer of (6) and assume that π satisfies Assumptions A1 and A2. Let \mathcal{C} be defined by $H(x) = 0$ and $G(x) \leq 0$, where $H : \mathbb{R}^n \rightarrow \mathbb{R}^{n_H}$ and $G : \mathbb{R}^n \rightarrow \mathbb{R}^{n_G}$ are continuously differentiable. Assume that $\varepsilon_{\text{sol}} > 0$, $\varepsilon_{\text{kkt}} > 0$, $\varepsilon_{\text{feas}} > 0$, and $\varepsilon_{\text{comp}} > 0$ are arbitrary. Then, there exist $x \in \mathbb{R}^n$, $\lambda_H \in \mathbb{R}^{n_H}$, $\mu_G \in \mathbb{R}_+^{n_G}$, and $\mu_j \in \mathbb{R}_+$ for all j such that $g_j(x^*) = 0$, such that:*

$$\|x - x^*\| \leq \varepsilon_{\text{sol}},$$

$$\|H(x)\| \leq \varepsilon_{\text{feas}}, \quad \|G(x)_+\| \leq \varepsilon_{\text{feas}},$$

$$g_j(x) \leq \varepsilon_{\text{feas}} \text{ for all } j \text{ such that } g_j(x^*) = 0,$$

$$\begin{aligned} & \left\| \nabla f(x) + \sigma \sum_{g_j(x^*) > 0} \pi'(g_j(x)) \nabla g_j(x) + \nabla H(x)\lambda_H + \nabla G(x)\mu_G \right. \\ & \left. + \sum_{g_j(x^*)=0} \mu_j \nabla g_j(x) \right\| \leq \varepsilon_{\text{kkt}}, \end{aligned}$$

$$[\mu_G]_j \leq \varepsilon_{\text{comp}} \text{ for all } j \text{ such that } G_j(x) < -\varepsilon_{\text{feas}},$$

and

$$\mu_j \leq \varepsilon_{\text{comp}} \text{ for all } j \text{ such that } g_j(x) < -\varepsilon_{\text{feas}}.$$

Proof By Theorem 3.1, x^* is a local minimizer of (16). Then, the thesis follows from (19), (20), and (21). \square

4 Model-intensive algorithms

As we mentioned in Section 2, problem (13) can be solved using standard methods for smooth constrained optimization. In many applications, the cost of evaluating $f(x)$ exceeds, by far, the cost of evaluating the constraints. For these situations, “model intensive” (MI) methods are recommendable. At each iteration of such an algorithm, a model of the objective function is built and this model is minimized using the original constraints (instead of their linear approximations as in SQP and other popular approaches). The idea is to exploit the model as much as possible so that, hopefully, it will not be necessary to evaluate the expensive function many times.

The most popular model for an arbitrary sufficiently smooth function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ comes from considering its p th Taylor approximation:

$$T_p(\bar{x}, x) := \sum_{j=1}^p \frac{1}{j!} \left((x - \bar{x}_1) \frac{\partial}{\partial x_1} + \cdots + (x - \bar{x}_n) \frac{\partial}{\partial x_n} \right)^j f(\bar{x}), \quad (22)$$

so that:

$$f(x) = f(\bar{x}) + T_p(\bar{x}, x) + o(\|x - \bar{x}\|^{p+1}). \quad (23)$$

Observe that $T_1(\bar{x}, x) = \nabla f(\bar{x})^T(x - \bar{x})$ and $T_2(\bar{x}, x) = \nabla f(\bar{x})^T(x - \bar{x}) + \frac{1}{2}(x - \bar{x})^T \nabla^2 f(\bar{x})(x - \bar{x})$. Consequently, the model for $f(x) + \sigma \sum_{j=1}^d y_j$ will be $f(\bar{x}) + T_p(\bar{x}, x) + \sigma \sum_{j=1}^d y_j$.

In this section, $M_{\bar{x}, \bar{y}}(x, y)$ will be a model of $\Phi(x, y) - \Phi(\bar{x}, \bar{y})$ for all $\bar{x}, x \in \mathbb{R}^n$ and $\bar{y}, y \in \mathbb{R}^d$. Therefore, $\Phi(x, y) \approx \Phi(\bar{x}, \bar{y}) + M_{\bar{x}, \bar{y}}(x, y)$. Let us define, for all $\delta \geq 0$:

$$\mathcal{C}_\delta = \{x \in \mathbb{R}^n \mid \|H(x)\|_\infty \leq \delta \text{ and } \|G(x)_+\|_\infty \leq \delta\}, \quad (24)$$

$$D_\delta = \{(x, y) \in \mathbb{R}^n \times \mathbb{R}^d \mid x \in \mathcal{C}_\delta, y \geq 0, \text{ and } g_j(x) - \pi^{-1}(y_j) \leq \delta, j = 1, \dots, d\}, \quad (25)$$

and $D = D_0$. The following algorithm is similar to Algorithm 3.1 of [11].

Algorithm 4.1 Assume that $p \in \{1, 2, 3, \dots\}$, $\eta > 0$, $\alpha > 0$, $\rho_{\min} > 0$, $\tau_2 \geq \tau_1 > 1$, $\theta > 0$, and $(x^0, y^0) \in D_{\eta/4}$ are given with $y_j^0 = \pi(g_j(x^0))$, $j = 1, \dots, d$. Initialize $k \leftarrow 0$.

Step 1. Set $\rho \leftarrow 0$.

Step 2. Compute $x^{\text{trial}} \in \mathbb{R}^n$ and $y^{\text{trial}} \in \mathbb{R}^d$ such that

$$(x^{\text{trial}}, y^{\text{trial}}) \in D_{\eta_k} \text{ with } \eta_k = \eta \left(\frac{k+1}{k+2} \right) \quad (26)$$

and

$$M_{x^k, y^k}(x^{\text{trial}}, y^{\text{trial}}) + \rho \|(x^{\text{trial}} - x^k, y^{\text{trial}} - y^k)\|^{p+1} \leq 0. \quad (27)$$

Step 3. Test the condition

$$\Phi(x^{\text{trial}}, y^{\text{trial}}) \leq \Phi(x^k, y^k) - \alpha \|(x^{\text{trial}} - x^k, y^{\text{trial}} - y^k)\|^{p+1}. \quad (28)$$

If (28) holds, accept the trial point $(x^{\text{trial}}, y^{\text{trial}})$, define $\rho_k = \rho$, $x^{k+1} = x^{\text{trial}}$, $y^{k+1} = y^{\text{trial}}$, set $k \leftarrow k + 1$, and go to Step 1. Otherwise, update $\rho \leftarrow \max\{\rho_{\min}, \tau \rho\}$ with $\tau \in [\tau_1, \tau_2]$, and go to Step 2.

Under the model-intensive strategy, conditions (26) and (27) may be obtained by solving the subproblem:

$$\text{Minimize } M_{x^k, y^k}(x, y) + \rho \|(x - x^k, y - y^k)\|^{p+1} \text{ subject to } (x, y) \in D, \quad (29)$$

with rather high precision, since we assume that the cost of evaluating the model is negligible with respect to the cost of evaluating f . Note that the feasibility required at x^{trial} is looser than the precision with which x^k satisfies those constraints. Therefore, it is plausible to require the fulfillment of (26) and (27) by means of a standard constrained optimization algorithm. This means that, although conditions (27) and (28) are trivially satisfied by $x^{\text{trial}} = x^k$ and $y^{\text{trial}} = y^k$, it is reasonable to expect that a nontrivial solution with decrease of Φ could be obtained with non-null increments.

