Operations Research Forum (2024) 5:41
https://doi.org/10.1007/s43069-024-00323-x

RESEARCH

On the Convergence of Inexact Alternate Minimization in Problems with ℓ_0 Penalties

Matteo Lapucci[1] · **Alessio Sortino**[1]

Received: 15 January 2024 / Accepted: 9 April 2024

Abstract
In this work, we consider unconstrained nonlinear optimization problems where the objective function presents a penalty term on the cardinality of a subset of the variables vector; specifically, we prove that an alternate minimization scheme has global asymptotic convergence guarantees towards points satisfying first-order optimality conditions, even when the optimization step with respect to one of the blocks of variables is inexact and without introducing proximal terms. This result, supported by numerical evidence, justifies the use of pure alternate minimization in applications, even in absence of convexity assumptions.

Keywords Sparse optimization · Block coordinate descent · Global convergence · Optimality conditions

Mathematics Subject Classification (2020) 90C26 · 90C30

1 Introduction

In this paper, we are interested in optimization problems of the form

$$\min_{x\in\mathbb{R}^n, y\in\mathbb{R}^m} f(x, y) + \eta\|y\|_0 \tag{1}$$

where $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is a bounded below, continuously differentiable, possibly nonconvex function, whereas $\|\cdot\|_0$ denotes the ℓ_0 pseudo-norm, i.e., $\|y\|_0 = |\{i \mid y_i \neq 0\}|$. Note that the ℓ_0 term makes the overall objective function not only nonconvex and not differentiable, but also discontinuous (lower semi-continuity is the only regularity property that can be stated). Problems of this form can commonly arise, for example, in the context of *compressed sensing* [1, 2]. Moreover, sparsity-inducing

✉ Matteo Lapucci
matteo.lapucci@unifi.it

[1] DINFO, Università di Firenze, Via di Santa Marta 3, 50139 Florence, Italy

Published online: 29 April 2024

regularizes are often introduced in *statistics and machine learning* to improve the generalization performance of models [3, 4] or to carry out best subset selection tasks [5–7]. We refer the readers to [8] for a thorough survey about sparse optimization problems and related algorithmic approaches, as the discussion of the vast literature on this topic goes far beyond the scope of this short paper.

In this work, we focus on the cases where the objective function can effectively be minimized up to global optimality w.r.t. the y block of variables, while keeping x fixed. This exact minimization may be doable either in closed form or by suitable reformulations and algorithmic schemes. In both cases, a two-block coordinate descent approach [9, 10], also referred to as *Alternate Minimization* (AM), becomes particularly appealing to tackle problem (1). As we remark more in detail later in this manuscript, there are recurrent application settings where the ℓ_0 term becomes easy to handle if it is decoupled from the nonlinear part of the objective. Lu and Zhang [11] even defined a Penalty Decomposition scheme for general ℓ_0-penalized optimization problems, where a sequence of subproblems of the form (1) is solved by exact alternate minimization.

Unfortunately, without convexity assumptions on f w.r.t. the x variables, the global optimization step for the x-update in the exact alternate minimization scheme is unreasonable. The contribution, of theoretical nature, of this paper is thus to prove that an *inexact version* of the AM algorithm, similarly as in the continuously differentiable case, converges to points satisfying the same optimality conditions as its exact counterpart, justifying the application of the approach in many practical cases. Differently from other relevant works from the literature [12], we are able to obtain the result without resorting to the introduction of a proximal term. We furthermore show that this theoretical property can actually be observed computationally.

The rest of the paper is organized as follows: in Sect. 2 we state and describe the considered algorithmic scheme; in Sect. 3 we provide the theoretical convergence analysis for the inexact AM algorithm applied to problem (1); then, in Sect. 4 we list some application settings fitting the proposed theoretical framework; computational experiments conducted on these classes of problems are finally shown in Sect. 5, which corroborate our theoretical analysis. The paper ends with some concluding remarks in Sect. 6.

