Active-Set Identification with Complexity Guarantees of an Almost Cyclic 2-Coordinate Descent Method with Armijo Line Search

Published version:
DOI: 10.1137/20M1328014

(Article begins on next page)

ACTIVE-SET IDENTIFICATION WITH COMPLEXITY GUARANTEES OF AN ALMOST CYCLIC 2-COORDINATE DESCENT METHOD WITH ARMIJO LINE SEARCH*

ANDREA CRISTOFARI†

Abstract. In this paper, it is established finite active-set identification of an almost cyclic 2-coordinate descent method for problems with one linear coupling constraint and simple bounds. First, general active-set identification results are stated for non-convex objective functions. Then, under convexity and a quadratic growth condition (satisfied by any strongly convex function), complexity results on the number of iterations required to identify the active set are given. In our analysis, a simple Armijo line search is used to compute the stepsize, thus not requiring exact minimizations or additional information.

Key words. active-set identification, surface identification, manifold identification, active-set complexity, block coordinate descent methods

AMS subject classifications. 90C06, 90C30, 65K05

1. Introduction. In many different contexts, a desirable property of an optimization algorithm is the ability to identify, in a finite number of iterations, a surface containing an optimal solution, in the sense that the points generated by the algorithm eventually remain on that surface. After such an identification, convergence can indeed be faster since the algorithm can work in a lower dimensional space and, under proper assumptions, it may also be possible to switch to methods with higher convergence rate. Furthermore, in certain problems one may only be interested in knowing the structure of an optimal solution, which can be revealed by identifying a surface where it lies, without the need of running the algorithm to convergence (for example, in lasso problems sparse solutions are promoted by the ℓ_1 norm and one may only be interested in knowing the support of an optimal solution).

In the literature, much effort has been devoted to proving identification properties of some algorithms for smooth optimization [3, 5, 6, 7, 8, 9, 10, 11, 19, 22, 25, 48, 50], non-smooth optimization [16, 24, 26, 30, 32, 36, 42, 43, 49, 51], stochastic optimization [18, 29, 47] and derivative-free optimization [31]. Moreover, a wide class of methods, known as *active-set methods*, has been object of extensive study from decades (see, e.g., [4, 13, 14, 17, 20, 23] and the references therein), making use of specific techniques to identify the so called *active set*, which is the set of constraints or variables that parametrizes a surface containing a solution.

The scope of the present paper is establishing finite active-set identification of a 2-coordinate descent method, proposed by the author in [12], for smooth minimization problems with one linear equality constraint and simple bounds on the variables. The main contributions of this paper can be summarized in the following points:

(i) The problem we consider here is not separable, due to a coupling constraint, and the method under analysis does not require first-order information to choose the working set, while guaranteeing deterministic convergence properties.
These features represent major differences with the analysis of other block coordinate descent methods for which active-set identification results have been proved [15, 17, 27, 34, 35, 42, 48, 51], since these methods either solve uncon-

*Submitted to the editors November 18, 2021.

†Department of Mathematics "Tullio Levi-Civita", University of Padua, Via Trieste, 63, 35121 Padua, Italy (andrea.cristofari@unipd.it).

strained problems where the objective function is the sum of a smooth term and a convex separable term (the latter might be an indicator function that enforces bound constraints), or allow for a non-separable structure but require full gradient evaluations to choose the working set, or have convergence results in expectation. In particular, active-set identification results are given in [48] for variants of the sequential minimal optimization algorithm applied to the Support Vector Machine problem, where the authors consider a random selection of the working-set, which therefore does not require first-order information, but leads to convergence results in expectation.

(ii) Besides stating finite active-set identification results in a general non-convex setting, complexity results are also given under convexity of the objective function and a quadratic growth condition (satisfied by any strongly convex function), allowing us to bound the maximum number of iterations needed to identify the active set.

Let us also remark that here we consider a simple Armijo line search for computing the stepsize along any search direction, thus not requiring exact minimizations, or the knowledge of the Lipschitz constant of the gradient, or other additional information. This makes our analysis of particular interest for realistic application to large-scale optimization problems.

2. Preliminaries and Notation. Let us first introduce part of the notation used in the paper. Given a function $f \colon \mathbb{R}^n \to \mathbb{R}$, we indicate the gradient of f by ∇f and we denote by $\nabla_i f$ its ith component (i.e., the ith partial derivative of f). For a vector $x \in \mathbb{R}^n$, we denote by x_i the ith component of x, we indicate by $\|x\|$ the Euclidean norm of x and we indicate by $\|x\|_\infty$ the sup-norm of x. We also denote by $e \in \mathbb{R}^n$ the vector made of all ones, and by $e_i \in \mathbb{R}^n$ the vector that has the ith component equal to 1 and all other components equal to 0. Given a scalar a, we indicate with $\lfloor a \rfloor$ the largest integer less than or equal to a.

Our analysis is concerned with the following problem:

$$\begin{aligned} &\min\ f(x) \\ &e^T x = b \\ &l_i \le x_i \le u_i, \quad i = 1,\dots,n, \end{aligned} \tag{2.1}$$

where $f \colon \mathbb{R}^n \to \mathbb{R}$ is a function with Lipschitz continuous gradient, $n \ge 2$, $b \in \mathbb{R}$ and, for all $i = 1,\dots,n$, we have $l_i < u_i$, $l_i \in \mathbb{R} \cup \{-\infty\}$, $u_i \in \mathbb{R} \cup \{+\infty\}$. The feasible set of problem (2.1) is denoted by $\mathcal{F}$.

Note that we may consider, instead of $e^T x = b$, any constraint of the form $a^T x = b$, with $a_i \neq 0$, $i = 1,\dots,n$. In such a case, problem (2.1) can be obtained by applying the variable transformation $x_i \leftarrow a_i x_i$ and setting the lower and the upper bound accordingly. (Examples of relevant applications where problem (2.1) arises can be found, e.g., in [12] and the references therein.)

The Lipschitz constant of ∇f over $\mathbb{R}^n$ is denoted by L, that is,

$$\|\nabla f(x) - \nabla f(y)\| \le L \|x - y\|, \quad \forall\, x, y \in \mathbb{R}^n.$$

It is possible to show [1] that there exist local Lipschitz constants

$$L_{i,j} \le 2L, \quad i,j = 1,\dots,n, \tag{2.2}$$

such that, for any $x \in \mathbb{R}^n$,

$$|\nabla f(x + s(e_i - e_j))^T (e_i - e_j) - \nabla f(x + t(e_i - e_j))^T (e_i - e_j)| \le L_{i,j} |s - t|, \quad \forall\, s, t \in \mathbb{R}.$$

Equivalently, defining $\phi_{i,j,x}(\alpha) = f(x + \alpha(e_i - e_j))$ and denoting its derivative by $\dot{\phi}_{i,j,x}$, we have that

$$|\dot{\phi}_{i,j,x}(s) - \dot{\phi}_{i,j,x}(t)| \le L_{i,j}|s-t|, \quad \forall\, s,t \in \mathbb{R}, \tag{2.3}$$

that is, each derivative $\dot{\phi}_{i,j,x}$ is Lipschitz continuous over $\mathbb{R}$ with constant $L_{i,j}$.

Without loss of generality, we assume that all $L_{i,j} > 0$, $i \ne j$ (if some of them are equal to zero, they can be replaced by positive overestimates) and that $L_{i,i} = 0$, $i = 1, \dots, n$. We also define the following constants:

$$L^{\max} = \max_{i,j=1,\dots,n} L_{i,j}, \tag{2.4}$$

$$L_j = \sum_{i=1}^{n} L_{i,j}, \quad j = 1, \dots, n, \tag{2.5}$$

$$\hat{L}^{\max} = \max_{j=1,\dots,n} L_j. \tag{2.6}$$

A characterization of stationary points for problem (2.1) follows from KKT conditions. In particular, a point $x^* \in \mathcal{F}$ is stationary for problem (2.1) if and only if there exists $\lambda^* \in \mathbb{R}$ such that, for all $i = 1, \dots, n$,

$$\nabla_i f(x^*) \begin{cases} \ge \lambda^*, & \text{if } x_i^* = l_i, \\ = \lambda^*, & \text{if } x_i^* \in (l_i, u_i), \\ \le \lambda^*, & \text{if } x_i^* = u_i. \end{cases} \tag{2.7}$$

Moreover, a variable $x_i^* \in \{l_i, u_i\}$ is said to satisfy the strict complementarity if $\nabla_i f(x^*) \ne \lambda^*$. We also say that x^* is non-degenerate if all variables x_i^* such that $x_i^* \in \{l_i, u_i\}$ satisfy the strict complementarity.

