On the complexity of solving feasibility problems [*]

L. F. Bueno [†] J. M. Martínez [‡]

November 17, 2018

Abstract

We consider feasibility problems defined by a set of constraints that exhibit gradient Hölder continuity plus additional constraints defined by the affordability of obtaining approximate minimizers of quadratic models onto the associated feasible set. Each iteration of the method introduced in this paper involves the approximate minimization of a two-norm regularized quadratic subject to the affordable constraints. We obtain complexity results regarding the number of iterations and evaluations that are necessary to obtain an approximate KKT point for the problem of minimizing the sum of squares of the first type of constraints subject to the second type of constraints. Such results are a consequence of recent developments on the minimization of functions with Hölder-continuity assumptions. In addition, we show that much stronger results may be obtained when a meaningful non-degeneracy assumption is satisfied. Examples indicate that the estimations so far obtained are reliable.

Key words: Complexity, Feasibility problem, First order methods, Reliable estimations.

1 Introduction

Many practical problems require the solution of systems of equations and inequations. For example, socio-economic models usually aim targets with respect to education, housing, erradication of poverty, sustainability, and public health. See, for example, [14]. Moreover, from the practical point of view, constrained optimization problems, multiobjective problems, and equilibrium problems can be frequently reformulated as feasibility problems. Very often, feasibility versions of those problems are much easier to solve than their original counterparts.

In this paper, the equations and inequations that define a feasibility problem will be called *constraints*. We will assume that it is relatively easy to maintain feasibility with respect to a subset of constraints. For example, that subset could involve only bounds on the variables, linear constraints, spherical and ball constraints, intersections of balls and polytopes, and matrices with some peculiarities as idempotency or semidefiniteness. For solving the feasibility problem it is

[*] This work was supported by CEPID-CEMeai FAPESP Grants 2015/02528-8 and 2013/05475-7 and CNPq.

[†] Department of Science and Technology, Federal University of São Paulo, São José dos Campos, SP, Brazil. E-mail: lfelipebueno@gmail.com

[‡] Professor, Department of Applied Mathematics, Institute of Mathematics, Statistics and Scientific Computing, University of Campinas, Campinas, SP, Brazil. E-mail: martinez@ime.unicamp.br

usual to minimize the sum of squares of difficult constraints subject to the fulfillment of easy constraints. The natural stopping criterion for algorithms that proceed in this way involves the values of the difficult constraints, in addition to the optimality conditions of the associated optimization problem.

The problem of minimizing a sum of squares subject to convex constraints was considered in [9] and [10]. In [9] a cubic regularization scheme of ARC-type [7, 8] was employed and complexity results were given, considering stopping criteria based on the residual norm and on the gradient of the residual norm. In [10] the approach of [9] was extended in order to consider arbitrary p-order Taylor approximations of the objective function in the sense of [6]. For technical reasons, the case $p = 1$ (as well as all cases in which p is odd) was not addressed in [10].

In this paper, for solving the feasibility problem, we introduce an algorithm based on the approximate minimization of quadratically-regularized quadratic models subject to generally nonconvex constraints. Only the first derivatives of the quadratic model employed at each iteration will be assumed to agree with the derivatives of the (sum of squares) objective function. The Hessian of the quadratic model will only be assumed to be bounded, opening the possibility of using opportunistic quasi-Newton approximations. Thus, in terms of [9], we will be involved with the case $p = 1$. Moreover, we will be concerned with the case in which the gradient of the sum of squares does not satisfy a Lipschitz condition but only a Hölder condition, as in [11, 13, 15]. Most complexity results will be proved following the lines of [15]. However, if a gradient-domination property holds (see [16]), the method finds a solution with residual norm smaller than ε employing $O(|\log \varepsilon|)$ computer work. A similar bound for the ARC-like method, with convex affordable constraints and using even-order Taylor model approximations has been presented in Theorem 3.6 of [10]. We will verify that, unlike many worst-case complexity results for nonlinear optimization, the $O(|\log \varepsilon|)$ result provides reliable estimations of computer work in practical cases.

Notation

$\| \cdot \|$ will denote the Euclidean norm.

If $v \in \mathbb{R}^n$ is a vector with components v_i, v_+ will be the vector with components $\max\{0, v_i\}$, $i = 1, \ldots, n$.

If $h(x)$ is a vectorial function, we denote $h'(x) = \left(\frac{\partial h_i}{\partial x_j}(x) \right)$.

$\mathbb{R}_+$ will denote the set of nonnegative elements of $\mathbb{R}$.

If v and w are vectors with componentes v_i and w_i, respectively, $\min\{v, w\}$ will denote the vector with components $\min\{v_i, w_i\}$.

2 Main results

The problem considered in this work consists of finding $x \in \mathbb{R}^n$ such that

$$h(x) = 0, g(x) \leq 0, \tag{1}$$

$$h_E(x) = 0, \text{ and } h_I(x) \leq 0, \tag{2}$$

where $h : \mathbb{R}^n \to \mathbb{R}^m$, $g : \mathbb{R}^n \to \mathbb{R}^q$, $h_E : \mathbb{R}^n \to \mathbb{R}^{\underline{m}}$, and $h_I : \mathbb{R}^n \to \mathbb{R}^{\underline{q}}$ have continuous first derivatives for all $x \in \mathbb{R}^n$. The constraints (1) are will be said to be "hard" constraints whereas the constraints (2) will be called "affordable" constraints. We will assume that C_0 is a convex and compact set that contains all the solutions of (2).

We define, for all $x \in \mathbb{R}^n$,

$$\Phi(x) = \frac{1}{2}(\|h(x)\|^2 + \|g(x)_+\|^2). \tag{3}$$

Clearly,

$$\nabla\Phi(x) = h'(x)^T h(x) + g'(x)^T g(x)_+ \tag{4}$$

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

The method for solving (1)–(2) will be iterative. At each iteration k we will consider subproblems of the form:

$$\text{Minimize } \nabla\Phi(x^k)^T(x - x^k) + \frac{1}{2}(x - x^k)^T B_k (x - x^k) + \sigma\|x - x^k\|^2 \tag{5}$$

subject to

$$h_E(x) = 0 \text{ and } h_I(x) \leq 0. \tag{6}$$

The constraints (2) are said to be affordable because we assume that obtaining approximate solutions of (5)–(6) is not hard. This is the case when h_E and h_I are affine functions, when the constraints (6) describe balls or interseccion of balls with polytopes, when they represent matricial properties as positive semidefiniteness or idempotency, and many other situations.