2 Inexact Alternate Minimization

The *block coordinate descent* (BCD) algorithm is a well-established decomposition scheme for optimization problems [9]; the method consists in solving a sequence of subproblems, minimizing at each iteration the objective function with respect to a subset, or block, of variables. The standard, *exact*, version of the BCD scheme requires to solve each subproblem up to global optimality in order to prove convergence to stationary points of the overall, original problem.

However, in the continuously differentiable case, convergence can still be proved even if optimization w.r.t. to one of the blocks of variables is carried out in an inexact fashion; in other words, a suitable step along a gradient-related direction is sufficient to guarantee convergence to stationary points [10].

The extension of the *inexact alternate minimization* scheme to problems of the form (1), with two blocks of variables and irregularity in the objective function, is reported in Algorithm 1.

Algorithm 1 `Inexact Alternate Minimization`

Input: $x^0 \in \mathbb{R}^n$, $y^0 \in \mathbb{R}^m$, $\beta \in (0,1)$, $\gamma \in (0,1)$
1: **for** $k = 0, 1, \ldots$ **do**
2: Compute

$$\alpha^k = \max_{j=0,1,\ldots} \{\beta^j \mid f(x^k - \beta^j \nabla_x f(x^k, y^k), y^k) \leq f(x^k, y^k) - \gamma \beta^j \|\nabla_x f(x^k, y^k)\|^2\}$$

3: Set $x^{k+1} = x^k - \alpha^k \nabla_x f(x^k, y^k)$
4: Compute $y^{k+1} \in \arg\min_y f(x^{k+1}, y) + \eta \|y\|_0$
5: **end for**

Basically, Algorithm 1 alternates exact minimization steps w.r.t. to y variables and descent steps along x variables. For the sake of simplicity, in this paper we assume that the inexact minimization w.r.t. the continuous function $f(\cdot, y^k)$ is carried out by performing a step along the negative gradients $-\nabla_x f(x, y^k)$ with a step size selected by the Armijo rule [13].

In fact, the steepest descent direction $-\nabla_x f(x, y^k)$ could be replaced by any gradient-related direction d_k [13]. Moreover, a number of descent steps could be consecutively performed, instead of a single one. It can be easily seen that such modifications do not spoil the theoretical analysis we are going to carry out in the following section, while they may bring significant benefits from a computational perspective.

The main benefit of Algorithm 1 w.r.t. its exact counterpart lies in the fact that convexity assumptions on f w.r.t. the x variables are no more required for the algorithm to be actually employable. In the next section we formally prove this last claim.

Remark 1 A convergent, inexact alternate minimization algorithm has been devised for a general class of problems, also covering the case of problem (1), in [12]. The convergence results obtained by the so-called Proximal Alternating Linearized Minimization (PALM) algorithm can be guaranteed under reasonable assumptions, even if inexact updates are carried out for both blocks of variables. However, this is possible thanks to the introduction of a proximal point term within the subproblems. Here, we show that if we are able to perform an exact update on just one of the two blocks, we can remove the proximal term altogether, improving numerical performance, while also simplifying the conceptual scheme and the implementation of the algorithm.

3 Convergence Analysis

We now prove the convergence of Algorithm 1 to points satisfying first-order necessary optimality conditions. First, we explicitly state the first-order necessary optimality conditions for problem (1) [6, 11]:

Lemma 1 *Let* $(x^\star, y^\star) \in \mathbb{R}^n \times \mathbb{R}^m$ *be an optimal solution of problem* (1)*. Then,* $(x^\star, y^\star)$ *satisfies the following properties:*

$$\nabla_x f(x^\star, y^\star) = 0, \qquad \frac{\partial f(x^\star, y^\star)}{\partial y_i} = 0 \ \forall i : y_i^\star \neq 0. \tag{2}$$

Proof Let $(x^\star, y^\star)$ be an optimal solution for problem (1). We now prove the two conditions separately.

Let $\xi_i = \frac{\partial f(x^\star, y^\star)}{\partial y_i}$ for all $i = 1, \ldots, m$. Assume by contradiction that there exists $j \in \{1, \ldots, m\}$ such that $y_j^\star \neq 0$ and $\xi_j \neq 0$.