Complexity results for Algorithm 4.1 are presented below. Before that, the standard assumptions (see, for example, [10, 14]) are given.

Assumption A3 *For all (x^k, y^k) generated by Algorithm 4.1, we have that*

$$\nabla \Phi(x^k, y^k) = \nabla M_{x^k, y^k}(x, y)|_{(x, y) = (x^k, y^k)}.$$

Assumption A4 *There exists $L_1 > 0$ such that, for every (x^k, y^k) calculated by Algorithm 4.1 and for every $(x^{\text{trial}}, y^{\text{trial}})$ satisfying (27), we have that:*

$$\Phi(x^{\text{trial}}, y^{\text{trial}}) \leq \Phi(x^k, y^k) + M_{x^k, y^k}(x^{\text{trial}}, y^{\text{trial}}) + L_1 \|(x - x^{\text{trial}}, y - y^{\text{trial}})\|^{p+1}.$$

Assumption A4 is valid if M is obtained using the p th Taylor approximation and the derivatives of order p of f are Lipschitz continuous. See [3, 10, 14].

Assumption A5 *There exists $L_2 > 0$ such that, for every x^k, y^k, x^{k+1} , and y^{k+1} obtained by Algorithm 4.1, we have that:*

$$\|\nabla_{x, y} \Phi(x^{k+1}, y^{k+1}) - \nabla_{x, y} M_{x^k, y^k}(x^{k+1}, y^{k+1})\| \leq L_2 \|(x^{k+1} - x^k, y^{k+1} - y^k)\|^p.$$

As in the case of Assumption A4, Assumption A5 also holds if we use Taylor p th approximations for computing M and the p th order derivatives of f are Lipschitz-continuous.

Assumption A6 below says that (29) must be approximately solved. Before stating this assumption let us recall that, assuming that $\|(x - x^k, y - y^k)\|^{p+1}$ is continuously differentiable with respect to (x, y) , (29) is a constrained optimization problem of the form (18).

Assumption A6 *We say that this assumption holds at iteration k of Algorithm 4.1 if the function $\|(x - x^k, y - y^k)\|^{p+1}$ is continuously differentiable with respect to (x, y) and $(x^{\text{trial}}, y^{\text{trial}})$ is an $(\varepsilon_{\text{feas}}/\varepsilon_{\text{comp}}/\varepsilon_{\text{kkt}})$ -AKKT point of (29) with $\varepsilon_{\text{feas}} = \eta_k$, $\varepsilon_{\text{comp}} = \eta$, and $\varepsilon_{\text{kkt}} = \theta \|(x^{\text{trial}} - x^k, y^{\text{trial}} - y^k)\|^p$.*

Theorem 4.2 Assume that the sequence $\{(x^k, y^k)\}$ is generated by Algorithm 4.1, $\Phi_{\text{target}} \in \mathbb{R}$, $\varepsilon > 0$, and $L = \max\{L_1, L_2\}$. Define:

$$K(x^0, y^0, \Phi_{\text{target}}, \alpha, p, L, \tau_2, \theta) = \left\lfloor \left(\Phi(x^0, y^0) - \Phi_{\text{target}} \right) \left(\frac{\alpha^{p/(p+1)} \varepsilon}{L + \tau_2 (L + \alpha) (p + 1) + \theta} \right)^{-(p+1)/p} \right\rfloor. \quad (30)$$

Suppose that Assumptions A3, A4, A5, and A6 hold for all:

$$k \leq K(x^0, y^0, \Phi_{\text{target}}, \alpha, p, L, \tau_2, \theta).$$

Then, the number of iterations k such that:

$$\Phi(x^{k+1}, y^{k+1}) > \Phi_{\text{target}}$$

and (x^{k+1}, y^{k+1}) is not an $(\eta_k/\eta/\varepsilon)$ -AKKT point of (13) is bounded above by $K(x^0, y^0, \Phi_{\text{target}}, \alpha, p, L, \tau_2, \theta)$. Moreover, the number of functional evaluations per iteration is bounded above by:

$$\left\lfloor \log_{\tau_1} \left(\frac{\tau_2 (L + \alpha)}{\rho_{\min}} \right) \right\rfloor + 1.$$

Proof The desired result follows as in [11, Thm.3.1]. \square

Theorem 4.3 Suppose that Assumptions A3, A4, and A5 hold, and the function $\|(x - x^k, y - y^k)\|^{p+1}$ is continuously differentiable with respect to (x, y) . Assume, moreover, that for all $k = 0, 1, 2, \dots$ and $\varepsilon_{\text{feas}} > 0$, $\varepsilon_{\text{comp}} > 0$, and $\varepsilon_{\text{kkt}} > 0$, we are able to compute, using a suitable algorithm, an $(\varepsilon_{\text{feas}}/\varepsilon_{\text{comp}}/\varepsilon_{\text{kkt}})$ -AKKT point of (29) such that (27) holds. Then, there are two possibilities:

1. Assumption A6 holds for all $k \leq K(x^0, y^0, \Phi_{\text{target}}, \alpha, p, L, \tau_2, \theta)$.
2. There exists $k \leq K(x^0, y^0, \Phi_{\text{target}}, \alpha, p, L, \tau_2, \theta)$ such that (x^k, y^k) is an AKKT point of (13).

Proof Assume that $(x^{k,\ell}, y^{k,\ell})$ is a sequence generated by an algorithm that, when applied to the problem:

$$\text{Minimize } M_{x^k, y^k}(x, y) + \rho \|(x - x^k, y - y^k)\|^{p+1} \text{ subject to } (x, y) \in D_{\eta_{k-1}}, \quad (31)$$

satisfies the hypotheses of the theorem. By construction, $(x^k, y^k) \in D_{\eta_{k-1}}$. Therefore the minimum of $M_{x^k, y^k}(x, y) + \rho \|(x - x^k, y - y^k)\|^{p+1}$ onto $D_{\eta_{k-1}}$ is non-positive, as well as the minimum of $M_{x^k, y^k}(x, y) + \rho \|(x - x^k, y - y^k)\|^{p+1}$ onto D_{η_k} .

By the fulfillment of the AKKT optimality conditions, we have two possibilities:

1. There exists an iterate $(x^{k,\ell}, y^{k,\ell})$ which is an $(\varepsilon_{\text{feas}}/\varepsilon_{\text{comp}}/\varepsilon_{\text{kkt}})$ -AKKT point of (29) with $\varepsilon_{\text{feas}} = \eta_k$, $\varepsilon_{\text{comp}} = \eta$, and $\varepsilon_{\text{kkt}} = \theta \|(x^{k,\ell} - x^k, y^{k,\ell} - y^k)\|^p$.