In the following, we will make use of a simple operator between vectors in $\mathbb{R}^n$, obtained from the usual dot product by discarding a certain component. More precisely, for any $j \in \{1, \dots, n\}$ we define the following positive semidefinite inner product:

$$\langle x, y \rangle_j = \sum_{i \ne j} x_i y_i, \quad \forall\, x, y \in \mathbb{R}^n.$$

We also define the following seminorm, induced by the above inner product:

$$\|x\|_{\langle j \rangle} = \sqrt{\langle x, x \rangle_j}, \quad \forall\, x \in \mathbb{R}^n.$$

Note that, by Cauchy-Bunyakovsky-Schwarz inequality, we have

$$\langle x, y \rangle_j \le \|x\|_{\langle j \rangle} \, \|y\|_{\langle j \rangle}, \quad \forall\, x, y \in \mathbb{R}^n. \tag{2.8}$$

In particular, (2.8) implies that

$$|x_i| \le \|x\|_{\langle j \rangle}, \quad i \ne j, \quad \forall\, x \in \mathbb{R}^n, \tag{2.9}$$

$$\sum_{i \ne j} |x_i| \le \sqrt{n-1}\|x\|_{\langle j \rangle}, \quad \forall\, x \in \mathbb{R}^n. \tag{2.10}$$

Moreover, it is straightforward to verify that

$$\|x\|_{\langle j \rangle} \le \|x\|, \quad \forall\, x \in \mathbb{R}^n. \tag{2.11}$$

3. Review of the algorithm. Let us briefly review the algorithm proposed in [12], named Almost Cyclic 2-Coordinate Descent (AC2CD) method, to solve problem (2.1). The main feature of AC2CD is an *almost cyclic* rule to choose the working set. This rule iteratively selects two variables: one is picked in a cyclic fashion, while the other one is chosen by considering the distance from the bounds in some points produced by the algorithm and remains in the working set until all the other variables have been picked. Note the difference from the so-called *essentially cyclic* rule, where all blocks of variables must be selected at least once within a certain number of steps.

More precisely, at the beginning of each outer iteration k of AC2CD we have a feasible point x^k and we select a variable index $j(k)$ such that $x^k_{j(k)}$ is "sufficiently far" from its nearest bound. Then, we set the point $z^{k,1} = x^k$ and start a cycle of inner iterations, which are denoted by $(k,1),\dots,(k,n)$. In each inner iteration (k,i), we choose a working set of two variables: one of them is selected in a cyclic fashion, while the other one remains the $j(k)$th variable. So, we produce a feasible point $z^{k,i+1}$ from $z^{k,i}$ by moving only the two variables in the working set. At the end of the last inner iteration we finally set $x^{k+1} = z^{k,n+1}$ and start a new outer iteration $k+1$.

Let us remark that our algorithm does not use first-order information to choose the working set. Moreover, as to be described later, only two partial derivatives are required to move each pair of variables. We can hence achieve high computational efficiency if partial derivative evaluation for the objective function is much cheaper than full gradient evaluation. For instance, this is the case when f is the sum of univariate functions (such as in the problems considered in [38] for large-scale network optimization). Other interesting examples, including the Support Vector Machine problem and the Chebyshev center problems, are those where the objective function is quadratic of the form $f(x) = x^T Q^T Q x - q^T x$, with Q being a given $m \times n$ matrix and q being a given vector. In this case, a partial derivative of $f(x)$ can be computed with a cost $\mathcal{O}(m)$, while computing the whole gradient has a cost $\mathcal{O}(mn)$ (see [12] for details).

Now, let us explain in more detail how the index $j(k)$ is chosen at the beginning of an outer iteration k and how the two variables in the working set are moved in the inner iterations $(k,1),\dots,(k,n)$.

For what concerns the choice of $j(k)$, for any $x \in \mathcal{F}$ let us first define

$$D_h(x) = \min\{x_h - l_h, u_h - x_h\}, \quad h = 1,\dots,n. \tag{3.1}$$

Namely, $D_h(x)$ returns the distance of x_h from its nearest bound. Moreover, for any point x^k produced by the algorithm, we define D^k as the maximum distance between each component of x^k and its nearest bound, that is,

$$D^k = \max_{h=1,\dots,n} D_h(x^k). \tag{3.2}$$

Then, $j(k)$ can be chosen as any index satisfying

$$D_{j(k)}(x^k) \ge \tau D^k, \tag{3.3}$$

where $\tau \in (0,1]$ is a fixed parameter. In other words, the distance between $x^k_{j(k)}$ and its nearest bound must be sufficiently large compared to D^k.

For what concerns the variable update, let us denote by p^k_i the variable index that is selected in a cyclic manner at an inner iteration (k,i) (note that the variables can be taken in any order). So, $z^{k,i}_{p^k_i}$ and $z^{k,i}_{j(k)}$ are the two variables that can be moved from $z^{k,i}$. To do this, we use the following search direction (which has at most two non-zero components and maintains feasibility for the equality constraint):

$$d^{k,i} = g^{k,i}(e_{p_i^k} - e_{j(k)}), \quad \text{where} \quad g^{k,i} = \nabla_{j(k)} f(z^{k,i}) - \nabla_{p_i^k} f(z^{k,i}), \tag{3.4}$$

and we set

$$z^{k,i+1} = z^{k,i} + \alpha^{k,i} d^{k,i},$$

where $\alpha^{k,i}$ is a suitably computed feasible stepsize. Note that

$$\nabla f(z^{k,i})^T d^{k,i} = -(g^{k,i})^2, \tag{3.5}$$

and then, every non-zero $d^{k,i}$ is a descent direction. The scheme of AC2CD is reported in Algorithm 3.1.

Algorithm 3.1 `Almost Cyclic 2-Coordinate Descent (AC2CD) method`

0 **Given** $x^0 \in \mathcal{F}$ and $\tau \in (0, 1]$
1 **For** $k = 0, 1, \dots$
2 Choose a variable index $j(k) \in \{1, \dots, n\}$ that satisfies (3.3)
3 Choose a permutation $\{p_1^k, \dots, p_n^k\}$ of $\{1, \dots, n\}$
4 Set $z^{k,1} = x^k$
5 **For** $i = 1, \dots, n$
6 Let $g^{k,i} = \nabla_{j(k)} f(z^{k,i}) - \nabla_{p_i^k} f(z^{k,i})$
7 Compute the search direction $d^{k,i} = g^{k,i}(e_{p_i^k} - e_{j(k)})$
8 Compute a feasible stepsize $\alpha^{k,i}$ and set $z^{k,i+1} = z^{k,i} + \alpha^{k,i} d^{k,i}$
9 **End for**
10 Set $x^{k+1} = z^{k,n+1}$
11 **End for**

3.1. Computation of the stepsize. Under a technical assumption (see Assumption 1 in the next section), global convergence of AC2CD to stationary points was established in [12] for different choices of the stepsize $\alpha^{k,i}$ (to be used at line 8 of Algorithm 3.1), including the Armijo stepsize, overestimates of the local Lipschitz constants of ∇f and the exact stepsize for strictly convex objective functions[1].

Here we focus on the case where, at every inner iteration (k, i), the stepsize $\alpha^{k,i}$ is computed by the Armijo line search, which is a backtracking procedure that computes a stepsize in a finite number of iterations. The scheme of the Armijo line search used in AC2CD is reported in Algorithm 3.2.

Algorithm 3.2 `Armijo line search` (to compute $\alpha^{k,i}$ at step 8 of AC2CD)

0 **Given** the search direction $d^{k,i}$ and two parameters $\gamma \in (0, 1)$, $\delta \in (0, 1)$
1 Choose a feasible stepsize $\Delta^{k,i} \geq 0$ and set $\alpha = \Delta^{k,i}$
2 **While** $f(z^{k,i} + \alpha d^{k,i}) > f(z^{k,i}) + \gamma\alpha\nabla f(z^{k,i})^T d^{k,i}$
3 Set $\alpha = \delta\alpha$
4 **End while**
5 Return $\alpha^{k,i} = \alpha$

[1]For general conditions on the stepsize, see SC (Stepsize Condition) 1 in [12]. A typo is present in point (i) of SC 1 in [12]: $f(z^{k,i+i})$ should be replaced by $f(z^{k,i+1})$.

We see that the considered Armijo line search is very simple and does not require exact minimizations or additional information (such as the knowledge of the Lipschitz constant of ∇f). For this reason, it can be an effective choice for non-convex large-scale problems and when no closed form is known for the stepsize.

To obtain global convergence of AC2CD to stationary points, an appropriate choice of the initial stepsize $\Delta^{k,i}$ at line 1 of Algorithm 3.2 is needed. In [12] it was shown that, at every inner iteration (k,i), a possible choice is the following:

$$\Delta^{k,i} = \min\{\bar{\alpha}^{k,i}, A^{k,i}\}, \tag{3.6}$$

where

- $\bar{\alpha}^{k,i}$ is the largest feasible stepsize along the direction $d^{k,i}$, that is,

$$\bar{\alpha}^{k,i} = \begin{cases} \dfrac{1}{g^{k,i}} \min\{u_{p_i^k} - z^{k,i}_{p_i^k}, z^{k,i}_{j(k)} - l_{j(k)}\}, & \text{if } g^{k,i} > 0, \\ \dfrac{1}{|g^{k,i}|} \min\{z^{k,i}_{p_i^k} - l_{p_i^k}, u_{j(k)} - z^{k,i}_{j(k)}\}, & \text{if } g^{k,i} < 0, \\ 0, & \text{if } g^{k,i} = 0; \end{cases} \tag{3.7}$$

- $A^{k,i}$ must be chosen between two finite positive constants, that is,

$$0 < A_l \le A^{k,i} \le A_u < \infty, \tag{3.8}$$

  with A_l and A_u being two fixed parameters.