Denote by e_j the j-th unit vector in $\mathbb{R}^m$. Since $-\xi_j e_j$ is a descent direction for f at $(x^\star, y^\star)$, there exists $\bar{t}$ such that $f(x^\star, y^\star - t\xi_j e_j) < f(x^\star, y^\star)$ for all $t \in (0, \bar{t}\,]$.

On the other hand, by continuity, if $|y_j^\star| > 0$ then there exists $\bar{\rho} > 0$ such that $|y_j^\star + \rho| > 0$ for all $\rho \in [-\bar{\rho}, \bar{\rho}]$.

Hence, for all $t \in (0, \min\{\bar{t}, \bar{\rho}/|\xi_j|\}]$ we have

$$\begin{aligned} f(x^\star, y^\star - t\xi_j e_j) + \eta\|y^\star - t\xi_j e_j\|_0 & \\ = f(x^\star, y^\star - t\xi_j e_j) + \eta\|y^\star\|_0 & \\ < f(x^\star, y^\star) + \eta\|y^\star\|_0, & \end{aligned}$$

which is absurd being $(x^\star, y^\star)$ optimal for problem (1).

Next, assume instead that $\nabla_x f(x^\star, y^\star) \neq 0$. Then for t sufficiently small we have

$$f(x^\star - t\nabla_x f(x^\star, y^\star), y^\star) < f(x^\star, y^\star)$$

and hence

$$f(x^\star - t\nabla_x f(x^\star, y^\star), y^\star) + \eta\|y^\star\|_0 < f(x^\star, y^\star) + \eta\|y^\star\|_0,$$

which is again absurd being $(x^\star, y^\star)$ optimal. □

We also state a suitable assumption on function f that we will consider satisfied throughout the rest of this section. Note that we do not need assumptions, e.g., Lipschitz continuity, on the gradient $\nabla f(x)$ or the partial gradients $\nabla_x f(x, y)$ and $\nabla_y f(x, y)$.

Assumption 1 Function f is coercive on $\mathbb{R}^n \times \mathbb{R}^m$, i.e., for any sequence $\{x^t, y^t\}$ such that $\|(x^t, y^t)\| \to \infty$ we have $f(x^t, y^t) \to \infty$.

Then, we recall some well-known properties for the Armijo-type line search, later used in the convergence analysis. These results can be deduced, for instance, by [14]. Firstly, it can be easily seen that step 2 of Algorithm 1 is well defined, i.e., there exists a finite integer j such that β^j satisfies Armijo acceptability condition. Moreover the following result holds.

Lemma 2 *Let $g : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ be a continuously differentiable function and $\{x^t, y^t\} \subseteq \mathbb{R}^n \times \mathbb{R}^m$. Let $T \subseteq \{0, 1, \dots, \}$ be an infinite subset such that*

$$\lim_{\substack{t\to\infty \\ t\in T}} (x^t, y^t) = (\bar{x}, \bar{y}).$$

Let $\{d^t\}$ be a sequence of directions such that

$$\nabla_x g(x^t, y^t)^\top d^t < 0$$

and assume that $\|d^t\| \leq M$ for some $M > 0$ and for all $t \in T$. If α^t is a sequence of step sizes satisfying Armijo condition and

$$\lim_{\substack{t\to\infty \\ t\in T}} g(x^t, y^t) - g(x^t + \alpha_t d^t, y^t) = 0,$$

then we have

$$\lim_{\substack{t\to\infty \\ t\in T}} \nabla_x g(x^t, y^t)^\top d^t = 0.$$

We can now prove the main convergence result.