2. There exist infinitely many iterates $(x^{k,\ell}, y^{k,\ell})$ that are $(\varepsilon_{\text{feas}}/\varepsilon_{\text{comp}}/\varepsilon_{\text{kkt}})$ -AKKT points of (29) with $\varepsilon_{\text{feas}} = 1/\ell < \eta_k$, $\varepsilon_{\text{comp}} = 1/\ell < \eta$, and $\varepsilon_{\text{kkt}} = 1/\ell$ but

$$\theta \|(x^{k,\ell} - x^k, y^{k,\ell} - y^k)\| < 1/\ell. \quad (32)$$

In the first case, we can choose $x^{\text{trial}} = x^{k,\ell}$ and $y^{\text{trial}} = y^{k,\ell}$, satisfying Assumption A6 and (27). In the second case, (32) implies that:

$$\lim_{\ell \rightarrow \infty} (x^{k,\ell}, y^{k,\ell}) = (x^k, y^k).$$

Therefore, (x^k, y^k) satisfies the AKKT optimality condition for (31). Since the first derivatives of $\Phi(x, y)$ at (x^k, y^k) coincide with the first derivatives of the objective function of (31), it turns out that (x^k, y^k) was an AKKT point for (31). In other words, Assumption A6 did not hold because (x^k, y^k) already was an approximate solution of the original problem. This completes the proof. \square

Theorems 4.2 and 4.3 say that after at most the number of iterations given by (30), one finds an iterate that satisfies the constraints of (13) with tolerance η such that the objective function value is smaller than or equal to Φ_{target} ; or, alternatively, we find an iterate that satisfies KKT conditions with tolerance η for feasibility, tolerance η for complementarity, and tolerance ε for optimality. The conclusion of these theorems is that, ultimately, Assumption A6 is not necessary since, if it holds for all $k \leq K(x^0, y^0, \Phi_{\text{target}}, \alpha, p, L, \tau_2, \theta)$, it guarantees that an approximate solution is found for some $k \leq K(x^0, y^0, \Phi_{\text{target}}, \alpha, p, L, \tau_2, \theta)$. But, if Assumption A6 does not hold for some $k \leq K(x^0, y^0, \Phi_{\text{target}}, \alpha, p, L, \tau_2, \theta)$, the iterate x^k is an approximate solution. This is stated in the following corollary.

Corollary 4.4 *Suppose that the hypotheses of Theorem 4.3 hold. Then, one of the following statements is true:*

1. *There exists $k \leq K(x^0, y^0, \Phi_{\text{target}}, \alpha, p, L, \tau_2, \theta)$, such that*

$$\Phi(x^{k+1}, y^{k+1}) \leq \Phi_{\text{target}}$$

and (x^{k+1}, y^{k+1}) is $(\eta_k/\eta/\varepsilon)$ -AKKT point of (13).
2. *There exists $k \leq K(x^0, y^0, \Phi_{\text{target}}, \alpha, p, L, \tau_2, \theta)$ such that (x^k, y^k) is an AKKT point of (13).*

5 Linearly constrained problems

In this section, we consider the case in which $G(x)$, $H(x)$, and $g_j(x)$, $j = 1, \dots, d$, are affine functions. Thus, their first derivatives are constant. Consequently, we denote:

$$A_G = G'(x), \quad A_H = H'(x), \quad a_j^T = g_j'(x), \quad j = 1, \dots, d,$$

for all $x \in \mathbb{R}^n$. Therefore, for all $\bar{x} \in \mathbb{R}^n$,

$$\begin{aligned} G(x) &= A_G(x - \bar{x}) + G(\bar{x}), \quad H(x) = A_H(x - \bar{x}) + H(\bar{x}), \\ g_j(x) &= a_j^T(x - \bar{x}) + g_j(\bar{x}), \quad j = 1, \dots, d. \end{aligned}$$

The constraints $H(x) = 0$ and $G(x) \leq 0$ are now linear but the constraints $g_j(x) - \pi^{-1}(y_j) \leq 0$ of (13) are not, due to the nonlinearity of π^{-1} . Therefore, the set \mathcal{C} is a polytope but the set D_0 , defined by (24) and (25), is not.

Of course, we may use Algorithm 4.1 for solving (13) in this case, but the presence of linearity suggests that linear constraints could be satisfied exactly at the solution of each subproblem; and that subproblems might be tackled by a linear-constraints optimization method. For achieving this improvement, the nonlinear constraints $g_j(x) - \pi^{-1}(y_j) \leq 0$ must be linearized. This amounts to replace, at each subproblem, each constraint $g_j(x) - \pi^{-1}(y_j) \leq 0$ with:

$$g_j(x) - [\pi^{-1}(y_j^k) + (\pi^{-1})'(y_j^k)(y_j - y_j^k)] \leq 0. \quad (33)$$

By the convexity of the function π^{-1} , the fulfillment of (33) implies that $g_j(x) - \pi^{-1}(y_j) \leq 0$. Therefore, the feasible set defined by $H(x) = 0$, $G(x) \leq 0$, and (33) for $j = 1, \dots, d$ is contained in D_0 .

Algorithm 5.1 This algorithm is identical to Algorithm 4.1 except that condition (26) is replaced by:

$$(x^{\text{trial}}, y^{\text{trial}}) \in \mathcal{C} \text{ and } g_j(x) - [\pi^{-1}(y_j^k) + (\pi^{-1})'(y_j^k)(y_j - y_j^k)] \leq 0, \quad j = 1, \dots, d. \quad (34)$$

Of course, the parameter η is not necessary anymore since we assume that linear constraints can be satisfied exactly using well-established constrained optimization methods.

Assumptions A3, A4, and A5 stand exactly in the same way as in Section 4. However, Assumption A6 needs to be replaced in order to take into account that now the subproblem has only linear constraints and that d constraints are linear approximations of the true ones.

Assumption A7 At each iteration k of Algorithm 5.1, the function $\|(x - x^k, y - y^k)\|^{p+1}$ is continuously differentiable with respect to (x, y) and (x^{k+1}, y^{k+1}) satisfies the KKT conditions for the linearly constrained problem that consists of minimizing $M_{x^k, y^k}(x, y) + \rho\|(x - x^k, y - y^k)\|^{p+1}$ subject to $H(x) = 0$, $G(x) \leq 0$, and $g_j(x) - [\pi^{-1}(y_j^k) + (\pi^{-1})'(y_j^k)(y_j - y_j^k)] \leq 0$, $j = 1, \dots, d$.

Assumption A7 is plausible because every minimizer of linearly constrained optimization problems satisfies KKT conditions.

According to Assumption A7, the increment (x^{k+1}, y^{k+1}) at each iteration of Algorithm 5.1 must satisfy the following conditions:

$$\begin{aligned} & \nabla_x \left[M_{x^k, y^k}(x^{k+1}, y^{k+1}) + \rho\|(x^{k+1} - x^k, y^{k+1} - y^k)\|^{p+1} \right] \\ & + A_H^T \lambda + A_G^T \mu_+ + \sum_{j=1}^d a_j \beta_j = 0, \end{aligned} \quad (35)$$

$$\rho \frac{\partial}{\partial y_j} \|(x^{k+1} - x^k, y^{k+1} - y^k)\|^{p+1} + \sigma_j - (\pi^{-1})'(y_j^{k+1})\beta_j - \gamma_j = 0, \quad j = 1, \dots, d, \quad (36)$$

$$H(x^{k+1}) = 0, \quad (37)$$

$$G(x^{k+1}) \leq 0, \quad (38)$$

$$g_j(x^{k+1}) - \left[\pi^{-1}(y_j^k) + (\pi^{-1})'(y^k)(y_j^{k+1} - y_j^k) \right] \leq 0, \quad j = 1, \dots, d, \quad (39)$$

$$\beta_j = 0 \text{ whenever } g_j(x^{k+1}) - [\pi^{-1}(y_j^k) + (\pi^{-1})'(y^k)(y_j^{k+1} - y_j^k)] < 0, \quad (40)$$

$$y^{k+1} \geq 0, \quad (41)$$

$$\mu_j = 0 \text{ for all } j \text{ such that } G(x^{k+1})_j < 0, \quad (42)$$

$$\gamma_j = 0 \text{ for all } j \text{ such that } y_j^{k+1} > 0, \quad (43)$$

$$\mu, \beta, \gamma \geq 0. \quad (44)$$

Recall that (x, y) satisfies the KKT conditions of (13) if conditions (35)–(44) hold with (x, y) replacing (x^{k+1}, y^{k+1}) , except that (35) and (36) must be replaced by:

$$\nabla f(x) + A_H^T \lambda + A_G^T \mu_+ \sum_{j=1}^d a_j \beta_j = 0 \quad (45)$$

and

$$\sigma_j - (\pi^{-1})'(y_j) \beta_j - \gamma_j = 0, \quad j = 1, \dots, d, \quad (46)$$

whereas (39) and (40) must be replaced by:

$$g_j(x) - \pi^{-1}(y_j) \leq 0 \quad (47)$$

and

$$\beta_j = 0 \text{ whenever } g_j(x) - \pi^{-1}(y_j) < 0. \quad (48)$$

Therefore, the question is whether (36)–(44) implies some relaxed version of (45)–(48) for $x = x^{k+1}$ and $y = y^{k+1}$.