We observe that, in (3.7), we set $\bar{\alpha}^{k,i} = 0$ when $g^{k,i} = 0$, i.e., when $d^{k,i} = 0$ (see (3.4)). Therefore, $\bar{\alpha}^{k,i}$ is not actually the largest feasible stepsize along $d^{k,i}$ when $d^{k,i} = 0$. This choice in the definition of $\bar{\alpha}^{k,i}$ simplifies the analysis and entails no loss of generality, since it stills guarantees that $z^{k,i+1} = z^{k,i}$ when $d^{k,i} = 0$. In particular, note that

$$d^{k,i} = 0 \quad \overset{(3.4)}{\Leftrightarrow} \quad g^{k,i} = 0 \quad \overset{(3.7)}{\Rightarrow} \quad \bar{\alpha}^{k,i} = 0 \quad \Leftrightarrow \quad z^{k,i+1} = z^{k,i}. \tag{3.9}$$

To obtain the last relation in (3.9), we can use (3.5), (3.6) and (3.8), leading to

$$\bar{\alpha}^{k,i} > 0 \quad \Leftrightarrow \quad \Delta^{k,i} > 0 \wedge \nabla f(z^{k,i})^T d^{k,i} < 0.$$

So, if $\bar{\alpha}^{k,i} > 0$, the Armijo line search returns a stepsize $\alpha^{k,i} > 0$, implying that $z^{k,i+1} \neq z^{k,i}$. Vice versa, if $\bar{\alpha}^{k,i} = 0$, the Armijo line search returns $\alpha^{k,i} = 0$, implying that $z^{k,i+1} = z^{k,i}$. Namely, the last relation in (3.9) holds.

4. Basic assumptions.

Let X^* be the set of all stationary points for problem (2.1) and also define the level set

$$\mathcal{L}^0 = \{x \in \mathcal{F} \colon f(x) \le f(x^0)\},$$

where $\mathcal{F}$ is the feasible set of problem (2.1) and x^0 is the starting point used in AC2CD. We assume that $\mathcal{L}^0$ is non-empty and compact (implying that both the feasible set $\mathcal{F}$ and the set of stationary points X^* are non-empty as well).

According to the results stated in [12], we also need the following assumption on the level set $\mathcal{L}^0$ to ensure global convergence of AC2CD (in the sense that every limit point of the sequence $\{x^k\}$ produced by the algorithm is stationary):

Assumption 1. $\forall\, x \in \mathcal{L}^0$, $\exists\, i \in \{1,\dots,n\} \colon x_i \in (l_i, u_i)$.

Namely, we require that every point of $\mathcal{L}^0$ has at least one component strictly between the lower and the upper bound. Note that Assumption 1 is automatically satisfied when $\mathcal{F}$ is the unit simplex (i.e., when in problem (2.1) we have $b = 1$, $l_i = 0$, $u_i = +\infty$, $i = 1, \dots, n$). Moreover, in [33] it is shown that Assumption 1 is also satisfied for the Support Vector Machine training problem if $f(x^0) < 0$ and the smallest eigenvalue of the Hessian matrix of $f(x)$ is sufficiently large. (Assumption 1 is satisfied also when at least one variable has are no finite bounds, provided $\mathcal{F}$ is not a singleton.)

Essentially, Assumption 1 is needed to prevent AC2CD from converging to a point x^* with all components at the lower or the upper bound. To be more specific, the convergence analysis of AC2CD (see [12]) relies on the fact that eventually $l_{j(k)} < x^k_{j(k)} < u_{j(k)}$ and that $\nabla_{j(k)} f(x^k)$ converges (over suitable subsequences) to the KKT multiplier λ^* appearing in (2.7). Also the analysis of the active-set identification reported later uses the same properties (see Proposition 6.2 and the proof of Theorem 6.4 below). Without Assumption 1, all these results do not hold, since $\{x^k\}$ may have limit points with all components at the lower or the upper bound.

We also observe that, for every outer iteration $k \ge 0$, Assumption 1 ensures that $x^{k+1} \neq x^k$ if and only if x^k is non-stationary. To see this, under Assumption 1 observe that $l_{j(k)} < x^k_{j(k)} < u_{j(k)}$ for all $k \ge 0$ (since $j(k)$ must satisfy (3.3) with $D^k > 0$). Then, from the KKT conditions (2.7), there exists a feasible descent direction in the inner iterations $(k, 1), \dots, (k, n)$ if and only if x^k is non-stationary.

On the contrary, without Assumption 1, the algorithm may end up in a non-stationary point x^k with all components at the lower or the upper bound. In such a case, even if every choice of $j(k) = 1, \dots, n$ satisfies (3.3) (since $D^k = D_h(x^k) = 0$, $h = 1, \dots, n$), for certain choices of $j(k)$ there may not exist a feasible descent direction in any inner iteration $(k, 1), \dots, (k, n)$. Namely, AC2CD may get stuck in a non-stationary point x^k. This issue can be overcome by introducing an anticycling rule to select $j(k)$ when such a point x^k is produced. Doing so, we may relax Assumption 1 by requiring only the stationary points in $\mathcal{L}^0$ not to have all components at the lower or the upper bound, but in our analysis we use Assumption 1 for simplicity.

Overcoming the limitation deriving from Assumption 1 by properly modifying the algorithm might be a challenging subject for future research.

In the rest of the paper, we will consider all the above assumptions always satisfied, even if not explicitly invoked. Namely, we will consider $\mathcal{L}^0$ non-empty and compact and we will consider Assumption 1 satisfied.

5. Technical results.

In this section, we fix a few concepts and give some technical results. First note that, for every inner iteration (k, i) of AC2CD,

$$p^k_i \neq j(k) \quad \Rightarrow \quad z^{k,i}_{p^k_i} = x^k_{p^k_i} \text{ and } x^{k+1}_{p^k_i} = z^{k,i+1}_{p^k_i}, \tag{5.1}$$

since each coordinate, except the $j(k)$th one, is moved (at most) once in a cycle of inner iterations.

Furthermore, there is a relation between the Armijo stepsize and the local Lipschitz constants of ∇f: at any inner iteration (k, i), every stepsize $\alpha \le 2(1-\gamma)/L_{p^k_i, j(k)}$ satisfies the so called Armijo condition, which is the exit condition in the while loop of Algorithm 3.2. Namely, $f(z^{k,i} + \alpha d^{k,i}) \le f(z^{k,i}) + \gamma\alpha\nabla f(z^{k,i})^T d^{k,i}$ for all $\alpha \in [0, 2(1-\gamma)/L_{p^k_i, j(k)}]$ (see the proof of Proposition 3 in [12]). Since, in our line search, α is multiplied by $\delta \in (0, 1)$ until the Armijo condition is satisfied (see line 3 in Algorithm 3.2), we immediately have the following result.

LEMMA 5.1. *At every inner iteration (k,i), the initial stepsize $\Delta^{k,i}$ used in the Armijo line search is such that*

$$\Delta^{k,i} \le \frac{2(1-\gamma)}{L_{p_i^k,j(k)}} \;\Rightarrow\; \alpha^{k,i} = \Delta^{k,i},$$
$$\Delta^{k,i} > \frac{2(1-\gamma)}{L_{p_i^k,j(k)}} \;\Rightarrow\; \alpha^{k,i} \in \left(\frac{2\delta(1-\gamma)}{L_{p_i^k,j(k)}}, \Delta^{k,i}\right],$$

where, in the Armijo line search, $\gamma \in (0,1)$ is the parameter for sufficient decrease and $\delta \in (0,1)$ is the reduction parameter. Therefore, $\alpha^{k,i} \ge \min\left\{\Delta^{k,i}, \dfrac{2\delta(1-\gamma)}{L_{p_i^k,j(k)}}\right\}$.

As a consequence of Lemma 1 in [12], we also have the following relation between the limit of $\{x^k\}$ and the limit of the sequences $\{z^{k,i}\}$, $i = 1,\dots,n$:

$$\lim_{k\to\infty} x^k = x^* \;\Leftrightarrow\; \lim_{k\to\infty} z^{k,i} = x^*, \quad i = 1,\dots,n. \tag{5.2}$$

Now we state some useful properties derived from the semidefinite inner product and the seminorm defined at the end of Section 2. In the following results, we use L_j as defined in (2.5). The proofs are reported in Appendix A.

LEMMA 5.2. *For any $j \in \{1,\dots,n\}$ we have that*

$$v^T(x'-x'') = \langle v - v_j e, x'-x''\rangle_j, \quad \forall\, x',x'' \in \mathcal{F}, \quad \forall\, v \in \mathbb{R}^n.$$

LEMMA 5.3. *If f is convex over $\mathbb{R}^n$, for any $j \in \{1,\dots,n\}$ we have that*

$$\big\|[\nabla f(x') - \nabla_j f(x')e] - [\nabla f(x'') - \nabla_j f(x'')e]\big\|_{\langle j\rangle} \le L_j \|x'-x''\|_{\langle j\rangle}, \quad \forall\, x',x'' \in \mathcal{F}$$

COROLLARY 5.4. *If f is convex over $\mathbb{R}^n$, at every inner iteration (k,i) of AC2CD we have that*

$$|\nabla_{p_i^k} f(v) - \nabla_{j(k)} f(v) + g^{k,i}| \le L_{j(k)} \|v - z^{k,i}\|_{\langle j(k)\rangle}, \quad \forall\, v \in \mathbb{R}^n.$$

LEMMA 5.5. *If f is convex over $\mathbb{R}^n$, for any $j \in \{1,\dots,n\}$ we have that*

$$f(x'') \le f(x') + \nabla f(x')^T(x''-x') + \frac{L_j}{2}\|x'-x''\|^2_{\langle j\rangle}, \quad \forall\, x',x'' \in \mathcal{F}.$$

6. Active-set identification in the non-convex case. In this section, we show that AC2CD identifies the active set of problem (2.1) in a finite number of iterations, without any assumption on the convexity of f.