Proposition 1 *Let (x^k, y^k) be the sequence generated by Algorithm 1. Then, (x^k, y^k) admits cluster points and each cluster point $(\bar{x}, \bar{y})$ satisfies conditions (2).*

Proof From the instructions of the algorithm, we have

$$f(x^{k+1}, y^k) < f(x^k, y^k)$$

and

$$f(x^{k+1}, y^{k+1}) + \eta\|y^{k+1}\|_0 \leq f(x^{k+1}, y^k) + \eta\|y^k\|_0.$$

Thus

$$f(x^{k+1}, y^{k+1}) + \eta\|y^{k+1}\|_0 < f(x^k, y^k) + \eta\|y^k\|_0.$$

The sequence $\{f(x^k, y^k) + \eta\|y^k\|_0\}$ is therefore monotone decreasing; thus it admits a limit $F^\star$ which is finite since $f(x, y) + \|y\|_0$ is obviously bounded below.

The level set $\{x, y \mid f(x, y) \leq f(x^0, y^0)\}$ is compact, being $f(x, y)$ coercive (Assumption 1); similarly, the set $\mathcal{L}_0 = \{x, y \mid f(x, y) \leq f(x^0, y^0) + \eta m\}$ is also compact. In addition, the term $\eta\|y\|_0$ is bounded above by ηm. Hence, for all k, we have

$$\begin{aligned} f(x^k, y^k) &\leq f(x^k, y^k) + \eta\|y^k\|_0 \\ &\leq f(x^0, y^0) + \eta\|y^0\|_0 \\ &\leq f(x^0, y^0) + \eta m, \end{aligned}$$

i.e., $\{x^k, y^k\}$ belongs to the compact set $\mathcal{L}_0$ and has thus cluster points.

Now, let $K \subseteq \{0, 1, \ldots\}$ be an infinite subsequence such that

$$\lim_{\substack{k\to\infty \\ k\in K}} (x^k, y^k) = (\bar{x}, \bar{y}).$$

Let $d^k = -\nabla_x f(x^k, y^k)$. Recalling the continuity of ∇f, we have $d^k \to -\nabla_x f(\bar{x}, \bar{y})$ as $k \to \infty, k \in K$, and hence there exists $M > 0$ such that $\|d^k\| \leq M$ for all $k \in K$ sufficiently large. Moreover, we know that $\nabla_x f(x^k, y^k)^\top d^k = -\|\nabla_x f(x^k, y^k)\|^2 < 0$.

We can also remark that

$$\begin{aligned} f(x^{k+1}, y^{k+1}) + \eta\|y^{k+1}\|_0 &\leq f(x^{k+1}, y^k) + \eta\|y^k\|_0 \\ &= f(x^k + \alpha_k d^k, y^k) + \eta\|y^k\|_0 \\ &< f(x^k, y^k) + \eta\|y^k\|_0. \end{aligned}$$

Since the rightmost and the leftmost terms in the above chain of inequalities both converge to $F^\star$ as $k \to \infty$, we have that

$$\begin{aligned} 0 &= \lim_{\substack{k\to\infty \\ k\in K}} f(x^k, y^k) + \eta\|y^k\|_0 - f(x^k + \alpha_k d^k, y^k) - \eta\|y^k\|_0 \\ &= \lim_{\substack{k\to\infty \\ k\in K}} f(x^k, y^k) - f(x^k + \alpha_k d^k, y^k). \end{aligned}$$

We thus have that all the conditions of Lemma 2 hold and thus

$$\lim_{\substack{k\to\infty \\ k\in K}} -\|\nabla_x f(x^k, y^k)\|^2 = -\|\nabla_x f(\bar{x}, \bar{y})\|^2 = 0,$$

i.e., the first condition in (2) is satisfied at $(\bar{x}, \bar{y})$.

Now, we know that, for all k, y^k is a global optimizer of $f(x^k, y) + \eta\|y\|_0$, hence it satisfies Lu-Zhang conditions [6], i.e., $\nabla_{y_i} f(x^k, y^k) = 0$ for all i such that $y_i^k \neq 0$.