By the continuity of the derivatives of $h(x)$ and $g(x)$, the function $\Phi(x)$ also has continuous first derivatives. However, in general, second derivatives of $\Phi(x)$ do not exist. In this paper we will not assume Lipschitz-continuity of $\nabla\Phi(x)$ either. In fact we will only assume the following Hölder-continuity property:

Assumption A1 *There exist $\beta \in (0, 1]$ and $L > 0$ such that for all $x, y \in C_0$,*

$$\|\nabla\Phi(x) - \nabla\Phi(y)\| \leq L\|x - y\|^\beta. \tag{7}$$

Assumption A1 implies that, for all $x, y \in C_0$,

$$|\Phi(y) - [\Phi(x) + \nabla\Phi(x)^T(y - x)]| \leq L\|x - y\|^{1+\beta}. \tag{8}$$

(See, for example, [5].)

Proposition 2.1 *Assume that Assumption A1 holds and B is a symmetric matrix such that $\|B\| \leq M$, then, for all $x, y \in C_0$,*

$$\Phi(y) \leq \Phi(x) + \nabla\Phi(x)^T(y - x) + \frac{1}{2}(y - x)^T B(y - x) + L_B\|x - y\|^{1+\beta}. \tag{9}$$

and

$$\|\nabla\Phi(x) + B(y - x) - \nabla\Phi(y)\| \leq L_B\|x - y\|^\beta, \tag{10}$$

where

$$L_B = L + M \max\{diam(C_0), 1\} \tag{11}$$

and $diam(C_0)$ is the diameter of the compact set C_0.

Proof By (7),

$$\|\nabla\Phi(x) + B(y-x) - \nabla\Phi(y)\| \le L\|x-y\|^\beta + M\|x-y\|. \tag{12}$$

If $\|x-y\| \le 1$, we have that $\|x-y\| \le \|x-y\|^\beta$ and, so,

$$\|\nabla\Phi(x) + B(y-x) - \nabla\Phi(y)\| \le (L+M)\|x-y\|^\beta.$$

If $\|x-y\| > 1$ we have that $\|x-y\| \le \|x-y\|^\beta\|x-y\| \le \|x-y\|^\beta diam(C_0)$. Therefore, by (12),

$$\|\nabla\Phi(x) + B(y-x) - \nabla\Phi(y)\| \le L\|x-y\|^\beta + Mdiam(C_0)\|x-y\|^\beta,$$

so (10) is proved.

Now, by (8),

$$\Phi(y) \le \Phi(x) + \nabla\Phi(x)^T(y-x) + \frac{1}{2}(y-x)^T B(y-x) + L\|x-y\|^{1+\beta} - \frac{1}{2}(y-x)^T B(y-x)$$

$$\le \Phi(x) + \nabla\Phi(x)^T(y-x) + \frac{1}{2}(y-x)^T B(y-x) + L\|x-y\|^{1+\beta} + M\|y-x\|^2.$$

$$\le \Phi(x) + \nabla\Phi(x)^T(y-x) + \frac{1}{2}(y-x)^T B(y-x) + L\|x-y\|^{1+\beta} + M\|y-x\|^{1+\beta}\|y-x\|^{1-\beta}.$$

If $\|y-x\| \le 1$,

$$\Phi(y) \le \Phi(x) + \nabla\Phi(x)^T(y-x) + \frac{1}{2}(y-x)^T B(y-x) + (L+M)\|x-y\|^{1+\beta}.$$

On the other hand, if $\|y-x\| > 1$,

$$\Phi(y) \le \Phi(x) + \nabla\Phi(x)^T(y-x) + \frac{1}{2}(y-x)^T B(y-x) + (L+Mdiam(C_0))\|x-y\|^{1+\beta}.$$

This completes the proof. □

Algorithm 2.1

Assume that $h_E(x^0) = 0$ and $g(x^0) \le 0$. Initialize $k \leftarrow 0$. Assume that $\varepsilon > 0, \gamma > 0, \alpha \in (0,1), \sigma_{min} > 0$, $\theta > 0$, and $M > 0$.

Step 1 Call Algorithm 2.2 with the objective of obtaining $x \in \mathbb{R}^n$ such that $h_E(x) = 0$, $h_I(x) \le 0$ and at least one of the following two conditions is satisfied:

$$\Phi(x) \le \frac{1}{2}\Phi(x^k); \tag{13}$$

or there exist $\lambda \in \mathbb{R}^m$ and $\mu \in \mathbb{R}^q_+$ such that

$$\|\nabla\Phi(x) + h'_E(x)^T\lambda + h'_I(x)^T\mu\| \le \gamma\sqrt{\Phi(x^k)} \tag{14}$$

with

$$\min\{\mu, -h_I(x)\} = 0. \tag{15}$$

Step 2 If (13) does not hold or $\Phi(x) \leq \varepsilon$, stop the execution of Algorithm 2.1. Otherwise, define $x^{k+1} = x$, update $k \leftarrow k + 1$ and go to Step 1.

Algorithm 2.2 below is Algorithm 2.1 of [15] applied to the minimization of $\Phi(x)$ with the initial point x^k and $p = 1$. We use the target $\Phi(x^k)/2$ and the tolerance for the KKT condition is $\min\{0.99, \gamma\Phi(x^k)\}$. Algorithm 2.2 is called at each iteration of Algorithm 2.1. As a consequence of the results proved in [15], in a finite number of iterations Algorithm 2.2 finds a point with functional value not bigger than the target or finds a point at which the KKT-like approximate condition required is satisfied. In this case, as the KKT condition holds with tolerance $\min\{0.99, \gamma\Phi(x^k)\}$, it also holds with tolerance $\gamma\Phi(x^k)$, as required in (14). The properties of Algorithm 2.2 will be deduced from the properties proved for Algorithm 2.1 in [15].