First of all, let us give the definition of active set for our problem.

DEFINITION 6.1. *Given a stationary point x^* of problem (2.1), we define the active set as*

$$\mathscr{Z}(x^*) = \{i\colon x_i^* = l_i\} \cup \{i\colon x_i^* = u_i\}.$$

We also define

$$\mathscr{Z}^+(x^*) = \mathscr{Z}(x^*) \cap \{i\colon \nabla_i f(x^*) \ne \lambda^*\},$$

where λ^* is the KKT multiplier associated with x^* appearing in (2.7).

We see that $\mathscr{Z}(x^*)$ is the set of indices of all the variables that are at the lower or the upper bound in a stationary point x^*, whereas $\mathscr{Z}^+(x^*)$ contains only the indices of the variables satisfying the strict complementarity. We notice that, from a geometric perspective, $\mathscr{Z}^+(x^*)$ defines the face of $\mathcal{F}$ exposed to $-\nabla f(x^*)$ [9].

The scope of this section is two-fold:

(i) Firstly, it will be shown that, given a sequence of points $\{x^k\} \to x^*$ produced by AC2CD, an iteration $\bar{k}$ exists such that, for all $k > \bar{k}$,

$$x^k_h = x^*_h, \quad \forall\, h \in \mathscr{Z}^+(x^*). \tag{6.1}$$

Namely, in a finite number of iterations AC2CD sets to the bounds all the variables that satisfy the strict complementarity at x^*.

(ii) Secondly, we will give a characterization of the neighborhood of x^* where (6.1) holds, which will be used in Section 7 to obtain an upper bound for $\bar{k}$ (under convexity of f and a quadratic growth condition).

Note that, as common when analyzing active-set identification properties of an optimization algorithm, here we require the whole sequence $\{x^k\}$ to converge. For AC2CD, in [12] it was shown that every limit point of $\{x^k\}$ is stationary and, if $\{f(x^k)\}$ converges, then $\lim_{k\to\infty}\|z^{k,i+1} - z^{k,i}\| = 0$, $i = 1,\dots,n$, implying that $\lim_{k\to\infty}\|x^{k+1} - x^k\| = 0$ if a limit point of $\{x^k\}$ exists. So, using the same arguments given in [45, Theorem 14.1.5], we get that the whole sequence $\{x^k\}$ converges if the number of stationary points in $\mathcal{L}^0$ is finite. By a more general result stated in [21, Proposition 8.3.10], we also have that the whole sequence $\{x^k\}$ converges if it has an isolated limit point. Other conditions can be obtained from [5, Theorem 4.3]: if f satisfies a suitable descent property along the search directions, then a strict local minimum with no other stationary points in its neighborhood attracts the whole sequence $\{x^k\}$.

Now, we start our analysis by giving an intermediate result stating that, in a neighborhood of x^*, the index $j(k)$ is such that $l_{j(k)} < x^*_{j(k)} < u_{j(k)}$.

PROPOSITION 6.2. *Let $\{x^k\}$ be a sequence of points produced by AC2CD and assume that $\lim_{k\to\infty} x^k = x^*$. Define the maximum distance from the bounds at x^* as*

$$D^{max}(x^*) = \max_{i=1,\dots,n} D_i(x^*),$$

which is positive by Assumption 1, and let k^j be the first outer iteration such that

$$\|x^k - x^*\|_\infty < \frac{\tau}{\tau+1} D^{max}(x^*), \quad \forall\, k \ge k^j,$$

where $\tau \in (0,1]$ is the parameter used to choose $j(k)$, satisfying (3.3). Then, for all $k \ge k^j$ we have that $j(k) \notin \mathscr{Z}(x^*)$.

Proof. Consider an outer iteration $k \ge k^j$ and let $\hat{\jmath}$ be an index such that $D_{\hat{\jmath}}(x^*) = D^{\max}(x^*)$. We have $|x^k_{\hat{\jmath}} - x^*_{\hat{\jmath}}| \le \|x^k - x^*\|_\infty < \dfrac{\tau}{\tau+1} D^{\max}(x^*)$, implying that

$$x^k_{\hat{\jmath}} - l_{\hat{\jmath}} > x^*_{\hat{\jmath}} - l_{\hat{\jmath}} - \frac{\tau}{\tau+1} D_{\hat{\jmath}}(x^*) \quad \text{and} \quad u_{\hat{\jmath}} - x^k_{\hat{\jmath}} > u_{\hat{\jmath}} - x^*_{\hat{\jmath}} - \frac{\tau}{\tau+1} D_{\hat{\jmath}}(x^*). \tag{6.2}$$

Therefore, we can write

$$\begin{aligned} D_{\hat{\jmath}}(x^k) &= \min\{x^k_{\hat{\jmath}} - l_{\hat{\jmath}}, u_{\hat{\jmath}} - x^k_{\hat{\jmath}}\} \overset{(6.2)}{>} \min\{x^*_{\hat{\jmath}} - l_{\hat{\jmath}}, u_{\hat{\jmath}} - x^*_{\hat{\jmath}}\} - \frac{\tau}{\tau+1} D_{\hat{\jmath}}(x^*) \\ &= D_{\hat{\jmath}}(x^*) - \frac{\tau}{\tau+1} D_{\hat{\jmath}}(x^*) = \frac{1}{\tau+1} D_{\hat{\jmath}}(x^*). \end{aligned} \tag{6.3}$$

Arguing by contradiction, assume now that $j(k) \in \mathscr{Z}(x^*)$, that is,

$$x^*_{j(k)} \in \{l_{j(k)}, u_{j(k)}\}. \tag{6.4}$$

We obtain

$$\begin{aligned} D_{j(k)}(x^k) &= \min\{x^k_{j(k)} - l_{j(k)}, u_{j(k)} - x^k_{j(k)}\} \overset{(6.4)}{\le} |x^k_{j(k)} - x^*_{j(k)}| \le \|x^k - x^*\|_\infty \\ &< \frac{\tau}{\tau+1} D^{\max}(x^*) = \frac{\tau}{\tau+1} D_{\hat{\jmath}}(x^*) \overset{(6.3)}{<} \tau D_{\hat{\jmath}}(x^k) \overset{(3.2)}{\le} \tau D^k, \end{aligned}$$

contradicting (3.3). ∎

Combining the above proposition with (5.2), the next result immediately follows.

PROPOSITION 6.3. *Let $\{x^k\}$ be a sequence of points produced by AC2CD and assume that $\lim_{k\to\infty} x^k = x^*$. There exists an iteration k^z such that, for all $k \ge k^z$,*

$$l_{j(k)} < z^{k,i}_{j(k)} < u_{j(k)}, \quad i = 1, \dots, n+1.$$

Now, we are ready to show that (6.1) holds for all sufficiently large iterations. Our analysis takes inspiration from the one in [43] for proximal gradient methods, where it is proved that the active set is identified in a neighborhood of the optimal solution under the non-degeneracy assumption. That neighborhood is defined in [43] by using a problem-dependent constant related on "the amount of degeneracy" of the optimal solution.

Here, for a stationary point x^* such that $\mathscr{Z}^+(x^*) \neq \emptyset$, we define the following positive constant, measuring the "minimum amount of strict complementarity" at x^*:

$$\zeta(x^*) = \min_{i \in \mathscr{Z}^+(x^*)} |\nabla_i f(x^*) - \lambda^*|, \tag{6.5}$$

where λ^* is the KKT multiplier associated to x^*, according to (2.7).

THEOREM 6.4. *Let $\{x^k\}$ be a sequence of points produced by AC2CD and assume that $\lim_{k\to\infty} x^k = x^*$. Let $\bar{k}$ be the first outer iteration such that*

$$\|z^{k,i} - x^*\| < \frac{\zeta(x^*)}{2L + \max\left\{\dfrac{1}{A_l}, \dfrac{L^{max}}{2(1-\gamma)}\right\}}, \quad i = 1, \dots, n, \quad \forall\, k \ge \bar{k}, \tag{6.6}$$

where $\zeta(x^*) > 0$ is the minimum strict complementarity measure at x^*, defined as in (6.5), L is the Lipschitz constant of ∇f, $A_l > 0$ is the lower bound on the parameter $A^{k,i}$ used to compute the initial stepsize $\Delta^{k,i}$ in the Armijo line search (see (3.6) and (3.8)), $\gamma \in (0,1)$ is the parameter for sufficient decrease in the Armijo line search and $L^{max} > 0$ is the maximum among the local Lipschitz constants $L_{i,j}$, defined as in (2.4).

Also assume that $\bar{k} \ge \max\{k^j, k^z\}$, where k^j is the first outer iteration such that $j(k) \notin \mathscr{Z}(x^*)$ for all $k \ge k^j$, defined as in Proposition 6.2, and k^z is the first outer iteration such that $l_{j(k)} < z^{k,i}_{j(k)} < u_{j(k)}$, $i = 1, \dots, n+1$, for all $k \ge k^z$, defined as in Proposition 6.3.