As in [11, Lemma 3.2], by Assumption A4, it follows that the values of ρ used at each subproblem of Algorithm 5.1 are bounded. By Assumption A5, (35) implies that:

$$\|\nabla f(x^{k+1}) + A_H^T \lambda + A_G^T \mu_+ \sum_{j=1}^d a_j \beta_j\| \leq O(\|(x^{k+1} - x^k, y^{k+1} - y^k)\|^p). \quad (49)$$

By the limitation of ρ and the fact that $|\frac{\partial}{\partial y_j} \|(x - x^k, y - y^k)\|^{p+1}| \leq O(\|(x^{k+1} - x^k, y^{k+1} - y^k)\|^p)$, (36) implies that:

$$|\sigma_j - (\pi^{-1})'(y_j^{k+1}) \beta_j - \gamma_j| \leq O(\|(x^{k+1} - x^k, y^{k+1} - y^k)\|^p), \quad j = 1, \dots, d. \quad (50)$$

By (39) and the convexity of π^{-1} , we have that (47) holds with $x = x^{k+1}$ and $y = y^{k+1}$. Thus,

$$g_j(x^{k+1}) - \pi^{-1}(y_j^{k+1}) \leq 0, \quad j = 1, \dots, d. \quad (51)$$

For proving approximate complementarity, we need a new assumption regarding the replacement of π^{-1} with its linear approximation.

Assumption A8 *There exists $c_\pi > 0$ such that, for all $z, y \geq 0$,*

$$|\pi^{-1}(y) - [\pi^{-1}(z) + (\pi^{-1})'(z)(z - y)]| \leq c_\pi |y - z|^2.$$

By Assumption A8, we have that $g_j(x^{k+1}) - \pi^{-1}(y_j^{k+1}) < -c_\pi |y_j^{k+1} - y_j^k|^2$ implies that $g_j(x^{k+1}) - [\pi^{-1}(y_j^k) + (\pi^{-1})'(y^k)(y_j^{k+1} - y_j^k)] < 0$; so, by (40), $\beta_j = 0$.

Theorem 5.2 *Suppose that Assumptions A3, A4, A5, A7, and A8 hold, $L = \max\{L_1, L_2\}$, the sequence $\{(x^k, y^k)\}$ is generated by Algorithm 5.1, $\Phi_{\text{target}} \in \mathbb{R}$, and $\varepsilon > 0$. Then, the number of iterations k such that:*

$$\Phi(x^{k+1}, y^{k+1}) > \Phi_{\text{target}}$$

and (x^{k+1}, y^{k+1}) is not an $(\eta_k/\eta/\varepsilon)$ -AKKT point of (13) is not greater than

$$\left\lfloor \left((\Phi(x^0, y^0) - \Phi_{\text{target}}) \right) \left(\frac{\alpha^{p/(p+1)} \varepsilon}{L + \tau_2 (L + \alpha) (p + 1) + \theta} \right)^{-(p+1)/p} \right\rfloor. \quad (52)$$

Moreover, the number of functional evaluations per iteration is bounded above by:

$$\left\lfloor \log_{\tau_1} \left(\frac{\tau_2 (L + \alpha)}{\rho_{\min}} \right) \right\rfloor + 1.$$

Proof As in [11, Lemma 3.1], we obtain that:

$$\lim_{k \rightarrow \infty} \|x^{k+1} - x^k\| = \lim_{k \rightarrow \infty} \|y^{k+1} - y^k\| = 0.$$

As in [11, Lemma 3.1], we prove that the sequence of penalty parameters is bounded.

By the arguments presented above, we have that, for all k , there exist $\lambda = \lambda^k \in \mathbb{R}^{n_H}$, $\mu = \mu_k \in \mathbb{R}_+^{n_G}$, $\beta = \beta^k \in \mathbb{R}_+^d$, and $\gamma = \gamma^k \in \mathbb{R}_+^d$ such that:

$$\|\nabla f(x^{k+1}, y^{k+1}) + A_H^T \lambda + A_G^T \mu + \sum_{j=1}^d a_j \beta_j\| \leq O(\|(x^{k+1}, y^{k+1}) - (x^k, y^k)\|^p), \quad (53)$$

$$\sigma_j - (\pi^{-1})'(y_j^{k+1}) \beta_j - \gamma_j = 0, \quad j = 1, \dots, d, \quad (54)$$

$$H(x^{k+1}) = 0, \quad (55)$$

$$G(x^{k+1}) \leq 0, \quad (56)$$

$$g_j(x^{k+1}) - \pi^{-1}(y_j^{k+1}) \leq 0, \quad j = 1, \dots, d, \quad (57)$$

$$\beta_j = 0 \text{ whenever } g_j(x^{k+1}) - \pi^{-1}(y_j^{k+1}) < -c_\pi |y_j^{k+1} - y_j^k|^2, \quad (58)$$

$$y^{k+1} \geq 0, \quad (59)$$

$$\mu_j = 0 \text{ for all } j \text{ such that } G(x^{k+1})_j < 0, \quad (60)$$

$$\gamma_j = 0 \text{ for all } j \text{ such that } y_j^{k+1} > 0, \quad (61)$$

$$\mu, \beta, \gamma \geq 0. \quad (62)$$

By the sufficient descent condition and (53), there exists $c_{\nabla} > 0$ such that, for all k :

$$\Phi(x^{k+1}, y^{k+1}) \leq \Phi(x^k, y^k) - c_{\nabla} \left\| \nabla f(x^{k+1}, y^{k+1}) + A_H^T \lambda + A_G^T \mu + \sum_{j=1}^d a_j \beta_j \right\|^{\frac{p+1}{p}}. \quad (63)$$

Therefore, given $\varepsilon > 0$, the number of iterations such that $\Phi(x^{k+1}, y^{k+1}) > \Phi_{\text{target}}$ and

$$\left\| \nabla f(x^{k+1}, y^{k+1}) + A_H^T \lambda + A_G^T \mu + \sum_{j=1}^d a_j \beta_j \right\| > \varepsilon \quad (64)$$

is, at most,

$$(\Phi(x^0, y^0) - \Phi_{\text{target}}) \times O(\varepsilon^{\frac{p+1}{p}}).$$