Assumption A2 below ensures that the subproblem invoked at each iteration of Algorithm 2.2 is solvable. A sufficient condition for this assumption is that all the points that satisfy the affordable constraints also fulfill a constraint qualification that guarantees that the minimizers of quadratics subject to those constraints satisfy KKT conditions.

Assumption A2 *For all $\theta > 0$, B symmetric, $\sigma \geq 0$, and $\bar{x} \in \mathbb{R}^n$ satisfying $h_E(\bar{x}) = 0$ and $h_I(\bar{x}) \leq 0$, there exist $x \in \mathbb{R}^n$, $\lambda \in \mathbb{R}^{\underline{m}}$, and $\mu \in \mathbb{R}^{\underline{q}}_+$ such that*

$$\nabla\Phi(\bar{x})^T(x - \bar{x}) + \frac{1}{2}(x - \bar{x})^T B(x - \bar{x}) + \sigma\|x - \bar{x}\|^2 \leq 0,$$

$$\|\nabla\Phi(\bar{x}) + (B + 2\sigma I)(x - \bar{x}) + h_E'(x)^T\lambda + h_I'(x)^T\mu\| \leq \theta\|x - \bar{x}\|,$$

$$\|h_E(x)\| = 0, \ \textit{and} \ \|\min\{\mu, -h_I(x)\}\| = 0.$$

Algorithm 2.2

Initialize $\ell \leftarrow 0$, $x^{k,0} = x^k$, and $\sigma_0 = \sigma_{min}$.

Step 1. Set $\sigma \leftarrow \sigma_\ell$. Let $B_{k,\ell} \in \mathbb{R}^{n\times n}$ be symmetric with $\|B_{k,\ell}\| \leq M$.

Step 2. Find $x \in \mathbb{R}^n$, $\lambda \in \mathbb{R}^{\underline{m}}$, and $\mu \in \mathbb{R}^{\underline{q}}_+$ such that

$$\nabla\Phi(x^{k,\ell})^T(x - x^{k,\ell}) + \frac{1}{2}(x - x^{k,\ell})^T B_{k,\ell}(x - x^{k,\ell}) + \sigma\|x - x^{k,\ell}\|^2 \leq 0, \tag{16}$$

$$\|\nabla\Phi(x^{k,\ell}) + (B_{k,\ell} + 2\sigma I)(x - x^{k,\ell}) + h_E'(x)^T\lambda + h_I'(x)^T\mu\| \leq \theta\|x - x^{k,\ell}\|, \tag{17}$$

where

$$\lambda \in \mathbb{R}^{\underline{m}},\ \mu \in \mathbb{R}^{\underline{q}}_+,\ \|\min\{\mu, -h_I(x)\}\| = 0, \tag{18}$$

$$\|h_E(x)\| = 0, \text{ and } h_I(x) \leq 0. \tag{19}$$

Step 3. If

$$\Phi(x) \le \Phi(x^k)/2 \tag{20}$$

or

$$\|\nabla\Phi(x) + h_E'(x)^T\lambda + h_I'(x)^T\mu\| \le \min\{0.99, \gamma\sqrt{\Phi(x^k)}\}, \tag{21}$$

return to Algorithm 2.1.

Step 4. Test the sufficient descent condition

$$\Phi(x) \le \Phi(x^{k,\ell}) - \frac{\alpha}{36}\frac{\min\{0.99, \gamma\sqrt{\Phi(x^k)}\}^2}{\sigma}. \tag{22}$$

If (22) does not hold, set $\sigma \leftarrow 2\sigma$ and go to Step 2. Else, continue at Step 5.

Step 5. Define $x^{k,\ell+1} = x$, $\ell \leftarrow \ell + 1$, $\sigma_k = \sigma$, and go to Step 1.

Remarks

- Note that the matrix $B_{k,\ell}$ at each iteration of Algorithm 2.2 is only required to be bounded. Even the null matrix is admissible. This opens the possibility of using problem-oriented quasi-Newton approximations of $\nabla^2\Phi(x)$. The Gauss-Newton approximation $\sum_{i=1}^m \nabla h_i(x)\nabla h_i(x)^T + \sum_{g_i(x)\ge 0}\nabla g_i(x)\nabla g_i(x)^T$ is an interesting alternative due to its positive semidefiniteness and its approximation to $\nabla^2\Phi(x)$ when x is almost feasible. Note that this approximation only involves first-order information of the derivatives of h and g. Other interesting possibilities in the quasi-Newton field are the rank-one correction SR1 and structured quasi-Newton approximations [12].

- Later, we will prove that, if (13) does not hold, then (14) necessarily holds. In this case, dividing both members of (14) by $\sqrt{\Phi(x)}$, we obtain:

$$\left\|\frac{\nabla\Phi(x)}{\sqrt{\Phi(x)}} + h_E'(x)^T\frac{\lambda}{\sqrt{\Phi(x)}} + h_I'(x)^T\frac{\mu}{\sqrt{\Phi(x)}}\right\| \le \gamma\frac{\sqrt{\Phi(x^k)}}{\sqrt{\Phi(x)}}. \tag{23}$$

But, since (13) does not hold, one has that $\sqrt{\Phi(x^k)}/\sqrt{\Phi(x)} \le \sqrt{2}$. Moreover, $\nabla\sqrt{\Phi(x)} = \frac{1}{2}\frac{\nabla\Phi(x)}{\sqrt{\Phi(x)}}$. Therefore, by (23),

$$\left\|2\nabla\sqrt{\Phi(x)} + h_E'(x)^T\frac{\lambda}{\sqrt{\Phi(x)}} + h_I'(x)^T\frac{\mu}{\sqrt{\Phi(x)}}\right\| \le \gamma\sqrt{2}. \tag{24}$$