Then, for all $k > \bar{k}$ we have that

$$x^k_h = x^*_h, \quad \forall\, h \in \mathscr{Z}^+(x^*).$$

Proof. Consider an outer iteration $k \ge \bar{k}$ and any index $h \in \mathscr{Z}^+(x^*)$. Moreover, let (k,i) be the inner iteration where $p_i^k = h$. Without loss of generality, let us assume that $x_h^* = l_h$ (the proof for the case where $x_h^* = u_h$ is analogous). Namely,

$$x_h^* = l_h \quad \text{and} \quad \nabla_h f(x^*) > \lambda^*, \tag{6.7}$$

where λ^* is the KKT multiplier associated to x^*, according to the stationary conditions (2.7). Since $k \ge k^j$, from Proposition 6.2 we have that

$$j(k) \notin \mathscr{Z}(x^*), \tag{6.8}$$

implying that $h \ne j(k)$. Then, using (6.8) and the stationary conditions (2.7), we get $\lambda^* = \nabla_{j(k)} f(x^*)$. Recalling the definition of $\zeta(x^*)$, it follows that

$$\zeta(x^*) \le \nabla_h f(x^*) - \nabla_{j(k)} f(x^*).$$

Moreover, from the definition of $g^{k,i}$ given in (3.4) we can write

$$\begin{aligned}\nabla_h f(x^*) - \nabla_{j(k)} f(x^*) + g^{k,i} &= \nabla_h f(x^*) - \nabla_{j(k)} f(x^*) + \nabla_{j(k)} f(z^{k,i}) - \nabla_h f(z^{k,i}) \\ &\le |\nabla_h f(x^*) - \nabla_h f(z^{k,i})| + |\nabla_{j(k)} f(z^{k,i}) - \nabla_{j(k)} f(x^*)| \\ &\le 2\|\nabla f(x^*) - \nabla f(z^{k,i})\| \le 2L\|x^* - z^{k,i}\|,\end{aligned}$$

and then,

$$\zeta(x^*) \le -g^{k,i} + 2L\|x^* - z^{k,i}\|. \tag{6.9}$$

Now, we can rewrite (6.6) by multiplying the numerator and the denominator of the right-hand side by $\max\left\{\frac{1}{A_l}, \frac{L^{\max}}{2(1-\gamma)}\right\}^{-1} = \min\left\{A_l, \frac{2(1-\gamma)}{L^{\max}}\right\}$, obtaining

$$\|z^{k,i} - x^*\| < \frac{\zeta(x^*) \min\left\{A_l, \frac{2(1-\gamma)}{L^{\max}}\right\}}{2L \min\left\{A_l, \frac{2(1-\gamma)}{L^{\max}}\right\} + 1}.$$

Multiplying both sides of this inequality by the denominator of the right-hand side, we can write

$$\begin{aligned}\|z^{k,i} - x^*\| &= (\zeta(x^*) - 2L\|z^{k,i} - x^*\|) \min\left\{A_l, \frac{2(1-\gamma)}{L^{\max}}\right\} \\ &\overset{(6.9)}{\le} -g^{k,i} \min\left\{A_l, \frac{2(1-\gamma)}{L^{\max}}\right\}.\end{aligned} \tag{6.10}$$

It follows that $g^{k,i} \le 0$. If $g^{k,i} = 0$, we have

$$x_h^{k+1} \overset{(5.1)}{=} z_h^{k,i+1} \overset{(3.4)}{=} z_h^{k,i} \overset{(6.10)}{=} x_h^*,$$

and the desired result is thus obtained. Now assume that $g^{k,i} < 0$. We can upper bound the largest feasible stepsize $\bar{\alpha}^{k,i}$ as follows:

$$\begin{aligned}\bar{\alpha}^{k,i} &\overset{(3.7)}{\le} -\frac{z_h^{k,i} - l_h}{g^{k,i}} \overset{(6.7)}{=} -\frac{z_h^{k,i} - x_h^*}{g^{k,i}} \overset{(6.10)}{<} \min\left\{A_l, \frac{2(1-\gamma)}{L^{\max}}\right\} \\ &\overset{(3.8)}{\le} \min\left\{A^{k,i}, \frac{2(1-\gamma)}{L^{\max}}\right\},\end{aligned} \tag{6.11}$$

implying that $\bar{\alpha}^{k,i} < A^{k,i}$. Taking into account that the initial stepsize $\Delta^{k,i}$ in the Armijo line search is chosen as in (3.6), we have that $\Delta^{k,i} = \bar{\alpha}^{k,i}$. So, using again (6.11) we obtain

$$\Delta^{k,i} = \bar{\alpha}^{k,i} < \frac{2(1-\gamma)}{L^{\max}} \le \frac{2(1-\gamma)}{L_{h,j(k)}}.$$

From Lemma 5.1 we get that $\alpha^{k,i} = \bar{\alpha}^{k,i}$. Since $\bar{\alpha}^{k,i}$ is the largest feasible stepsize along $d^{k,i}$, (at least) one variable between $z_h^{k,i+1}$ and $z_{j(k)}^{k,i+1}$ will be at the lower or the upper bound. Using the fact that $k \ge k^z$, from Proposition 6.3 we have that $z_{j(k)}^{k,i+1} \in (l_{j(k)}, u_{j(k)})$, and then $z_h^{k,i+1}$ will be necessarily at the lower or the upper bound. Since $g^{k,i} < 0$, from the definition of the search direction given in (3.4) it follows that $z_h^{k,i+1} = l_h$. Using (5.1) and (6.7), we finally have that $z_h^{k,i+1} = x_h^{k+1}$ and $l_h = x_h^*$, yielding to the desired result. □

Remark 6.5. From (5.2), there must exist an outer iteration $\bar{k}$ such that (6.6) holds, provided the whole sequence $\{x^k\}$ converges to x^* and $\mathscr{Z}^+(x^*) \neq \emptyset$.

7. Active-set complexity.

In this section, the main result of the paper is presented: under convexity of f and a quadratic growth condition (satisfied by any strongly convex function), it is possible to compute the maximum number of iterations required by AC2CD to identify the active set, thus extending what obtained in the previous section. Using the definition given in [43], we refer to the maximum number of iterations required to identify the active set as "active-set complexity".

To obtain the desired result, we first show how choosing the initial stepsize in the Armijo line search, in order to meet an additional requirement. Then, we will show non-asymptotic sublinear convergence rate of AC2CD, which, combined with Theorem 6.4, will lead to the active-set complexity of the algorithm.

7.1. Initial stepsize in the Armijo line search.

To obtain non-asymptotic sublinear convergence rate of AC2CD, for all $k \ge 0$ we need to satisfy

$$l_{j(k)} < z_{j(k)}^{k,i} < u_{j(k)}, \quad i = 1, \dots, n+1. \tag{7.1}$$

Note that, in general, (7.1) holds only for sufficiently large k (see the proof of Theorem 1 in [12]). To satisfy (7.1) for all $k \ge 0$ we can use sufficiently small stepsizes in all the inner iterations, exploiting the fact that $x_{j(k)}^k = z_{j(k)}^{k,1} \in (l_{j(k)}, u_{j(k)})$ for all $k \ge 0$. In particular, to obtain a small stepsize $\alpha^{k,i}$ from the Armijo line search we must choose a small value of the initial stepsize $\Delta^{k,i}$. Taking into account (3.6), this means that we must use a small value of $A^{k,i}$. Anyway, we have to keep in mind that $A^{k,i}$ must satisfy (3.8) as well. A possible strategy is setting $A_u > 0$, $\epsilon \in (0,1)$ and, at every inner iteration (k,i), computing

$$A^{k,i} = \begin{cases} \min\{\hat{\alpha}^{k,i}, A_u\}, & \text{if } g^{k,i} \neq 0 \text{ (i.e., if } d^{k,i} \text{ is a non-zero direction)}, \\ A_u, & \text{otherwise}, \end{cases} \tag{7.2}$$

where $\hat{\alpha}^{k,i}$ is the stepsize such that $D_{j(k)}(z^{k,i} + \hat{\alpha}^{k,i} d^{k,i}) = \epsilon D_{j(k)}(z^{k,i})$ when $g^{k,i} \neq 0$. Note that $\hat{\alpha}^{k,i}$ may be infeasible and/or infinity. Since $\alpha^{k,i} \le A^{k,i} \le \hat{\alpha}^{k,i}$, it follows that $D_{j(k)}(z^{k,i+1}) = D_{j(k)}(z^{k,i} + \alpha^{k,i} d^{k,i}) \ge D_{j(k)}(z^{k,i} + \hat{\alpha}^{k,i} d^{k,i}) \ge \epsilon D_{j(k)}(z^{k,i})$. Consequently,

$$D_{j(k)}(z^{k,i+1}) \ge \epsilon^i D_{j(k)}(z^{k,1}) = \epsilon^i D_{j(k)}(x^k) > 0, \quad i = 1, \dots, n. \tag{7.3}$$

Then, this choice of $A^{k,i}$ satisfies (7.1) for all $k \ge 0$. To show that it also satisfies (3.8), we have to explicitly write the expression of $\hat{\alpha}^{k,i}$, which can be obtained by simple calculations (recall that $\hat{\alpha}^{k,i}$ is defined only when $g^{k,i} \neq 0$):