Let $S_k = \{i \mid y_i^k \neq 0\}$ and $\bar{S} = \{i \mid \bar{y}_i \neq 0\}$. Since $y^k \to \bar{y}$ as $k \to \infty, k \in K$, by continuity, if $\bar{y}_i \neq 0$ then $y_i^k \neq 0$ for k sufficiently large; hence $S_k \supseteq \bar{S}$ for $k \in K$ sufficiently large. We can conclude that $\nabla_{y_{\bar{S}}} f(x^k, y^k) = 0$ for all $k \in K$ sufficiently large and therefore, by the continuity of $\nabla f(x, y)$, $\nabla_{y_{\bar{S}}} f(\bar{x}, \bar{y}) = 0$. The second condition in (2) is thus satisfied at $(\bar{x}, \bar{y})$. □

4 Applications

The convergence results from Sect. 3 can be exploited in application settings where the y-update step can effectively be carried out by a global optimization step; hereafter we discuss in detail some of the most relevant ones.

4.1 Inexact Penalty Decomposition

The Penalty Decomposition (PD) approach was proposed in [11] to tackle generic optimization problems with ℓ_0 penalty terms, i.e., problems of the form

$$\min_{x \in C} \; f(x) + \eta \|x\|_0,$$

where $C \subseteq \mathbb{R}^n$ is a regular feasible set.

In essence, the PD method sequentially considers penalty functions of the form

$$q_{\tau_\kappa}(x, y) = f(x) + \eta \|y\|_0 + \tau_\kappa (\text{dist}^2(x, C) + \|x - y\|^2), \tag{3}$$

and seeks at each iteration points $(x^{\kappa+1}, y^{\kappa+1})$ such that

$$\|\nabla_x q_{\tau_\kappa}(x^{\kappa+1}, y^{\kappa+1})\| \leq \epsilon_\kappa \qquad y^{\kappa+1} \in \arg\min_y \; \|x^{\kappa+1} - y\|^2 + \eta \|y\|_0. \tag{4}$$

To produce such points, an (exact) alternate minimization scheme is employed as subprocedure.

We can note that the unconstrained minimization of function (3) perfectly fits the framework (1) considered in this work: function $f(x) + \tau_\kappa(\text{dist}^2(x, C) + \|x - y\|^2)$ is coercive if f is coercive [15, 16] and y-update subproblem can be solved in closed form up to global optimality [11].

Therefore, the result of Proposition 1 implies that the inexact version (Algorithm 1) of the AM procedure can be used within the PD approach to generate in finite time, for each κ, a point $(x^{\kappa+1}, y^{\kappa+1})$ satisfying (4). Indeed, by Proposition 1, Algorithm 1 is able to produce a (sub)sequence (x^t, y^t) such that $(x^t, y^t) \to (\bar{x}, \bar{y})$ with $\|\nabla_x q_{\tau_\kappa}(\bar{x}, \bar{y})\| = 0$; by the continuity of $\nabla_x q_{\tau_\kappa}$, we are guaranteed that the first property in (4) is satisfied by (x^t, y^t) for t sufficiently large; as for the second property in (4), it directly follows for any t from the instructions of Algorithm 1.

In conclusion, similarly as what was shown for the case of cardinality constrained problems [15, 16], we have basically proved here that the PD scheme can be employed, with no loss of convergence guarantees, solving the subproblems by inexact AM; this result also justifies the use of the PD method in the very common settings where the function f is not convex.

4.2 Sparse RBF Regression Problems

Let $\mathcal{D} = \{(u_i, v_i) \mid u_i \in \mathbb{R}^p, v_i \in \mathbb{R}, i = 1, \ldots, N\}$ be a dataset of observations. The problem of *radial basis functions (RBF) regression* [17–20] involves the solution of optimization problems of the form

$$\min_{\substack{\lambda \in \mathbb{R}^K \\ c \in \mathbb{R}^{K \times p}}} \; E(c, \lambda) = \|\Phi(c)\lambda - v\|^2 + \mu \left(\|\lambda\|^2 + \|c\|^2 \right), \tag{5}$$

to obtain the approximating model $f(u) = \sum_{i=1}^{K} \lambda_i \phi(\|u_i - c_i\|)$. Here, $K \leq N$ is the number of RBF centers, $\mu \geq 0$ is a regularization parameter and the matrix $\Phi(c)$ is defined as a function of centers c:

$$\Phi(c) = \begin{pmatrix} \phi(\|u_1 - c_1\|) & \phi(\|u_1 - c_2\|) & \cdots & \phi(\|u_1 - c_K\|) \\ \phi(\|u_2 - c_1\|) & \phi(\|u_2 - c_2\|) & \cdots & \phi(\|u_2 - c_K\|) \\ \vdots & \vdots & \ddots & \vdots \\ \phi(\|u_N - c_1\|) & \phi(\|u_N - c_2\|) & \cdots & \phi(\|u_N - c_K\|) \end{pmatrix} \in \mathbb{R}^{N \times K}.$$