Given $\xi > 0$, by the descent condition (28), the number of iterations such that $\|(x^{k+1} - x^k, y^{k+1} - y^k)\|^{p+1} > \xi$ cannot exceed the quantity $(\Phi(x^0, y^0) - \Phi_{\text{target}})/(\alpha\xi)$. Therefore, the number of iterations such that $\|(x^{k+1} - x^k, y^{k+1} - y^k)\|^2 > \xi^{2/(p+1)}$ cannot exceed the quantity $(\Phi(x^0, y^0) - \Phi_{\text{target}})/(\alpha\xi)$. Therefore, the number of iterations such that $\|(x^{k+1} - x^k, y^{k+1} - y^k)\|^2 > \xi$ cannot exceed the quantity $(\Phi(x^0, y^0) - \Phi_{\text{target}})/(\alpha\xi^{\frac{p+1}{2}})$. Therefore, the number of iterations such that $c_{\pi} \|(x^{k+1} - x^k, y^{k+1} - y^k)\|^2 > c_{\pi}\xi$ cannot exceed the quantity $(\Phi(x^0, y^0) - \Phi_{\text{target}})/(\alpha\xi^{\frac{p+1}{2}})$. Therefore, the number of iterations such that $c_{\pi} \|(x^{k+1} - x^k, y^{k+1} - y^k)\|^2 > \xi$ cannot exceed the quantity $(\Phi(x^0, y^0) - \Phi_{\text{target}})/(\alpha(\xi/c_{\pi})^{\frac{p+1}{2}})$. Therefore, after at most

$$\frac{(\Phi(x^0, y^0) - \Phi_{\text{target}})}{\alpha(\xi/c_{\pi})^{\frac{p+1}{2}}}$$

iterations, we have that all the iterates satisfy the approximate complementarity condition:

$$\beta_j = 0 \text{ whenever } g_j(x^{k+1}) - \pi^{-1}(y_j^{k+1}) < -\xi$$

for all $j = 1, \dots, d$. This completes the proof. \square

6 Numerical experiments

In this section, we present numerical experiments with the penalty function $\pi(\cdot) = \sqrt{\cdot}$. In Section 6.1, we aim to illustrate in which way the use of this penalty function promotes sparsity. In Section 6.2, we consider all the minimization problems from the Moré-Garbow-Hillstrom collection [27], penalized with the term $\pi(\cdot) = \sqrt{\cdot}$. Problems are solved in two different ways. On the one hand, reformulated problems are solved with the nonlinear programming solver Algencan [1, 7]. On the other hand, problems are solved with Algorithm 4.1. Experiments aim to simulate the situation in which a model-intensive algorithm is used to solve a problem with a costly objective function and a nonconvex regularization term.

6.1 Numerical illustration

In this section, we present numerical experiments that illustrate in which way the use of the penalty function $\pi(\cdot) = \sqrt{\cdot}$ promotes sparsity. Three toy problems are reformulated and solved with the general-use nonlinear programming solver Algen- can [1, 7] that applies to smooth problems.

Problem 1. Regularized gradient minimization

The continuous version of this problem is to find a function $u : [0, 1] \rightarrow \mathbb{R}$ that solves the problem:

$$\text{Minimize } \|\nabla u\|^2 \text{ subject to } \int_0^1 u(t) dt = \int_0^1 \bar{u}(t),$$

where $\bar{u}(t) = 400$ if $t \in [0, 0.5]$ and $\bar{u}(t) = 0$, otherwise. In addition, we aim u to coincide with \bar{u} in “as much as possible.” After discretization, the problem becomes:

$$\text{Minimize } \frac{1}{2} \sum_{i=2}^{99} (u_{i-1} - u_{i+1})^2 \text{ subject to } \sum_{i=1}^{100} u_i = \sum_{i=1}^{100} \bar{u}_i,$$

where $\bar{u}_i = 400$ for $i = 1, \dots, 50$ and $\bar{u}_i = 0$ for $i = 51, \dots, 100$, with the additional constraint of having $u_i = \bar{u}_i$, for $i = 1, \dots, 100$, as many times as possible.

Solutions to the discretized version of the problem can be found by solving the reformulated problem given by:

$$\begin{aligned} \text{Minimize } & \frac{1}{2} \sum_{i=2}^{99} (u_{i-1} - u_{i+1})^2 + \sigma \sum_{i=1}^{100} y_i \\ \text{subject to } & \sum_{i=1}^{100} u_i = \sum_{i=1}^{100} \bar{u}_i \\ & -y_i^2 \leq u_i - \bar{u}_i \leq y_i^2, \quad i = 1, \dots, 100 \end{aligned}$$

for different values of $\sigma > 0$. Note that the reformulation corresponds to penalizing $\pi(|u_i - \bar{u}_i|)$, $i = 1, \dots, 100$, with $\pi(\cdot) = \sqrt{\cdot}$. Figure 1 shows a graphical representation of solutions to problems with $\sigma \in \{0, 1, 2, 3, 4, 5, 10, 20, 30, 40, 50, 100\}$. Note that, the larger σ is, the larger the number of constraints of the form $u_i = \bar{u}_i$ being satisfied.

Problem 2. Two-dimensional dam

This problem is a two-dimensional version of Problem 1. The discretized version of the problem is given by:

$$\begin{aligned} \text{Minimize } & \frac{1}{2} \sum_{i=2}^{29} \sum_{j=2}^{19} (4u_{ij} - u_{i,j-1} - u_{i,j+1} - u_{i-1,j} - u_{i+1,j})^2 \\ \text{subject to } & \sum_{i=1}^{30} \sum_{j=1}^{20} u_{ij} = \sum_{i=1}^{30} \sum_{j=1}^{20} \bar{u}_{ij}, \end{aligned}$$

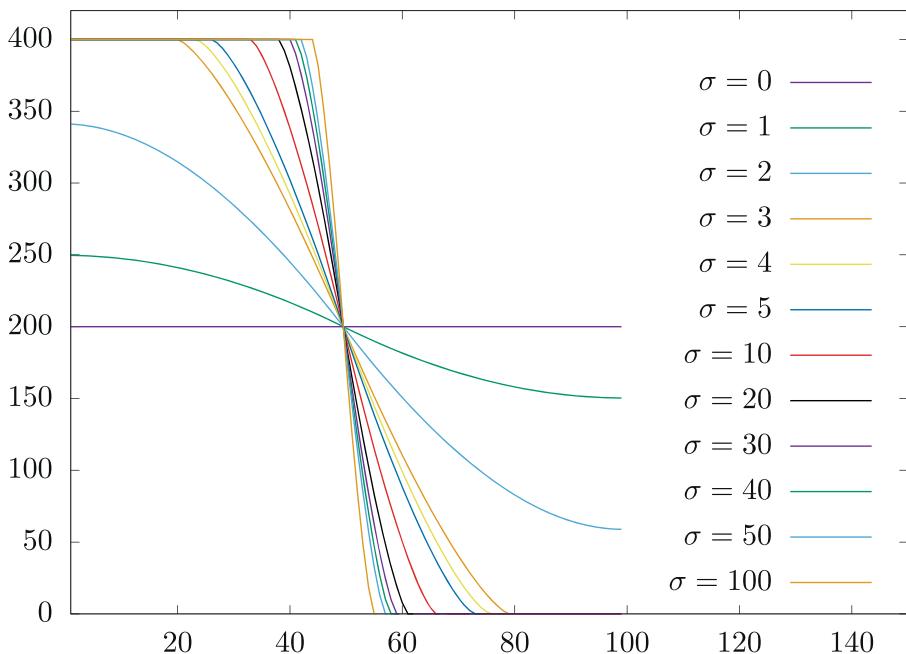


Fig. 1 Graphical representation of solutions to problem 1 for different values of σ

where

$$u_{ij} = \begin{cases} 400, & \text{if } j \leq 30 \left[\frac{1}{2} + \frac{1}{10} \sin \left(2\pi \left(\frac{i-1}{30-1} \right) \right) \right] \\ 0, & \text{otherwise,} \end{cases}$$