If $g^{k,i} > 0$,

$$\hat{\alpha}^{k,i} = \begin{cases} (1-\epsilon)D_{j(k)}(z^{k,i})/g^{k,i}, & \text{if } D_{j(k)}(z^{k,i}) = z^{k,i}_{j(k)} - l_{j(k)}, \\ [z^{k,i}_{j(k)} - l_{j(k)} - \epsilon D_{j(k)}(z^{k,i})]/g^{k,i}, & \text{otherwise;} \end{cases}$$

else if $g^{k,i} < 0$,

$$\hat{\alpha}^{k,i} = \begin{cases} (1-\epsilon)D_{j(k)}(z^{k,i})/|g^{k,i}|, & \text{if } D_{j(k)}(z^{k,i}) = u_{j(k)} - z^{k,i}_{j(k)}, \\ [u_{j(k)} - z^{k,i}_{j(k)} - \epsilon D_{j(k)}(z^{k,i})]/|g^{k,i}|, & \text{otherwise.} \end{cases}$$

We see that, when $g^{k,i} \neq 0$, we have $\hat{\alpha}^{k,i} \ge (1-\epsilon)D_{j(k)}(z^{k,i})/|g^{k,i}|$. Using (7.2) and (7.3), it follows that

$$\min\left\{\frac{(1-\epsilon)\epsilon^{i-1}D_{j(k)}(x^k)}{|g^{k,i}|}, A_u\right\} \le A^{k,i} \le A_u, \quad \text{if } g^{k,i} \neq 0.$$

Then, (3.8) is satisfied with a proper value of A_l which can be easily obtained, since any non-zero $|g^{k,i}|$ is less than or equal to $\max_{i,j=1,\dots,n}\{\nabla_j f(x) - \nabla_i f(x) \colon x \in \mathcal{L}^0\}$ (which is finite by the assumption that the level set $\mathcal{L}^0$ is compact) and, from (3.3), we have $D_{j(k)}(x^k) \ge \tau \min_{x\in\mathcal{L}^0} \max_{i=1,\dots,n} D_i(x)$ (which is positive by Assumption 1).

Many other strategies can be used to compute a value of $A^{k,i}$ that satisfies all the required conditions. It is important to note that, in practice, $A^{k,i}$ should not be too small compared to the largest feasible stepsize $\bar{\alpha}^{k,i}$ (for a non-zero direction $d^{k,i}$), otherwise the Armijo line search may produce extremely small stepsizes which can dramatically slow down the algorithm. For example, ϵ should be sufficiently smaller than 1 in the above described strategy.

In the rest of this section, we will assume $A^{k,i}$ to be computed in order to satisfy, together with (3.8), condition (7.1) for all $k \ge 0$.

7.2. Convergence rate analysis. In this subsection we show that, when f is convex, AC2CD has a non-asymptotic sublinear convergence rate. Let us remark that the results reported here are completely different from those given in [12], where a linear rate was obtained, but asymptotically, whereas a non-asymptotic linear rate was proved only when there are no bounds on the variables (both results are not useful in the analysis of the active-set complexity).

Our results here are obtained by adapting the analysis of the block coordinate gradient projection method in [2] for minimization problems over the Cartesian product of closed convex sets. In particular, with respect to [2], major difficulties in our analysis come from the presence of the coupling constraint in the problem and the absence of projection operations in the algorithm. In such a context, the next lemma establishes a useful property of AC2CD.

LEMMA 7.1. *For all $x^* \in X^*$, at every inner iteration (k,i) of AC2CD we have that*

$$g^{k,i}(x^*_{p^k_i} - z^{k,i+1}_{p^k_i}) \le \max\left\{\frac{1}{A_l}, \frac{L^{max}}{2\delta(1-\gamma)}\right\}|z^{k,i+1}_{p^k_i} - z^{k,i}_{p^k_i}||x^*_{p^k_i} - z^{k,i+1}_{p^k_i}|,$$

where $A_l > 0$ is the lower bound on the parameter $A^{k,i}$ used to compute the initial stepsize $\Delta^{k,i}$ in the Armijo line search (see (3.6) and (3.8)), $\gamma \in (0,1)$ is the parameter *for sufficient decrease in the Armijo line search,* $\delta \in (0,1)$ *is the reduction parameter in the Armijo line search and* $L^{max} > 0$ *is the maximum among the local Lipschitz constants* $L_{i,j}$*, defined in* (2.4).

Proof. Consider any inner iteration (k,i). The result is trivial if $g^{k,i} = 0$, so we assume that $g^{k,i} \neq 0$ and distinguish two possible cases.

(i) First, assume that $\alpha^{k,i} = \bar{\alpha}^{k,i}$, that is, the largest feasible stepsize is used. This means that (at least) one variable between $z^{k,i+1}_{p^k_i}$ and $z^{k,i+1}_{j(k)}$ will be at the lower or the upper bound. Recalling that (7.1) holds for all $k \geq 0$, necessarily $z^{k,i+1}_{p^k_i}$ will be at the lower or the upper bound. Using the definition of $\bar{\alpha}^{k,i}$ given in (3.7), it follows that either $z^{k,i+1}_{p^k_i} = u_{p^k_i}$ if $g^{k,i} > 0$, or $z^{k,i+1}_{p^k_i} = l_{p^k_i}$ if $g^{k,i} < 0$, implying that $g^{k,i}(x^*_{p^k_i} - z^{k,i+1}_{p^k_i}) \leq 0$ and the desired result is obtained.

(ii) Now, assume that $\alpha^{k,i} < \bar{\alpha}^{k,i}$, which implies that $\bar{\alpha}^{k,i} > 0$ and, from (3.9), that $z^{k,i+1} \neq z^{k,i}$. Since $z^{k,i+1}_{p^k_i} = z^{k,i}_{p^k_i} + \alpha^{k,i} g^{k,i}$, it follows that $\alpha^{k,i} g^{k,i} \neq 0$. Recalling the definition of $g^{k,i}$ given in (3.4), this implies that $p^k_i \neq j(k)$. Moreover, we can write

$$g^{k,i}(x^*_{p^k_i} - z^{k,i+1}_{p^k_i}) = \frac{(z^{k,i+1}_{p^k_i} - z^{k,i}_{p^k_i})(x^*_{p^k_i} - z^{k,i+1}_{p^k_i})}{\alpha^{k,i}} \leq \frac{|z^{k,i+1}_{p^k_i} - z^{k,i}_{p^k_i}||x^*_{p^k_i} - z^{k,i+1}_{p^k_i}|}{\alpha^{k,i}}.$$

So, to obtain the desired result we have to show that

$$\alpha^{k,i} \geq \min\left\{A_l, \frac{2\delta(1-\gamma)}{L^{\max}}\right\}.$$

To this extent, let us distinguish two further subcases, depending on whether $\Delta^{k,i} = \bar{\alpha}^{k,i}$ or $\Delta^{k,i} < \bar{\alpha}^{k,i}$, according to the definition of $\Delta^{k,i}$ given in (3.6).

- If $\Delta^{k,i} = \bar{\alpha}^{k,i}$, then $\alpha^{k,i} < \Delta^{k,i}$ (recall that we are considering the case $\alpha^{k,i} < \bar{\alpha}^{k,i}$) and, from Lemma 5.1, it follows that

$$\alpha^{k,i} > \frac{2\delta(1-\gamma)}{L_{p^k_i,j(k)}} \geq \min\left\{A_l, \frac{2\delta(1-\gamma)}{L_{p^k_i,j(k)}}\right\} \geq \min\left\{A_l, \frac{2\delta(1-\gamma)}{L^{\max}}\right\}.$$

- If $\Delta^{k,i} < \bar{\alpha}^{k,i}$, from (3.6) we have $\Delta^{k,i} = A^{k,i}$. Using Lemma 5.1 it follows that

$$\begin{aligned}\alpha^{k,i} &\geq \min\left\{\Delta^{k,i}, \frac{2\delta(1-\gamma)}{L_{p^k_i,j(k)}}\right\} = \min\left\{A^{k,i}, \frac{2\delta(1-\gamma)}{L_{p^k_i,j(k)}}\right\} \\ &\geq \min\left\{A_l, \frac{2\delta(1-\gamma)}{L_{p^k_i,j(k)}}\right\} \geq \min\left\{A_l, \frac{2\delta(1-\gamma)}{L^{\max}}\right\}.\end{aligned}$$

∎

Now, we give a first result on the decrease in the objective function at every outer iteration.