Arguably, the most popular radial basis function is the Gaussian one, i.e., $\phi(r) = \exp(-\gamma r^2)$, but many other valid options exist. Anyhow, the resulting radial regressor is expressed as a linear combination of RBFs, each term of which is radially symmetric around its center.

Problem (5) is particularly well understood in the field of RBF Neural Networks training, where decomposition schemes were studied for its solution [21].

Since sparsity is well known to induce robustness in regression models [4], we might be tempted to consider a sparse version of the problem, i.e.,

$$\min_{\substack{\lambda \in \mathbb{R}^K \\ c \in \mathbb{R}^{K \times p}}} E(c, \lambda) = \|\Phi(c)\lambda - v\|^2 + \mu \left(\|\lambda\|^2 + \|c\|^2 \right) + \eta \|\lambda\|_0. \tag{6}$$

Basically, imposing sparsity on the vector of coefficients λ has the effect of constructing a regression model based on few radial bases in cases when the number of RBF centers to be used is unknown.

We can notice that the function $E(c, \lambda)$ is coercive, as it is the sum of a bounded below function, $\|\Phi(c)\lambda - v\|^2$, and a coercive term, $\mu(\|\lambda\|^2 + \|c\|^2)$; moreover, for fixed c, the problem can be equivalently reformulated as a mixed-integer quadratic programming problem:

$$\begin{aligned} \arg\min_{\lambda \in \mathbb{R}^N,\ z \in \{0,1\}^N} \quad & \|\Phi(\bar{c})\lambda - v\|^2 + \mu \left(\|\lambda\|^2 + \|\bar{c}\|^2 \right) + \eta(\mathbb{1}^\top z) \\ \text{s.t.} \quad & -Mz_i \leq \lambda_i \leq Mz_i, \qquad i = 1, \ldots, N, \end{aligned} \tag{7}$$

where $M > 0$ is a sufficiently large constant. This problem is solvable up to global optimality with standard software solvers for mixed-integer optimization.

Therefore, we are again in the setting covered by the framework (1) considered in this paper. Problem (6) can thus be solved by Algorithm 1 with the guarantee of obtaining a point satisfying first-order optimality conditions. Interestingly, Algorithm 1 can be seen as the extension to the ℓ_0-penalized problem (6) of the alternate minimization scheme proposed and analyzed in [21] for problem (5).

Finally, we shall also observe that the convergence result still holds if centers associated to zero coefficients are progressively removed from the optimization problem. Indeed, we are recursively applying the same algorithm to a smaller problem. In fact, the gradients w.r.t. a removed center remain zero no matter the value of the other variables, while the corresponding weight is set to zero and thus does not affect the necessary optimality condition.

5 Numerical Validation

In this section, we report the results of numerical experiments aimed at corroborating computationally the theoretical properties described and proved in this paper. Specifically, we carried out tests on instances of the problems discussed in Sect. 4.

In particular, we considered 48 instances of penalty problems with objective functions of the form (3), where $C = \mathbb{R}^n$, $n = 1000$. We considered all possible combinations of values of $\tau_k \in \{1, 10, 100\}$ and $\eta \in \{0.5, 2\}$, whereas for f we took the following functions from the CUTEst test suite [22]: `morebv`, `arwhead`, `noncvxun`, `noncvxu2`, `powellsg`, `power`, `tointgss`, `engval1`. For all problems, the starting point $x^0 = y^0$ is the one defined in CUTEst. Algorithm 1 has been implemented in Python v3.9, using `numpy` library; the inexact x-update step is carried out performing 20 iterations of the conjugate gradient method, for which we exploit the `scipy` library implementation. In each run, the algorithm stops as soon as condition (4) is satisfied for $\epsilon_\kappa = 0.001$.