with the additional constraint of having $u_{ij} = \bar{u}_{ij}$, for $i = 1, \dots, 30$ and $j = 1, \dots, 20$, as many times as possible. Solutions to the problem can be found by solving the reformulated problem given by:

$$\begin{aligned} & \text{Minimize} \quad \frac{1}{2} \sum_{i=2}^{99} (u_{i-1} - u_{i+1})^2 + \sigma \sum_{i=1}^{100} y_i \\ & \text{subject to} \quad \sum_{i=1}^{100} u_i = \sum_{i=1}^{30} \sum_{j=1}^{20} \bar{u}_{ij} \\ & \quad -y_{ij}^2 \leq u_{ij} - \bar{u}_{ij} \leq y_{ij}^2, \quad i = 1, \dots, 30, \quad j = 1, \dots, 20, \end{aligned}$$

for different values of $\sigma > 0$. Note that the reformulation corresponds to penalizing $\pi(|u_{ij} - \bar{u}_{ij}|)$, $i = 1, \dots, 30$, $j = 1, \dots, 20$, with $\pi(\cdot) = \sqrt{\cdot}$. Figure 2 shows a graphical representation of solutions to problems with $\sigma \in \{10^8, 10^7, \dots, 1, 10^{-1}, \dots, 10^{-8}\}$. Once again, the larger σ is, the larger the number of constraints of the form $u_{ij} = \bar{u}_{ij}$ being satisfied.

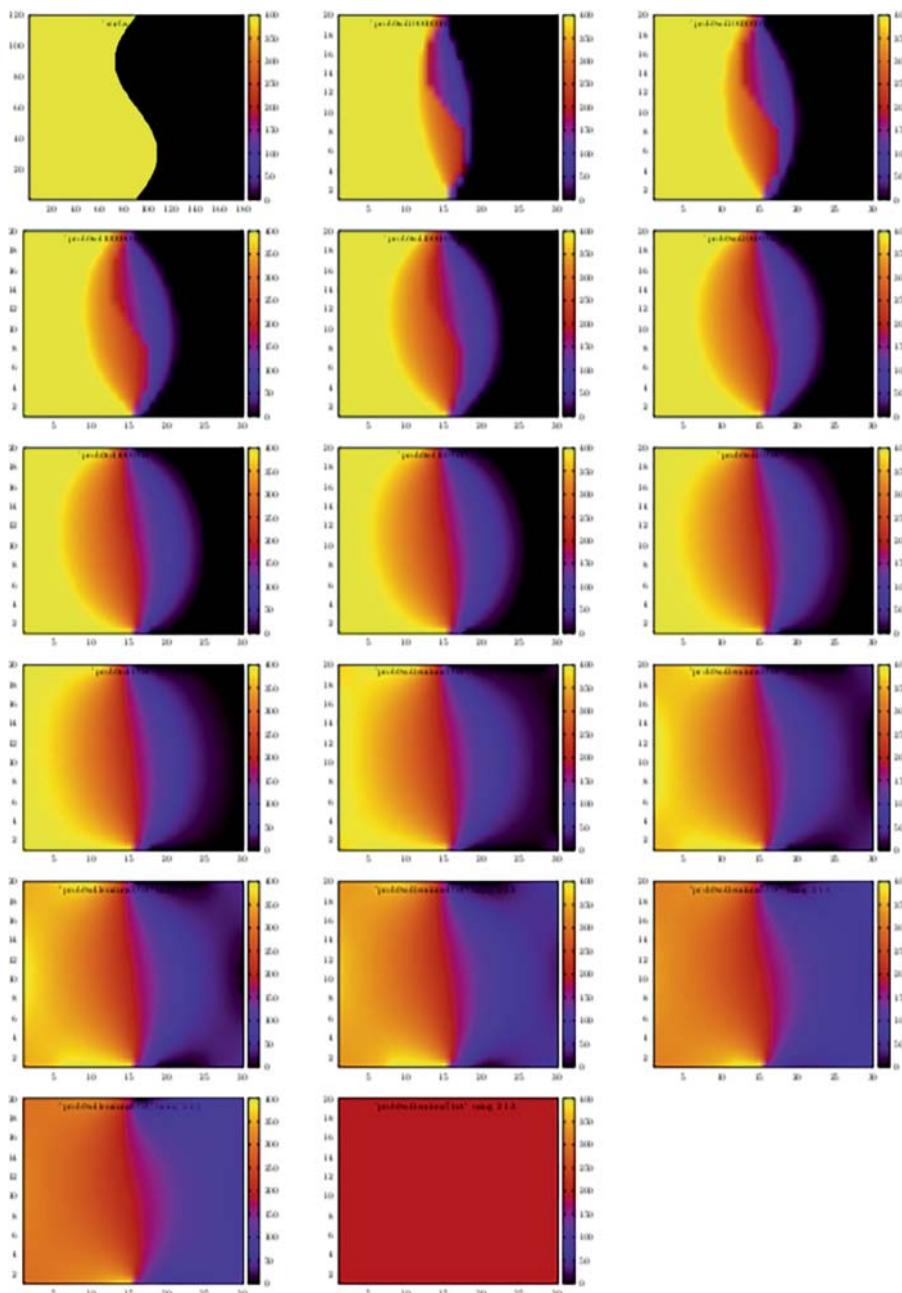


Fig. 2 Graphical representation of solutions to Problem 2 for different values of σ

Problem 3. Mass transportation

Assume that in a rectangle of $n_x \times n_y$ pixels, to each pixel (i, j) there are associated functions $P(i, j) \geq 0$ and $Q(i, j) \geq 0$ such that:

$$\sum_{i=1}^{n_x} \sum_{j=1}^{n_y} P(i, j) = \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} Q(i, j).$$

$P(i, j)$ represents white mass and $Q(i, j)$ represents red mass. The goal is to find a function $u((i, j), (k, \ell)) \geq 0$ that represents the transportation of white mass from pixel (i, j) to pixel (k, ℓ) . There is no transportation of red mass. The function of transported mass u must be such that, at every pixel, the amount of white mass be equal to the amount of red mass. Therefore, for every pixel $(i, j) \in V$, we must have that:

$$Q(i, j) = P(i, j) + \sum_{(k, \ell) \in V, (k, \ell) \neq (i, j)} u((k, \ell), (i, j)) - u((i, j), (k, \ell)),$$

where $V = \{(1, 2, \dots, n_x) \times (1, 2, \dots, n_y)\}$. With these constraints, the objective is to minimize the number of non-null transports, i.e., to minimize the number of $u((i, j), (k, \ell))$, for $(i, j) \neq (k, \ell) \in V$, that are positive.

The reformulated problem is given by:

$$\begin{aligned} \text{Minimize} \quad & \sigma \sum_{(i, j) \neq (k, \ell) \in V} y_{ijk\ell} \\ \text{subject to} \quad & Q(i, j) = P(i, j) + \sum_{(k, \ell) \in V, (k, \ell) \neq (i, j)} u((k, \ell), (i, j)) \\ & - u((i, j), (k, \ell)), \quad \forall (i, j) \in V \\ & 0 \leq u((i, j), (k, \ell)) \leq y_{ijk\ell}^2, \quad \forall (i, j) \neq (k, \ell) \in V. \end{aligned}$$

It corresponds to penalize $\sqrt{\sum_{(i, j) \neq (k, \ell) \in V} u((i, j), (k, \ell))}$ that represents the cost of the transportation process, its concavity representing economy of scale (decreasing marginal cost of transportation). Figure 3 illustrates solutions to a small instance of Problem 3 with $\sigma = 0$ and $\sigma > 0$.