Thus,

$$\left\|\nabla\sqrt{\Phi(x)} + h_E'(x)^T\frac{\lambda}{2\sqrt{\Phi(x)}} + h_I'(x)^T\frac{\mu}{2\sqrt{\Phi(x)}}\right\| \le \gamma\frac{\sqrt{2}}{2} < \gamma. \tag{25}$$

By (18) and (19), this means that the fulfillment of (14) implies that the point x is a KKT point for the minimization of $\sqrt{\Phi(x)}$ subject to $h(x) = 0$ and $g(x) \le 0$ with tolerance

γ. Cartis, Gould, and Toint [9, 10] addressed the problem (1)–(2) in the case that the constraints (2) are convex by means of ARC-like algorithms. They suggested two different stopping criteria. The first one requires that $\sqrt{\Phi(x)}$ be smaller than a given tolerance. Alternatively, their algorithms also stop when the minimum of $(z-x)^T \nabla \sqrt{\Phi(x)}$, subject to the convex constraints and $\|z-x\| \leq 1$, is, in modulus, smaller than a given tolerance. By (25), in the unconstrained case, our criterion (14) corresponds to the requirement $\|\nabla \sqrt{\Phi(x)}\| \leq \gamma$, as so happens to be with the second criterion of [9, 10]. On the other hand, the criterion [9, 10] for convexly constrained problems requires the minimization of a linear function with convex constraints. If (14) holds, we have, by (25), that

$$\nabla \sqrt{\Phi(x)} + h'_E(x)^T \frac{\lambda}{2\sqrt{\Phi(x)}} + h'_I(x)^T \frac{\mu}{2\sqrt{\Phi(x)}} = v, \tag{26}$$

with $\|v\| \leq \gamma$. Pre-multiplying (26) by $(z-x)^T$, we get:

$$(z-x)^T \nabla \sqrt{\Phi(x)} + (z-x)^T h'_E(x)^T \frac{\lambda}{2\sqrt{\Phi(x)}} + (z-x)^T h'_I(x)^T \frac{\mu}{2\sqrt{\Phi(x)}} = (z-x)^T v. \tag{27}$$

If the constraints (2) define a convex set, the points that satisfy the affordable equality constraints define an affine subspace, and so, by the feasibility of x and z with respect to (2), $(z-x)^T h'_E(x)^T = 0$. Moreover, by the convexity and the complementarity (15), we have that, $\mu_i = 0$ if $(h_I)_i(x) < 0$, and $\nabla (h_I)_i(x)^T (z-x) \leq 0$ if $(h_I)_i(x) = 0$. Therefore,

$$(z-x)^T \nabla \sqrt{\Phi(x)} \geq (z-x)^T v. \tag{28}$$

This inequality holds, in particular, when z is minimizer of $(z-x)^T \nabla \sqrt{\Phi(x)}$ subject to $\|z-x\| \leq 1$ and the affordable constraints. In this is the case, we have that $(z-x)^T \nabla \sqrt{\Phi(x)} \leq 0$, therefore, by (28),

$$|(z-x)^T \nabla \sqrt{\Phi(x)}| \leq -(z-x)^T v \leq \|z-x\| \|v\| \leq \gamma. \tag{29}$$

This means that the fulfillment of (14) in the case that the affordable constraints define a convex set implies the second stopping criterion employed by the ARC-like algorithms defined in [9] and [10]. The reciprocal is not true. Take, for example, the problem defined by $\Phi(x) = (x+1)^2$, $h_E(x) = x^2$, $n = \underline{m} = 1$, and $\underline{q} = 0$. If $x = 0$ it is easy to see that the stopping criterion of [9, 10] holds but (25) does not hold for $\gamma < 1$. This is due to the fact that the constraint $x^2 = 0$, in spite of defining a convex set, does not satisfy any constraint qualification. This means that, essentially, in the case of convex affordable constraints, the criterion (14) is stricter that the analogous one employed in [9] and [10].

Theorem 2.1 will follow as a consequence of Theorem 2.2 of [15]. The idea is to consider that Algorithm 2.2 addresses the minimization of $\Phi(x)$ subject to $h_E(x) = 0$ and $h_I(x) \leq 0$, generating feasible points at every iteration and stopping when the target $\Phi(x^k)/2$ is achieved or when the KKT condition is satisfied with precision $\min\{0.99, \gamma\sqrt{\Phi(x^k)}\}$. In this case, thanks to the fulfillment of Assumption A1, Algorithm 2.2 is a particular case of Algorithm 2.1 of [15] for any value of $\delta \geq 0$ and $p = 1$.

Theorem 2.1 *Assume that A1 and A2 hold. Then, there exists $c_p > 0$, only dependent of $\alpha, \beta, \theta, L, M,$ and $diam(C_0)$, such that, after at most*

$$\frac{\Phi(x^k)}{2} \frac{\left(\min\{0.99, \gamma\sqrt{\Phi(x^k)}\}\right)^{-\frac{1+\beta}{\beta}}}{\alpha c_p} \tag{30}$$

iterations of Algorithm 2.2, $x \in \mathbb{R}^n$, $\lambda \in \mathbb{R}^{\underline{m}}$, and $\mu \in \mathbb{R}^q_+$ are computed verifying

$$\|h_E(x)\| = 0,\ \|h_I(x)_+\| \leq 0,\ \textit{and}\ \Phi(x) \leq \Phi(x^k)/2 \tag{31}$$

or

$$\|\nabla\Phi(x) + h'_E(x)^T\lambda + h'_I(x)^T\mu\| \leq \min\{0.99, \gamma\sqrt{\Phi(x^k)}\}, \tag{32}$$

$$\|h_E(x)\| = 0,\ \|h_I(x)_+\| \leq 0,\ \textit{and}\ \|\min\{\mu, -h_I(x)\}\| = 0. \tag{33}$$

Proof. The proof follows from Theorem 2.2 of [15]. The function Φ corresponds to the function f of [15]. The condition (4) of [15] is our condition (16) with $p = 1$. The condition (5) of [15] is our condition (17) and the conditions (6) and (7) of [15] are (18) and (19) of the present paper with $\delta = 0$. The Hölder-like conditions (2) and (3) of [15] are proved in Proposition 2.1. The sufficient descent condition (12) of [15] is our condition (22). The value of f_{target} is $\frac{1}{2}\Phi(x^k)$ in our case, and c_p is given by formula (26) of [15], with constant L in [15] replaced with L_B given by (11). □

The following corollary corresponds to the case in which Assumption A1 holds with $\beta = 1$, that is, $\nabla\Phi$ satisfies a Lipschitz condition.