PROPOSITION 7.2. *At every outer iteration* k *of AC2CD we have that*

$$f(x^k) - f(x^{k+1}) \geq \frac{\gamma}{A_u}\|x^{k+1} - x^k\|^2_{\langle j(k)\rangle},$$

where $A_u > 0$ *is the upper bound on the parameter* $A^{k,i}$ *used to compute the initial stepsize* $\Delta^{k,i}$ *in the Armijo line search (see* (3.6) *and* (3.8)*) and* $\gamma \in (0,1)$ *is the parameter for sufficient decrease in the Armijo line search.*

Proof. First we show that at, every inner iteration (k, i), we have

$$f(z^{k,i}) - f(z^{k,i+1}) \geq \frac{\gamma}{A_u}(z^{k,i+1}_{p^k_i} - z^{k,i}_{p^k_i})^2. \tag{7.4}$$

If $\alpha^{k,i} = 0$, then $z^{k,i+1} = z^{k,i}$ and (7.4) trivially holds. If $\alpha^{k,i} > 0$, from the instructions of the Armijo line search it follows that $f(z^{k,i+1}) \leq f(z^{k,i}) + \gamma\alpha^{k,i}\nabla f(z^{k,i})^T d^{k,i}$. Using (3.5), we can write

$$f(z^{k,i+1}) \leq f(z^{k,i}) - \gamma\alpha^{k,i}(g^{k,i})^2 = f(z^{k,i}) - \frac{\gamma}{\alpha^{k,i}}(\alpha^{k,i}g^{k,i})^2.$$

Since $z^{k,i+1}_{p^k_i} = z^{k,i} + \alpha^{k,i}g^{k,i}$ and $\alpha^{k,i} \leq A_u$, we obtain (7.4). Hence, we have

$$\begin{aligned}
f(x^k) - f(x^{k+1}) &= \sum_{i=1}^{n}[f(z^{k,i}) - f(z^{k,i+1})] \overset{(7.4)}{\geq} \frac{\gamma}{A_u}\sum_{i=1}^{n}(z^{k,i+1}_{p^k_i} - z^{k,i}_{p^k_i})^2 \\
&= \frac{\gamma}{A_u}\sum_{i\colon p^k_i \neq j(k)}(z^{k,i+1}_{p^k_i} - z^{k,i}_{p^k_i})^2 \overset{(5.1)}{=} \frac{\gamma}{A_u}\sum_{i\colon p^k_i \neq j(k)}(x^{k+1}_{p^k_i} - x^{k}_{p^k_i})^2 \\
&= \frac{\gamma}{A_u}\|x^{k+1} - x^k\|^2_{\langle j(k)\rangle},
\end{aligned}$$

where, in the second equality, we have used the fact that $z^{k,i+1}_{j(k)} = z^{k,i}_{j(k)}$ when $p^k_i = j(k)$, according to the definition of the search direction $d^{k,i}$ given in (3.4). ∎

In the rest of this section, the objective function will be required to be convex over $\mathbb{R}^n$ and its optimal value for problem (2.1) will be denoted by f^*. Let us also define the following constants (which are finite under convexity of f, since this implies $X^* \subseteq \mathcal{L}^0$, where the level set $\mathcal{L}^0$ is assumed to be non-empty and compact):

$$R^0 = \max_{\substack{j=1,\dots,n \\ x\in\mathcal{L}^0 \\ x^*\in X^*}} \|x - x^*\|_{\langle j\rangle}, \tag{7.5}$$

$$G^* = \max_{\substack{i,j=1,\dots,n \\ x^*\in X^*}} [\nabla_j f(x^*) - \nabla_i f(x^*)]. \tag{7.6}$$

We see that R^0 is the maximum distance between a point in the level set $\mathcal{L}^0$ and a point in X^*, where the distance is measured in terms of the pseudometrics induced by the seminorms $\|\cdot\|_{\langle j\rangle}$ (the latter can be upper bounded by the Euclidean norm, see (2.11)). From the KKT conditions (2.7), we also note that G^* is related to the minimum strict complementarity measure $\zeta(x^*)$ defined in (6.5), in the sense that, if $\mathscr{Z}^+(x^*) \neq \emptyset$ for some $x^* \in X^*$, then $G^* \geq \zeta(x^*) > 0$, while, if $\mathscr{Z}^+(x^*) = \emptyset$ for all $x^* \in X^*$, then $G^* = 0$ and $\zeta(x^*)$ is not defined for any $x^* \in X^*$. We can interpret G^* as a measure of the "maximum amount of strict complementarity" over the set X^*.

We now state a result which, for every outer iteration, relates the decrease in the objective function with the optimization error.

PROPOSITION 7.3. *Assume that f is convex over $\mathbb{R}^n$. Then, at every outer iteration k of AC2CD we have that*

$$f(x^k) - f(x^{k+1}) \geq \frac{\gamma(f(x^{k+1}) - f^*)^2}{A_u(n-1)\left[\left(\max\left\{\dfrac{1}{A_l}, \dfrac{L^{max}}{2\delta(1-\gamma)}\right\} + 2\hat{L}^{max}\right)R^0 + G^*\right]^2},$$

where $A_l > 0$ and $A_u > 0$ are the lower and the upper bound, respectively, on the parameter $A^{k,i}$ used to compute the initial stepsize $\Delta^{k,i}$ in the Armijo line search (see (3.6) *and* (3.8)*), $\delta \in (0,1)$ is the reduction parameter in the Armijo line search, $\gamma \in (0,1)$ is the parameter for sufficient decrease in the Armijo line search, $L^{max} > 0$ is the maximum among the local Lipschitz constants $L_{i,j}$, defined as in* (2.4)*, $\hat{L}^{max} > 0$ is the maximum among the constants $L_j = \sum_{i=1}^n L_{i,j}$, defined as in* (2.6)*, $R^0 \ge 0$ is the maximum distance between a point in the level set $\mathcal{L}^0$ and an optimal solution, defined as in* (7.5)*, and $G^* \ge 0$ is the maximum strict complementarity measure over X^*, defined as in* (7.6)*.*

Proof. Let x^* be an optimal solution of problem (2.1) and consider any inner iteration (k,i). From the definition of the search direction $d^{k,i}$ given in (3.4), we have that $z^{k,i+1}_{p^k_i} \ge z^{k,i}_{p^k_i}$ if $g^{k,i} \ge 0$, and $z^{k,i+1}_{p^k_i} \le z^{k,i}_{p^k_i}$ if $g^{k,i} \le 0$. Namely, $g^{k,i}(z^{k,i}_{p^k_i} - z^{k,i+1}_{p^k_i}) \le 0$ and, using (5.1), we can write $g^{k,i}(x^k_{p^k_i} - x^{k+1}_{p^k_i}) \le 0$. Then,

$$\begin{aligned} g^{k,i}(x^k_{p^k_i} - x^*_{p^k_i}) &\le g^{k,i}(x^{k+1}_{p^k_i} - x^*_{p^k_i}) \\ &= [\nabla_{p^k_i} f(x^{k+1}) - \nabla_{j(k)} f(x^{k+1})](x^*_{p^k_i} - x^{k+1}_{p^k_i}) \\ &\quad + [\nabla_{p^k_i} f(x^{k+1}) - \nabla_{j(k)} f(x^{k+1}) + g^{k,i}](x^{k+1}_{p^k_i} - x^*_{p^k_i}). \end{aligned}$$

Using Corollary 5.4 with $v = x^{k+1}$, we have that

$$\begin{aligned} \nabla_{p^k_i} f(x^{k+1}) - \nabla_{j(k)} f(x^{k+1}) + g^{k,i} &\le \hat{L}^{\max} \|z^{k,i} - x^{k+1}\|_{\langle j(k) \rangle} \\ &\overset{(5.1)}{\le} \hat{L}^{\max} \|x^k - x^{k+1}\|_{\langle j(k) \rangle}. \end{aligned}$$

It follows that

$$\begin{aligned} g^{k,i}(x^k_{p^k_i} - x^*_{p^k_i}) \le{}& [\nabla_{p^k_i} f(x^{k+1}) - \nabla_{j(k)} f(x^{k+1})](x^*_{p^k_i} - x^{k+1}_{p^k_i}) \\ & + \hat{L}^{\max} \|x^k - x^{k+1}\|_{\langle j(k) \rangle} |x^*_{p^k_i} - x^{k+1}_{p^k_i}|. \end{aligned}$$

Summing these inequalities, we obtain

$$\begin{aligned} \sum_{i:\, p^k_i \ne j(k)} g^{k,i}(x^k_{p^k_i} - x^*_{p^k_i}) &\le \sum_{i:\, p^k_i \ne j(k)} [\nabla_{p^k_i} f(x^{k+1}) - \nabla_{j(k)} f(x^{k+1})](x^*_{p^k_i} - x^{k+1}_{p^k_i}) \\ &\quad + \hat{L}^{\max} \|x^k - x^{k+1}\|_{\langle j(k) \rangle} \sum_{i:\, p^k_i \ne j(k)} |x^*_{p^k_i} - x^{k+1}_{p^k_i}| \\ &\overset{(2.10)}{\le} \sum_{i:\, p^k_i \ne j(k)} [\nabla_{p^k_i} f(x^{k+1}) - \nabla_{j(k)} f(x^{k+1})](x^*_{p^k_i} - x^{k+1}_{p^k_i}) \\ &\quad + \sqrt{n-1}\, R^0 \hat{L}^{\max} \|x^k - x^{k+1}\|_{\langle j(k) \rangle}. \end{aligned}$$

Using Lemma 5.2 with $v = \nabla f(x^{k+1})$, $x' = x^*$ and $x'' = x^{k+1}$, we can write

$$\begin{aligned} \nabla f(x^{k+1})^T (x^* - x^{k+1}) &= \langle \nabla f(x^{k+1}) - \nabla_{j(k)} f(x^{k+1}) e, x^* - x^{k+1} \rangle_{j(k)} \\ &= \sum_{i:\, p^k_i \ne j(k)} [\nabla_{p^k_i} f(x^{k+1}) - \nabla_{j(k)} f(x^{k+1})](x^*_{p^k_i} - x^{k+1}_{p^k_i}), \end{aligned}$$

and then,

$$\sum_{i\colon p_i^k \neq j(k)} g^{k,i}(x^k_{p_i^k} - x^*_{p_i^k}) \le \nabla f(x^{k+1})^T(x^* - x^{k+1}) + \sqrt{n-1}\, R^0 \hat{L}^{\max} \|x^k - x^{k+1}\|_{\langle j(k)\rangle}.$$