6.2 Numerical comparison

We implemented the model-intensive Algorithm 4.1 in Fortran. At Step 2, a pair $(x^{\text{trial}}, y^{\text{trial}})$ satisfying (26) and (27) is computed by tackling problem:

$$\begin{aligned} \text{Minimize} \quad & M_{x^k, y^k}(x, y) + \rho \|x - x^k, y - y^k\|^{p+1} \\ \text{subject to} \quad & -y_i^2 \leq x_i \leq y_i^2 \text{ for } i = 1, \dots, n \text{ and } y \geq 0, \end{aligned} \quad (65)$$

where

$$M_{x^k, y^k}(x, y) = T_p(x^k, x) - f(x^k) + \sum_{i=1}^n [y - y^k]_i,$$

with the help of Algencan [1, 7]. By definition, $\eta_k \in [\frac{1}{2}\eta, \eta)$ for all k . Thus, in order to satisfy (26), it is enough to ask to Algencan to stop at a point satisfying

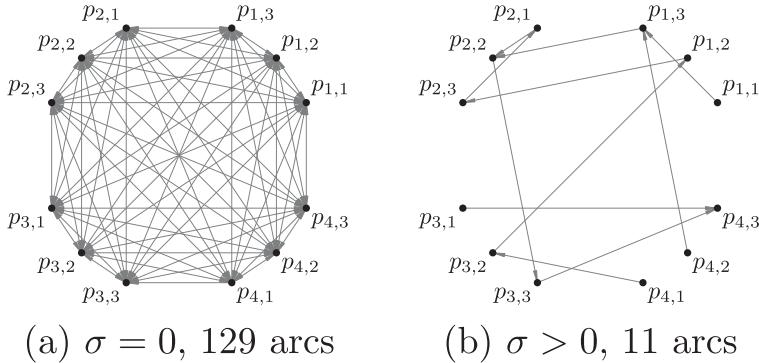


Fig. 3 Graphical representation of solutions to a small instance of Problem 3

the constraints with precision $\varepsilon_{\text{feas}}$ for any $\varepsilon_{\text{feas}} \leq \frac{1}{2}\eta$. Algencan also requires tolerances $\varepsilon_{\text{compl}}$ and ε_{opt} for complementarity and optimality measures, respectively. (See [8, §5.1] for details.) Since (27) is satisfied by $(x, y) = (0, 0)$, that is a feasible point of problem (65), in practice, it is expected to be always satisfied at the final iterate of Algencan. In the numerical experiments, in Algorithm 4.1, we arbitrarily set $p \in \{2, 3\}$, $\tau_1 = \tau_2 = 100$, $\rho_{\min} = 10^{-8}$, $\alpha = 10^{-8}$, $\eta = 10^{-6}$, and $\theta = 10^6$; and, in Algencan, $\varepsilon_{\text{feas}} = \varepsilon_{\text{compl}} = \varepsilon_{\text{opt}} = 10^{-8}$. In order to guarantee the fulfillment of Assumption A6, Algencan should have been modified to accept a tolerance ε_{opt} depending on its current iterate, i.e. not a constant. Instead of doing that, we keep Algencan as it is and we observed in practice that Assumption A6 would have been satisfied setting $\theta = 10^6$, which justifies that choice.

In the numerical experiments, we considered the 18 unconstrained minimization problems of the Moré-Garbow-Hillstrom collection [27] that consist in minimizing $f(x)$, penalized with the term $\sigma \sum_{i=1}^n \pi(|x_i|)$ with $\pi(\cdot) = \sqrt{\cdot}$ and $\sigma = 10^{-8}$. Note that, when applying Algorithm 4.1 with $p = 3$, evaluating $T_p(x^k, x)$ requires the third-order derivatives of f that were taken from [9]. These problems were solved in two different ways. On the one hand, problems were reformulated as in Section 6.1 and solved with Algencan. On the other hand, they were solved with Algorithm 4.1 with $p \in \{2, 3\}$. The same stopping criterion was adopted in both cases, namely, the approximate satisfaction of the optimality conditions given by:

$$\begin{aligned} \|g(x)_+\|_\infty &\leq \varepsilon \\ \|P_\Omega \left(x - \left[\nabla f(x) + \sum_{j=1}^p \mu_j \nabla g_j(x) \right] \right) - x\|_\infty &\leq \varepsilon \\ \max_{j=1, \dots, p} \{\min\{-g_j(x), \mu_j\}\} &\leq \varepsilon \end{aligned}$$

where $p = 2n$, $g_j(x, y) := x_j - y_j^2$ and $g_{n+j}(x, y) := -x_j - y_j^2$ for $j = 1, \dots, n$, P_Ω is the projector operator onto Ω , and $\Omega = \{(x, y) \in \mathbb{R}^{2n} \mid y \geq 0\}$, with $\varepsilon = 10^{-6}$. Emulating the situation to which Algorithm 4.1 is applicable, we considered

the number of evaluations of f as performance metric. Equivalent solutions were found with the two approaches in all the problems. Table 1 and Fig. 4 show the results. Figures show that Algorithm 4.1 with $p = 2$ and $p = 3$ used, in average 72% and 82% of the number of functional evaluations used by Algencan, respectively. Algorithm 4.1 with $p = 2$ used a larger number of evaluations than Algencan in 2 problems, the same number in 3 problems, and less evaluations in all the other 13 problems. The apparent inferiority of Algorithm 4.1 with $p = 3$, with respect to the case with $p = 2$, can be attributed to the (lack of) parameter tuning, since the arbitrary value of the (dimensional) sufficient decrease parameter $\alpha = 10^{-8}$ in (28) has a different meaning when $p = 2$ and $p = 3$.

7 Final remarks

In this work, we have introduced a smooth reformulation for constrained smooth problems with nonsmooth regularizations. The reformulation is entirely equivalent to the original problem both from the global and the local points of view. Moreover, we were able to prove optimality conditions in which auxiliary variables do not appear at all.

Table 1 Comparison of the performance of Algencan and Algorithm 4.1 with $p \in \{2, 3\}$ when applied to solving the Moré-Garbow-Hillstrom minimization problems with the penalty term given by $\pi(\cdot) = \sqrt{\cdot}$

Problem	$f(x^*)$	Number of functional evaluations		
		Algencan	Alg. 4.1 (2nd)	Alg. 4.1 (3rd)
Helical valley	3.00e-05	39	13	9
Biggs EXP6	5.91e-05	211	41	47
Gaussian	3.00e-05	3	3	2
Powell badly scaled	2.00e-05	254	90	85
Box three-dimensional	3.00e-05	49	21	31
Variably dimensioned	1.00e-04	14	15	20
Watson	2.35e-03	13	13	22
Penalty I	6.23e-05	46	38	44
Penalty II	4.86e-05	74	124	9
Brown badly scaled	3.07e-03	487	21	44
Brown and Dennis	8.58e+04	8	9	15
Gulf research and development	2.98e-05	36	35	35
Trigonometric	1.28e-04	15	10	14
Extended Rosenbrock	9.97e-05	34	27	43
Extended Powell singular	1.20e-04	17	18	27
Beale	2.00e-05	28	9	12
Wood	4.00e-05	90	53	72
Chebyquad	3.60e-03	31	19	21