Corollary 2.1 *Assume that A1 holds with $\beta = 1$ and A2 also holds. Then, there exists $c_p > 0$, only dependent of $\theta, L, M, diam(C_0)$, and α, such that, after at most*

$$\frac{1}{2\alpha c_p} \max\{1.03\Phi(x^0), 1/\gamma^2\} \tag{34}$$

iterations of Algorithm 2.2, $x \in \mathbb{R}^n$, $\lambda \in \mathbb{R}^{\underline{m}}$, and $\mu \in \mathbb{R}^q_+$ are computed verifying (31) or (32)–(33).

Proof Straightforward, from Theorem 2.1. □

Theorem 2.2 *Assume that A1 and A2 hold. Then, the number of evaluations of Φ and its derivatives at each call of Algorithm 2.2 is bounded above by*

$$\frac{\Phi(x^k)}{2} \frac{\left(\min\{0.99, \gamma\sqrt{\Phi(x^k)}\}\right)^{-\frac{1+\beta}{\beta}}}{\alpha c_p} + \frac{1-\beta}{\beta}|\log_2(\gamma\sqrt{\varepsilon})| + c_a. \tag{35}$$

where c_p is given in Theorem 2.1 and c_a only depends of $\alpha, \beta, \theta, \sigma_{min}, L, M,$ and $diam(C_0)$.

Proof By Theorem 2.3 of [15] with the definitions given in the proof of Theorem 2.1 above, at each call of Algorithm 2.2, a maximum of

$$\frac{\Phi(x^k)}{2}\frac{(\min\{0.99, \gamma\sqrt{\Phi(x^k)}\})^{-\frac{1+\beta}{\beta}}}{\alpha c_p} \tag{36}$$

iterations are performed and the maximum number of function and gradient evaluations of Φ is given by (36) plus

$$\max\left\{\log_2(\theta), \left(\frac{1-\beta}{\beta}\log_2\left(\left(\min\left\{0.99, \gamma\sqrt{\Phi(x^k)}\right\}\right)^{-1}\right) + c_e\right)\right\} + |\log_2(\sigma_{min})| + 1. \tag{37}$$

where c_p is given in Theorem 2.1 and c_e only depends of α, β, L, M, and $diam(C_0)$. Let us define

$$c_a = c_e + |\log_2(\theta)| + \frac{1-\beta}{\beta}\log_2(0.99^{-1}) + |\log_2(\sigma_{min})| + 1. \tag{38}$$

Then, (35) follows from straightforward calculations using that, before termination of Algorithm 2.1, one has that $\Phi(x^k) > \varepsilon$. □

Corollary 2.2 *Assume that A1 holds with $\beta = 1$ and A2 holds. Then, the number of evaluations of Φ and its derivatives at each call of Algorithm 2.2 is bounded above by*

$$\frac{1}{2\alpha c_p}\max\{1.03\Phi(x^0), 1/\gamma^2\} + c_a$$

where c_a, defined in (38), only depends on $\alpha, \beta, \theta, \sigma_{min}, L, M,$ and $diam(C_0)$.

Proof As in the case of Corollary 2.1, the thesis follows from Theorem 2.2 taking $\beta = 1$. □

Theorem 2.3 *Assume that A1 and A2 hold. Then, the total number of iterations of Algorithm 2.2 at the execution of Algorithm 2.1 is bounded above by*

$$c|\log_2(\varepsilon/\Phi(x^0))|\max\left\{\Phi(x^0)0.99^{-\frac{1+\beta}{\beta}}, \gamma^{-\frac{1+\beta}{\beta}}\varepsilon^{\frac{\beta-1}{2\beta}}\right\}, \tag{39}$$

where c only depends on $\alpha, \beta, \theta, L, M,$ and $diam(C_0)$.

Moreover, the total number of evaluations of Φ and its derivatives is bounded above by:

$$|\log_2(\varepsilon/\Phi(x^0))|\left(c\max\left\{\Phi(x^0)0.99^{-\frac{1+\beta}{\beta}}, \gamma^{-\frac{1+\beta}{\beta}}\varepsilon^{\frac{\beta-1}{2\beta}}\right\} + \frac{1-\beta}{\beta}|\log_2(\gamma\sqrt{\varepsilon})| + c_a\right) \tag{40}$$

where c_a only depends on $\alpha, \beta, \theta, \sigma_{min}, L, M,$ and $diam(C_0)$.

Proof By Theorem 2.2, at each call of Algorithm 2.2, the maximal number of iterations is

$$\frac{1}{2\alpha c_p}\max\{\Phi(x^0)0.99^{-\frac{1+\beta}{\beta}},\gamma^{-\frac{1+\beta}{\beta}}\varepsilon^{\frac{\beta-1}{2\beta}}\}.$$

Therefore, the total number of iterations of Algorithm 2.2 cannot exceed this bound times the number of calls of Algorithm 2.2. By (13) and the stopping criterion $\Phi(x) \leq \varepsilon$, Algorithm 2.2 cannot be called more than $|\log_2 \varepsilon/\Phi(x^0)|$ times. Defining $c = \frac{1}{2\alpha c_p}$, this completes the proof of (39).