The results are shown, in the form of cumulative distributions, in Fig. 1. In particular, in Fig. 1a we can observe the cumulative distribution, computed over the 48 problem instances, of the number of AM iterations required to activate the stopping condition. We observe that the condition is actually attained in a finite number of iterations in all cases and that often this number rather low. Figure 1b concerns the fraction of nonzero variables in the y part of the final solution, i.e., vector $\|y^*\|_0/n$. Of course this value is very strongly tied to the choice of η. The important thing to note here is that the benchmark is not entirely made up of possibly trivial cases where the final solution has either n or 0 active variables. Finally, in Fig. 1c we can see the distribution of the relative reduction attained of the objective function, i.e., $(q_{\tau_\kappa}(x^0, y^0) - q_{\tau_\kappa}(x^*, y^*))/q_{\tau_\kappa}(x^*, y^*)$. This result indicates that the algorithm has an actual, substantial impact at cutting down the objective function, possibly of many orders of magnitude.

Similar experiments have been carried out on instances of problem (6), with Gaussian kernel ($\gamma = 1$). Again, the code was implemented in Python 3.9, this time doing 20 iterations of the L-BFGS method [23] for the inexact update of variables c, whereas variables λ are updated solving problem (7) with `Gurobi` solver.

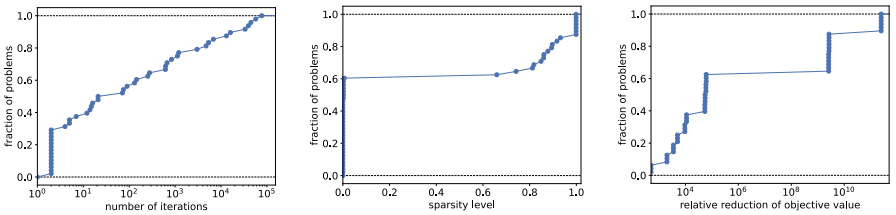

Fig. 1 Cumulative distributions for results obtained by the inexact AM algorithm on 48 penalty problems with objective of the form (3). Metrics considered are: number of iterations to satisfy stopping condition (**a**); sparsity level, i.e., fraction of nonzero variables of y block of variables at final solution (**b**); relative decrease of objective function (**c**)

Table 1 Results obtained on sparse RBF regression problems

Problem	p	$E(c^0, \lambda^0)$	$E(c^*, \lambda^*)$	$\ell_0(\lambda^*)$	# AM iterations
Adjiman	2	25.800	0.937	7	5
Ackley	2	24.339	3.205	2	6
Branin	2	23.516	2.766	3	7
CamelSixHumps	2	23.262	3.875	24	6
Rosenbrock4	4	17.135	0.182	53	6
Hartman6	6	52.918	0.032	59	26

Here, we considered the following 6 global optimization test functions: `Brainin`, `CamelSixHumps`, `Hartman6`, `Adjiman`, `Ackley`, `Rosenbrock4`. For each one, we randomly sampled $N = 80$ observations to build the sparse RBF approximation according to (6). For each problem we set $\mu = 10^{-6}$ and $\eta = 0.0005$. Algorithm 1 stops as soon as $\|\nabla_c E(c^{k+1}, \lambda^{k+1})\| \leq 0.01$. The results, reported in Table 1, highlight again that we obtained pretty fast convergence (towards nontrivial solutions) of the algorithm.

6 Conclusions

In this paper, we studied the convergence properties of the inexact Alternate Minimization method in problems with ℓ_0-penalty terms in the objective function. We showed that convergence can be attained without recurring to proximal terms even if the update of one of the two blocks of variables is carried out inexactly. This result has relevant implications in application settings. In particular, the considered algorithm can be used, with no loss of convergence guarantees for the overall method, within the popular Penalty Decomposition framework; the latter can thus be effectively used even without convexity assumptions on the objective function. Also, the proposed approach is suited to solve sparse RBF regression problems. Numerical experiments on the aforementioned classes of problems corroborated the theoretical finding discussed in this manuscript.