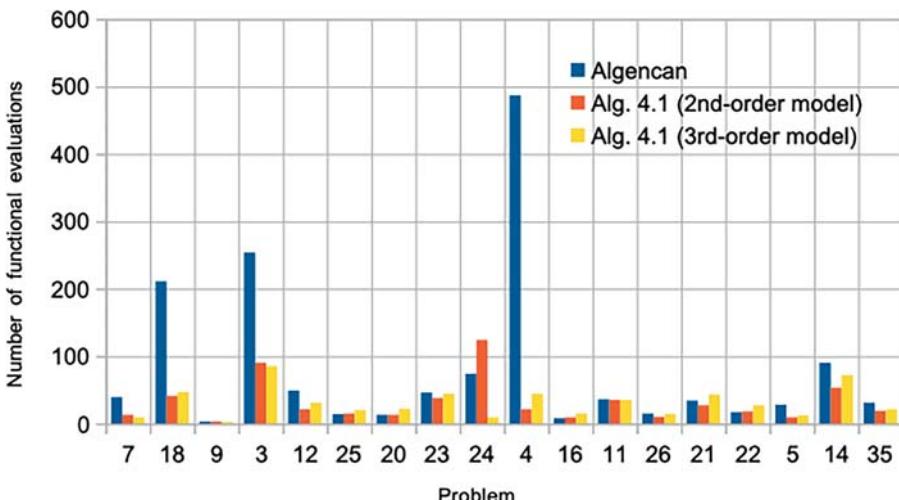


Fig. 4 Graphical representation of the performance of Algencan and Algorithm 4.1 with $p \in \{2, 3\}$ when applied to solving the Moré-Garbow-Hillstrom minimization problems with the penalty term given by $\pi(\cdot) = \sqrt{\cdot}$

Our main interest relies now in the application of these techniques to real problems in which the evaluation of the objective function is overwhelmingly more expensive than the evaluation of the constraints. This type of functions appear when PDE calculations are involved in the objective function evaluation and when the objective function is related to some phenomenon that takes place in real time. In these cases, model-intensive algorithms as the ones introduced in this paper may be useful. So, it is interesting to show that, from the theoretical point of view (convergence and complexity), these algorithms are well supported. Of course, smooth reformulations are valuable in these cases because one may take advantage of well-established constrained optimization software.

Funding information This work was supported by FAPESP (grants 2013/07375-0, 2016/01860-1, and 2018/24293-0) and CNPq (grants 438185/2018-8, 302538/2019-4, and 302682/2019-8).

References

1. Andreani, R., Birgin, E.G., Martínez, M.J., Schuverdt, M.L.: On augmented Lagrangian methods with general lower-level constraints. *SIAM J. Optim.* **18**, 1286–1309 (2008)
2. Andreani, R., Haeser, G., Martínez, M.J.: On sequential optimality conditions for smooth constrained optimization. *Optimization* **60**, 627–641 (2011)
3. Bertsekas, D.P.: *Nonlinear programming*, Athena Scientific (1999)
4. Bian, W., Chen, X.: From sparse solutions of systems of equations to sparse modeling of signals and image. *SIAM Rev.* **51**, 34–81 (2009)
5. Bian, W., Chen, X.: Linearly constrained non Lipschitz optimization for image restoration. *SIAM J. Imaging Sci.* **8**, 2294–2322 (2015)
6. Bian, W., Chen, X., Ye, Y.: Complexity analysis of interior point algorithms for non-Lipschitz and nonconvex minimization. *Math. Program.* **149**, 301–327 (2015)

7. Birgin, E.G., Martínez, M.J.: Practical augmented Lagrangian methods for constrained optimization society for industrial and applied mathematics, Philadelphia (2014)
8. Birgin E.G., Martínez, M.J.: Complexity and performance of an augmented Lagrangian algorithm, Optimization Methods and Software, to appear. <https://doi.org/10.1080/10556788.2020.1746962>
9. Birgin, E.G., Gardenghi, J.L., Martínez, M.J., Santos, S.A.: On the use of third-order models with fourth-order regularization for unconstrained optimization, Optimization Letters, to appear. <https://doi.org/10.1007/s11590-019-01395-z>
10. Birgin, E.G., Gardenghi, J.L., Martínez, M.J., Santos, S.A., Toint, P.L.: Worst-case evaluation complexity for unconstrained nonlinear optimization using higher order regularized models. *Math. Progr.* **163**, 359–368 (2017)
11. Birgin, E.G., Martínez, M.J.: On regularization and active-set methods with complexity for constrained optimization. *SIAM J. Optim.* **28**, 1367–1395 (2018)
12. Browne, S.: The risk and rewards of minimizing shortfall probability. *J. Portf. Manag.* **25**, 76–85 (1999)
13. Candes, E., Wakin, M., Boyd, S.: Enhancing sparsity by reweighted L_1 minimization. *J. Fourier Anal. Appl.* **14**, 877–905 (2008)
14. Cartis, C., Gould, N.I.M., Toint, P.H.L.: Second-order optimality and beyond: characterization and evaluation complexity in convexly-constrained nonlinear optimization. *Found. Comput. Math.* **18**, 1073–1107 (2018)
15. Chen, X., Guo, L., Lu, Z., Ye, J.: An augmented Lagrangian method for non-Lipschitz nonconvex programming. *SIAM J. Numer. Anal.* **55**, 168–193 (2017)
16. Chen, X., Niu, L., Yuan, Y.: Optimality conditions and a smoothing trust region Newton method for non-Lipschitz optimization. *SIAM J. Optim.* **23**, 1528–1552 (2013)
17. Chen, X., Zhou, W.: Convergence of the reweighted L_1 minimization algorithm for L_2 - L_p minimization. *Comput. Optim. Appl.* **59**, 47–61 (2014)
18. Cortes, C., Vapnik, V.: Support vector networks. *Mach. Learn.* **20**, 273–297 (1995)
19. Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* **96**, 1348–1360 (2001)
20. Fang, J., Peng, H.: Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Stat.* **32**, 928–961 (2004)
21. Frank, L.E., Friedman, J.H.: A statistical view of some chemometrics regression tools. *Technometrics* **35**, 109–135 (1993)
22. Haeser, G., Liu, H., Ye, Y.: Optimality condition and complexity analysis for linearly-constrained optimization without differentiability on the boundary, *Mathematical Programming*, to appear. <https://doi.org/10.1007/s10107-018-1290-4>
23. Lai, M., Wang, J.: An unconstrained L_q minimization with $0 < q \leq 1$ for sparse solution of underdetermined linear systems. *SIAM J. Optim.* **21**, 82–101 (2010)
24. Liu, Y.F., Ma, S., Dai, Y.H., Zhang, S.: A smoothing SQP framework for a class of composite L_q minimization over polyhedron. *Math. Program.* **158**, 467–500 (2016)
25. Lu, Z.: Iterative reweighted minimization methods for L_p regularized unconstrained nonlinear programming. *Math. Program.* **147**, 277–307 (2014)
26. Martínez, M.J.: On high-order model regularization for constrained optimization. *SIAM J. Optim.* **27**, 2447–2458 (2017)
27. Moré, J.J., Garbow, B.S., Hillstrom, K.E.: Testing unconstrained optimization software. *ACM Trans. Math. Softw.* **7**, 17–41 (1981)
28. Nikolova, M., Ng, M.K., Zhang, S., Ching, W.K.: Efficient reconstruction of piecewise constant images using nonsmooth nonconvex minimization. *SIAM J. Imaging Sci.* **1**, 2–25 (2008)
29. Nocedal, J., Wright, S.J.: Numerical optimization. Springer-Verlag, New York (2006)
30. Vogel, C.R.: Computational methods for inverse problems, Society for industrial and applied mathematics, Philadelphia (2002)
31. Zhang, C.H.: Nearly unbiased variable selection under minimax nonconcave penalty. *Ann. Stat.* **38**, 894–942 (2010)