The bound (40) follows from the number of calls of Algorithm 2.2 times the number of evaluations of Φ and its derivatives, which is bounded by (35). □

Corollary 2.3 *If Assumption A2 holds and A1 holds with $\beta = 1$, the total number of iterations of Algorithm 2.2 during an execution of Algorithm 2.1 is bounded above by*

$$c|\log_2(\varepsilon/\Phi(x^0))|\max\{1.03\Phi(x^0),1/\gamma^2\}$$

and the number of evaluations of Φ and its derivatives is bounded above by

$$|\log_2(\varepsilon/\Phi(x^0))|\left[c\max\{1.03\Phi(x^0),1/\gamma^2\}+c_a\right].$$

Let us examine the result of Theorem 2.3 under the light of a new assumption A3.

Assumption A3 *There exists $\kappa > 0$ such that, for all $x \in \mathbb{R}^n$ such that $h_E(x) = 0$, $h_I(x) \leq 0$ and $\Phi(x) > 0$, if $\lambda \in \mathbb{R}^m$ and $\mu \in \mathbb{R}^q_+$ is such that* $\min\{\mu, -h_I(x)\} = 0$*, we have that*

$$\left\|\frac{\nabla\Phi(x)}{\sqrt{\Phi(x)}} + h'_E(x)^T\lambda + h'_I(x)^T\mu\right\| \geq \kappa. \tag{41}$$

Since $\nabla\sqrt{\Phi(x)} = \frac{1}{2}\frac{\nabla\Phi(x)}{\sqrt{\Phi(x)}}$, Assumption A3 means that, if x is a feasible point with respect to (2), but not with respect to (1), the 2-norm of the infeasibility with respect to (1) is bounded away from satisfying KKT conditions with the constraints (2). Thus, these KKT conditions can be satisfied only when x is feasible both with respect to (2) and (1). This implies that the problem of solving (1)–(2) is not very hard, since stationary points, that could be attractive for minimization algorithms, are necessarily solutions of the problem. In the case that the set of affordable constraints (2) is empty, (41) represents an uniform regularity assumption of the constraints (1).

If $f : \mathbb{R}^n \to \mathbb{R}$ admits a global minimizer at x^* and we define $\Phi(x) = f(x) - f(x^*)$ for all $x \in \mathbb{R}^n$ we obviously have that x^* is a global minimizer of $\sqrt{f(x) - f(x^*)}$. The condition 41, in this case, says that, for all x such that $f(x) \neq f(x^*)$, the norm of the gradient of $\sqrt{f(x) - f(x^*)}$ is not smaller than 2κ. This excludes the existence of local minimizers different from x^*.

An immediate consequence of (41) is given in the following Lemma.

Lemma 2.1 *Assume that A3 holds. Then, for all $x \in \mathbb{R}^n$ such that $h_E(x) = 0$, $h_I(x) \leq 0$, if $\lambda \in \mathbb{R}^m$ and $\mu \in \mathbb{R}^q_+$ is such that* $\min\{\mu, -h_I(x)\} = 0$, *we have that*

$$\left\| \nabla \Phi(x) + h'_E(x)^T \lambda + h'_I(x)^T \mu \right\| \geq \kappa \sqrt{\Phi(x)}. \tag{42}$$

Proof The result is trivial if $\Phi(x) = 0$. Otherwise, assume, by contradiction that (42) does not hold. Then, there exist $\lambda \in \mathbb{R}^m$ and $\mu \in \mathbb{R}^q_+$ is such that $\min\{\mu, -h_I(x)\} = 0$ and

$$\left\| \nabla \Phi(x) + h'_E(x)^T \lambda + h'_I(x)^T \mu \right\| < \kappa \sqrt{\Phi(x)}. \tag{43}$$

Dividing both members of (43) by $\sqrt{\Phi(x)}$ we obtain the negation of Assumption A3 with the multipliers $\lambda / \sqrt{\Phi(x)}$ and $\mu / \sqrt{\Phi(x)}$. □

Let us now analyze the consequences of Assumption A3 for Algorithms 2.1 and 2.2.

The consequence of Lemma 2.1 is that, under Assumption A3, if γ is smaller than $\kappa / \sqrt{2}$, (14) will never hold.

Lemma 2.2 *Assume that A1, A2, and A3 hold and that, in Algorithm 2.1, one chooses $\gamma \leq \kappa / \sqrt{2}$. Then, condition (14) only holds when (13) holds. Analogously, in Algorithm 2.2, (21) only holds if (20) holds.*

Proof If (14) holds we have that there exist $\lambda \in \mathbb{R}^m$ and $\mu \in \mathbb{R}^q_+$ such that

$$\| \nabla \Phi(x) + h'_E(x)^T \lambda + h'_I(x)^T \mu \| \leq \gamma \sqrt{\Phi(x^k)}, \tag{44}$$

with

$$\min\{\mu, -h_I(x)\} = 0. \tag{45}$$

Assume, by contradiction, that (13) does not hold. Then, $\Phi(x) > \frac{1}{2} \Phi(x^k)$. Thus, $\sqrt{\Phi(x^k)} < \sqrt{2} \sqrt{\Phi(x)}$. Then, by (44),

$$\| \nabla \Phi(x) + h'_E(x)^T \lambda + h'_I(x)^T \mu \| \leq \gamma \sqrt{2} \sqrt{\Phi(x)} < \kappa \sqrt{\Phi(x)}. \tag{46}$$

This contradicts (42). □

Lemma 2.2 has a clear algorithmic consequence. Namely, if Assumption A3 holds we may skip the test (21) in Algorithm 2.2 and the alternative (14)–(15) in Algorithm 2.1 because those criteria would be reached only if (13) holds. As a consequence, the following theorem follows.

Theorem 2.4 *Suppose that Assumption A1, A2, and A3 hold, and that, in Algorithm 2.1, one chooses $\gamma \leq \kappa / \sqrt{2}$. Then, the total number of iterations of Algorithm 2.2 at the execution of Algorithm 2.1 is bounded above by*

$$c |\log_2(\varepsilon / \Phi(x^0))| \max \left\{ \Phi\left(x^0\right) 0.99^{-\frac{1+\beta}{\beta}}, \left(\frac{\kappa}{\sqrt{2}} \right)^{-\frac{1+\beta}{\beta}} \varepsilon^{\frac{\beta-1}{2\beta}} \right\} \tag{47}$$

where c only depends on $\alpha, \beta, \theta, L, M,$ and $diam(C_0)$.