Author Contribution M.L. devised the paper concept and carried out the theoretical analysis; both authors carried out the literature review, designed the algorithmic scheme, identified applications, carried out numerical experiments, and wrote the manuscript.

Funding No funding was received for conducting this study.

Data Availability No datasets were generated or analyzed during the current study.

Declarations

Competing Interest The authors declare no competing interests.

References

1. Blumensath T, Davies ME (2009) Iterative hard thresholding for compressed sensing. Appl Comput Harmon Anal 27(3):265–274
2. Foucart S, Rauhut H (2013) An invitation to compressive sensing. In: A Mathematical Introduction to Compressive Sensing. Springer, pp 1–39
3. Bach F, Jenatton R, Mairal J, Obozinski G (2011) Optimization with sparsity-inducing penalties. Preprint at http://arxiv.org/abs/1108.0775
4. Weston J, Elisseeff A, Schölkopf B, Tipping M (2003) Use of the zero norm with linear models and kernel methods. J Mach Learn Res 3:1439–1461
5. Bertsimas D, King A, Mazumder R et al (2016) Best subset selection via a modern optimization lens. Ann Stat 44(2):813–852
6. Civitelli E, Lapucci M, Schoen F, Sortino A (2021) An effective procedure for feature subset selection in logistic regression based on information criteria. Comput Optim Appl 1–32
7. Di Gangi L, Lapucci M, Schoen F, Sortino A (2019) An efficient optimization approach for best subset selection in linear regression, with application to model selection and fitting in autoregressive time-series. Comput Optim Appl 74(3):919–948
8. Tillmann AM, Bienstock D, Lodi A, Schwartz A (2021) Cardinality minimization, constraints, and regularization: a survey. Preprint at http://arxiv.org/abs/2106.09606
9. Bertsekas D, Tsitsiklis J (2015) Parallel and distributed computation: Numerical methods. Athena Scientific
10. Grippo L, Sciandrone M (1999) Globally convergent block-coordinate techniques for unconstrained optimization. Optimization methods and software 10(4):587–637
11. Lu Z, Zhang Y (2013) Sparse approximation via penalty decomposition methods. SIAM J Optim 23(4):2448–2478
12. Bolte J, Sabach S, Teboulle M (2014) Proximal alternating linearized minimization for nonconvex and nonsmooth problems. Math Program 146(1):459–494
13. Bertsekas DP, Hager W, Mangasarian O (1999) Nonlinear programming. Athena Scientific Belmont: Massachusets, USA
14. Bertsekas DP (1997) Nonlinear programming. J Oper Res Soc 48(3):334–334
15. Kanzow C, Lapucci M (2023) Inexact penalty decomposition methods for optimization problems with geometric constraints. Comput Optim Appl 1–35
16. Lapucci M, Levato T, Sciandrone M (2021) Convergent inexact penalty decomposition methods for cardinality-constrained problems. J Optim Theory Appl 188(2):473–496
17. Bishop CM et al (1995) Neural networks for pattern recognition. Oxford University Press
18. Fasshauer GE (2007) Meshfree approximation methods with MATLAB, vol. 6. World Scientific
19. Poggio T, Girosi F (1990) Networks for approximation and learning. Proc IEEE 78(9):1481–1497
20. Wendland H (2004) Scattered data approximation, vol. 17. Cambridge University Press
21. Buzzi C, Grippo L, Sciandrone M (2001) Convergent decomposition techniques for training RBF neural networks. Neural Comput 13(8):1891–1920
22. Gould NI, Orban D, Toint PL (2015) Cutest: a constrained and unconstrained testing environment with safe threads for mathematical optimization. Comput Optim Appl 60:545–557
23. Liu DC, Nocedal J (1989) On the limited memory bfgs method for large scale optimization. Math Program 45(1):503–528