Moreover, the total number of evaluations of Φ and its derivatives is bounded above by:

$$|\log_2(\varepsilon/\Phi(x^0))| \left(c \max\left\{ \Phi(x^0)0.99^{-\frac{1+\beta}{\beta}}, \left(\frac{\kappa}{\sqrt{2}}\right)^{-\frac{1+\beta}{\beta}} \varepsilon^{\frac{\beta-1}{2\beta}} \right\} + \frac{1-\beta}{\beta} \left|\log_2\left(\frac{\kappa\sqrt{\varepsilon}}{\sqrt{2}}\right)\right| + c_a \right) \quad (48)$$

where c_a, defined in (38)*, only depends on $\alpha, \beta, \theta, \sigma_{min}, L, M,$ and $diam(C_0)$.*

Proof By Lemma 2.2, running Algorithms 2.1 and 2.2 with $\gamma \leq \kappa/\sqrt{2}$ produces the same effect as skipping the test (21) and the alternative (14)–(15). Therefore, the obtained results are the same as considering $\gamma = \kappa/\sqrt{2}$. So the thesis follows from Theorem 2.3. □

The following results concern the case in which, not only Assumption A3 holds, but also the gradient $\nabla\Phi(x)$ is Lipschitz-continuous. This means that (7) holds with $\beta = 1$.

Corollary 2.1 *Under the hypotheses of Theorem 2.4, if A1 holds with $\beta = 1$, the total number of iterations of Algorithm 2.2 at the execution of Algorithm 2.1 is bounded above by*

$$c_b |\log_2(\varepsilon/\Phi(x^0))|, \quad (49)$$

where c_b only depends on $\Phi(x^0), \alpha, \beta, \theta, \kappa, L, M,$ and $diam(C_0)$.

Moreover, the total number of evaluations of Φ and its derivatives is bounded above by:

$$c_c |\log_2(\varepsilon/\Phi(x^0))| \quad (50)$$

where c_c only depends on $\Phi(x^0), \alpha, \beta, \theta, \sigma_{min}, \kappa, L, M,$ and $diam(C_0)$.

It is easy to see that the complexity bound given by Corollary 2.1 is sharp. For that purpose, consider $\Phi(x) = x^2$ (so $\nabla\Phi(x) = 2x$), $\sigma = 2$, $B_k = 0$ for all $k \in \mathbb{N}$, and x^0 arbitrary. Given $x^k \in \mathbb{R}$, $x^{k,0} = x^k$, defining $x = x^{k,0} - \frac{1}{2\sigma}\nabla\Phi(x^{k,0})$, we have that x satifies (16) and (17), therefore $x^{k,1} = x$. But $x^{k,1} = x^k/2$, therefore, by the definition of Φ, (20) holds. This implies that, $x^{k+1} = x^k/2$ and $\Phi(x^{k+1}) = \Phi(x^k)/4$ for all $k \in \mathbb{N}$. Then, for all $k \in \mathbb{N}$, $\Phi(x^k) = \Phi(x^0)/4^k$. So, $\Phi(x^k) \leq \varepsilon$ if and only if $\Phi(x^0)/4^k \leq \varepsilon$. Equivalently $4^k \geq \Phi(x^0)/\varepsilon$. Thus, taking logarithms, $k \geq \frac{1}{2}\log_2(\Phi(x^0)/\varepsilon) = \frac{1}{2}|\log_2(\varepsilon/\Phi(x^0))|$.

2.1 Examples

- If $\Phi(x)$ is the popular Rosenbrock function $100(x_2 - x_1^2)^2 + (x_1 - 1)^2$ and the affordable constraints (2) are $-2 \leq x_i \leq 2$, $i = 1, 2$, Assumption A3 holds with $\kappa = 0.66$.

  We employed Algorithm 2.1 for minimizing $\Phi(x)$ with $x_1, x_2 \in [-2, 2]$ using $B_{k,\ell}$ as the Barzilai-Borwein-Raydan diagonal estimation of the Hessian as in [2, 4, 17] and $\varepsilon = 10, 1, 10^{-1}, \ldots, 10^{-17}$. The correlation between the values of $\log(\varepsilon)$ and the number of iterations performed by Algorithm 2.2 was 0.962, corroborating that the logarithmic estimation is reliable.

- Let A be symmetric and positive definite and $\Phi(x) = \frac{1}{2}x^T Ax + b^T x$. Therefore, $\Phi(x) = \frac{1}{2}(x - \bar{x})^T A(x - \bar{x})$, where $\bar{x} = -A^{-1}b$. Thus, $\nabla\Phi(x) = A(x - \bar{x})$ and $\|\nabla\Phi(x)\| = \sqrt{(x - \bar{x})^T A^T A(x - \bar{x})}$. Then, for all $x \in \mathbb{R}^n$

  $$\|\nabla\Phi(x)\| \geq \lambda_{min}\|x - \bar{x}\|$$

  and

  $$\sqrt{\Phi(x))} \leq \lambda_{max}\|x - \bar{x}\|,$$

  where λ_{min} and λ_{max} are the smallest and the biggest eigenvalues of A, respectively. Thus, for all $x \neq \bar{x}$,

  $$\frac{\|\nabla\Phi(x)\|}{\sqrt{\Phi(x)}} \geq \frac{\lambda_{min}}{\lambda_{max}}.$$

  Therefore, in the problem of minimizing $\Phi(x)$ without affordable constraints the value of κ is the inverse of the condition number of A.

  We solved this problem being A the diagonal matrix with eigenvalues $1/2, 1/3, \ldots, n/(n+1)$ employing Algorithm 2.1 as in the Rosenbrock case. The correlation between $\log(\varepsilon)$ and number of iterations of Algorithm 2.2 was 0.99 when $n = 10$.

3 Final remarks

The objective of Optimization is to find a point in the feasible region at which the objective function takes a value as small as possible. Optimality conditions, which relate the local variation of the objective function with the local variations of the constraints, are tools for recognizing that a point is close to a solution or not, but do not have an intrinsic value for most users. Lagrange multipliers are generally used to estimate the variation of the minimum with respect to the variation of different constraints but this utility is challenged in the case that constraint qualifications do not hold or when the set of Lagrange multipliers is infinity. As a matter of fact, the variation of the minimum with respect to constraints is more reliably estimated by means of running the solver with the desired modification of constraints, that does not need to be small. This is the reason why the default version of many constrained optimization solvers, after satisfying a (successful or unsucessful) stopping criterion at some iterate x, try to find a very accurate point in the feasible region, starting with x as initial approximation. Alternatively, these solvers address the problem of finding a feasible point subject to the additional feasibility constraint $f(z) \leq f(x)$.

Constrained optimization problems are usually formulated in the form

$$\text{Minimize } f(x) \tag{51}$$

subject to

$$h(x) = 0, g(x) \leq 0, h_E(x) = 0, h_I(x) = 0, \tag{52}$$

where the constraints $h_E(x) = 0, h_I(x) = 0$ are affordable in the sense discussed in this paper. Augmented Lagrangian (AL) methods are appropriate for these formulations. At each outer iteration of an AL method the augmented Lagrangian function, which combines f, h, and g, is approximately minimized subject to the affordable constraints. See, for example, [3] and [1], where this approach is developed and analyzed. The complexity of solving each subproblem by means or regularization methods is similar to the complexity of solving unconstrained optimization problems. The difficulty of extending this result to the whole constrained minimization process relies on the fact that, in the worst situation, penalty parameters could grow indefinitely, affecting the Lipschitz constants associated to each subproblem. However, in many practical cases, users do not need to "minimize" $f(x)$ and would be happy after finding a feasible point for which $f(x)$ is smaller than a given target. In this case, the requirement (51) may be replaced by the inequality $f(x) - f_{target} \leq 0$, so far defining a feasibility problem together with the constraints (52). In this paper we showed that solving this feasibility problem may be overwhelmingly easier than solving (51)–(52) by means of constrained optimization solvers. This suggests that the obvious idea of reducing (51)–(52) to a sequence of feasibility problems, with the approximation of the "correct" f_{target} determined by means of bissection, could be plausible. It is necessary to exploit variations of this idea, aiming to discover good methods both from the complexity point of view and practical performance.

References

[1] R. Andreani, E. G. Birgin, J. M. Martínez, and M. L. Schuverdt, On augmented Lagrangian methods with general lower-level constraints, *SIAM Journal on Optimization* 18, pp. 1286–1309, 2007.

[2] J. Barzilai and J. M. Borwein, Two point step size gradient methods, *IMA Journal of Numerical Analysis* 8, pp. 141–148, 1988.

[3] E. G. Birgin and J. M. Martínez, Practical Augmented Lagrangian Methods for Constrained Optimization, SIAM Publications, Series: Fundamentals of Algorithms, Philadelphia, 2014.

[4] E. G. Birgin, J. M. Martínez, and M. Raydan, Nonmonotone spectral projected gradient methods on convex sets, *SIAM Journal on Optimization* 10, pp. 1196–1211, 2000.

[5] D. P. Bertsekas, Nonlinear Programming, Athena Scientific, Belmont, Massachussets, USA, 2nd Edition, 1999.

[6] E. G. Birgin, J. L. Gardenghi, J. M. Martínez, S. A. Santos, and Ph. L. Toint, Worst-case evaluation complexity for unconstrained nonlinear optimization using high-order regularized models, *Mathematical Programming* 163, pp. 359-368, 2017 (DOI: 10.1007/s10107-016-1065-8).

[7] C. Cartis, N. I. M. Gould, and Ph. L. Toint, Adaptive cubic regularization methods for unconstrained optimization. Part I: motivation motivation, convergence and numerical results, *Mathematical Programming* 127, pp. 245–295, 2011.

[8] C. Cartis, N. I. M. Gould, and Ph. L. Toint, Adaptive cubic regularization methods for unconstrained optimization. Part II: worst-case function and derivative complexity, *Mathematical Programming* 130, pp. 295–319, 2011.

[9] C. Cartis, N. I. M. Gould, and Ph. L. Toint, On the evaluation complexity of cubic regularization methods for potentially rank-deficient nonlinear least-squares problems and its relevance to constrained nonlinear optimization, SIAM Journal on Optimization 23, pp. 1553–1574, 2013.

[10] C. Cartis, N. I. M. Gould, and Ph. L. Toint, Improved worst-case evaluation complexity for potentially rank-deficient nonlinear least-Euclidean-norm problems using higher-order regularized models, Report NAXYS-12-2015, Namur Center for Complex Systems, University of Namur, Belgium, 2015.

[11] C. Cartis, N. I. M. Gould, and Ph. L. Toint, Universal regularization methods - varying the power, the smoothness and the accuracy, Preprint RAL-P-2016-010, Rutherford Appleton Laboratory, Chilton, England, 2016.

[12] J. E. Dennis and R. S. Schnabel, Numerical methods for unconstrained optimization and nonlinear equations, Prentice Hall, New Jersey, 1983.

[13] G. N. Grapiglia and Yu. Nesterov, Regularized Newton methods for minimizing functions with Hölder continuous Hessians, SIAM Journal on Optimization, 27 (2017), pp. 478–506.

[14] A. O. Herrera, H. D. Scolnik, G. Chichilnisky, G. C. Gallopin, J. E. Hardoy, *Catastrophe or New Society? A Latin America world model*, IDRC, Ottawa, ON, CA, 1976.

[15] J. M. Martínez, On high-order model regularization for constrained optimization, SIAM Journal on Optimization, 27 (2017), pp. 2447–2458.

[16] Yu. Nesterov and B. T. Polyak, Cubic regularization of Newton method and its global performance, Mathematical Programming 108 (2006), pp. 177–205.

[17] M. Raydan, The Barzilai and Borwein gradient method for the large scale unconstrained minimization problem, *SIAM Journal on Optimization* 7, pp. 26–33, 1997.