From the convexity of f we have that $f(x^{k+1}) - f^* \le \nabla f(x^{k+1})^T(x^{k+1} - x^*)$. Hence,

$$\begin{aligned} f(x^{k+1}) - f^* &\le \sum_{i\colon p_i^k \neq j(k)} g^{k,i}(x^*_{p_i^k} - x^k_{p_i^k}) + \sqrt{n-1}\, R^0 \hat{L}^{\max} \|x^k - x^{k+1}\|_{\langle j(k)\rangle} \\ &= \sum_{i\colon p_i^k \neq j(k)} g^{k,i}(x^*_{p_i^k} - x^{k+1}_{p_i^k}) + \sum_{i\colon p_i^k \neq j(k)} g^{k,i}(x^{k+1}_{p_i^k} - x^k_{p_i^k}) \\ &\quad + \sqrt{n-1}\, R^0 \hat{L}^{\max} \|x^k - x^{k+1}\|_{\langle j(k)\rangle}. \end{aligned}$$

Using (5.1) and Lemma 7.1, for all i such that $p_i^k \neq j(k)$ we can write

$$g^{k,i}(x^*_{p_i^k} - x^{k+1}_{p_i^k}) \le \max\left\{\frac{1}{A_l}, \frac{L^{\max}}{2\delta(1-\gamma)}\right\} |x^{k+1}_{p_i^k} - x^k_{p_i^k}||x^*_{p_i^k} - x^{k+1}_{p_i^k}|.$$

Therefore,

$$\begin{aligned} f(x^{k+1}) - f^* \le& \max\left\{\frac{1}{A_l}, \frac{L^{\max}}{2\delta(1-\gamma)}\right\} \sum_{i\colon p_i^k \neq j(k)} |x^{k+1}_{p_i^k} - x^k_{p_i^k}||x^*_{p_i^k} - x^{k+1}_{p_i^k}| \\ &+ \sum_{i\colon p_i^k \neq j(k)} g^{k,i}(x^{k+1}_{p_i^k} - x^k_{p_i^k}) \\ &+ \sqrt{n-1} R^0 \hat{L}^{\max} \|x^k - x^{k+1}\|_{\langle j(k)\rangle}. \end{aligned} \tag{7.7}$$

To obtain the desired result, now we upper bound the two summations in the right-hand side of (7.7) by appropriate constants.

- As for the first summation in the right-hand side of (7.7), using the fact that $|x^*_{p_i^k} - x^{k+1}_{p_i^k}| \le \|x^* - x^{k+1}\|_{\langle j(k)\rangle}$ by (2.9), we can write

$$\begin{aligned} \sum_{i\colon p_i^k \neq j(k)} |x^{k+1}_{p_i^k} - x^k_{p_i^k}||x^*_{p_i^k} - x^{k+1}_{p_i^k}| \quad &\le \quad R^0 \sum_{i\colon p_i^k \neq j(k)} |x^{k+1}_{p_i^k} - x^k_{p_i^k}| \\ &\overset{(2.10)}{\le} \sqrt{n-1} R^0 \|x^{k+1} - x^k\|_{\langle j(k)\rangle}. \end{aligned} \tag{7.8}$$

- As for the second summation in the right-hand side of (7.7), from the triangular inequality we have that

$$g^{k,i} \le |g^{k,i} + \nabla_{p_i^k} f(x^*) - \nabla_{j(k)} f(x^*)| + |\nabla_{p_i^k} f(x^*) - \nabla_{j(k)} f(x^*)|,$$

  and then, using Corollary 5.4 with $v = x^*$,

$$g^{k,i} \le L_{j(k)} \|x^* - z^{k,i}\|_{\langle j(k)\rangle} + |\nabla_{p_i^k} f(x^*) - \nabla_{j(k)} f(x^*)| \le \hat{L}^{\max} R^0 + G^*.$$

  Taking into account (2.10), we get

$$\sum_{i\colon p_i^k \neq j(k)} g^{k,i}(x^{k+1}_{p_i^k} - x^k_{p_i^k}) \le (\hat{L}^{\max} R^0 + G^*)\sqrt{n-1} \|x^{k+1} - x^k\|_{\langle j(k)\rangle}. \tag{7.9}$$

Combining (7.7) with (7.8) and (7.9), we have that

$$f(x^{k+1}) - f^* \le \sqrt{n-1}\left[\left(\max\left\{\frac{1}{A_l}, \frac{L^{\max}}{2\delta(1-\gamma)}\right\} + 2\hat{L}^{\max}\right)R^0 + G^*\right]\|x^{k+1} - x^k\|_{\langle j(k)\rangle}.$$

Using Proposition 7.2, the desired result is finally obtained. ∎

We are now ready to show the non-asymptotic sublinear convergence rate of AC2CD.

THEOREM 7.4. *Assume that f is convex over $\mathbb{R}^n$. Then, at every outer iteration $k \ge 1$ of AC2CD we have that*

$$f(x^k) - f^* \le \frac{C}{k},$$

where C is equal to

$$\sqrt{n-1}\max\left\{\frac{3A_u\sqrt{n-1}}{2\gamma}, \frac{1}{L^{max}}\right\}\left[\left(\max\left\{\frac{1}{A_l}, \frac{L^{max}}{2\delta(1-\gamma)}\right\} + 2\hat{L}^{max}\right)R^0 + 2G^*\right]^2,$$

$A_l > 0$ and $A_u > 0$ are the lower and the upper bound, respectively, on the parameter $A^{k,i}$ used to compute the initial stepsize $\Delta^{k,i}$ in the Armijo line search (see (3.6) and (3.8)), $\delta \in (0,1)$ is the reduction parameter in the Armijo line search, $\gamma \in (0,1)$ is the parameter for sufficient decrease in the Armijo line search, $L^{max} > 0$ is the maximum among the local Lipschitz constants $L_{i,j}$, defined as in (2.4), $\hat{L}^{max} > 0$ is the maximum among the constants $L_j = \sum_{i=1}^n L_{i,j}$, defined as in (2.6), $R^0 \ge 0$ is the maximum distance between a point in the level set $\mathcal{L}^0$ and an optimal solution, defined as in (7.5), and $G^* \ge 0$ is the maximum strict complementarity measure over X^*, defined as in (7.6).

Proof. Consider a sequence $\{a^k\}$ of nonnegative scalars such that $a^k - a^{k+1} \ge \beta(a^{k+1})^2$, for all $k \ge 0$, with $\beta > 0$. From Lemma 6.2 in [2] we have that, if $a^1 \le 3/(2\beta)$ and $a^2 \le 3/(4\beta)$, then $a^k \le 3/(2\beta k)$, for all $k \ge 1$. Using $a^k = f(x^k) - f^*$, in view of Proposition 7.3 we have that $a^k - a^{k+1} \ge \beta(a^{k+1})^2$ with $\beta \ge 3/(2C)$. It follows that the desired result is obtained if

$$f(x^1) - f^* \le C \quad \text{and} \quad f(x^2) - f^* \le \frac{C}{2}. \tag{7.10}$$

To show that (7.10) holds, by definition of C we first write

$$\begin{aligned} C &\ge \frac{\sqrt{n-1}}{L^{\max}}\left[\left(\max\left\{\frac{1}{A_l}, \frac{L^{\max}}{2\delta(1-\gamma)}\right\} + 2\hat{L}^{\max}\right)R^0 + 2G^*\right]^2 \\ &\ge \frac{\sqrt{n-1}}{L^{\max}}\left[\left(\frac{L^{\max}}{2} + 2\hat{L}^{\max}\right)R^0 + 2G^*\right]^2, \end{aligned} \tag{7.11}$$

where the last inequality follows from the fact that $2\delta(1-\gamma) \le 2$, since $\delta, \gamma \in (0,1)$. Now, we use the trivial inequality $(\theta_1 + \theta_2 + \theta_3)^2 \ge 2\theta_1(\theta_2 + \theta_3)$, holding for all $\theta_1, \theta_2, \theta_3 \in \mathbb{R}$, with the choice $\theta_1 = L^{\max}R^0/2, \theta_2 = 2\hat{L}^{\max}R^0, \theta_3 = 2G^*$. We get

$$C \ge 2\sqrt{n-1}R^0(\hat{L}^{\max}R^0 + G^*) \ge 2\left[\frac{\hat{L}^{\max}}{2}(R^0)^2 + \sqrt{n-1}G^*R^0\right], \tag{7.12}$$

where the last inequality follows from the fact that we are assuming $n \ge 2$.

Now consider an outer iteration $k \geq 1$, picking any $x^* \in X^*$ and any $j \in \{1,\dots,n\}$. Using Lemma 5.2 with $v = \nabla f(x^*)$, $x' = x^k$ and $x'' = x^*$, we have

$$\begin{aligned}\nabla f(x^*)^T(x^k - x^*) &= \langle \nabla f(x^*) - \nabla_j f(x^*) e, x^k - x^* \rangle_j \\ &= \sum_{h \neq j} [\nabla_h f(x^*) - \nabla_j f(x^*)](x^k_h - x^*_h) \\ &\overset{(2.10)}{\leq} \sqrt{n-1}\, G^* \|x^k - x^*\|_{\langle j \rangle}.\end{aligned}$$

So, using Lemma 5.5 with $x' = x^*$ and $x'' = x^k$, we get

$$f(x^k) - f^* \leq \nabla f(x^*)^T(x^k - x^*) + \frac{L_j}{2}\|x^* - x^k\|^2_{\langle j \rangle} \leq \sqrt{n-1}\, G^* R^0 + \frac{\hat{L}^{\max}}{2}(R^0)^2.$$

In view of (7.12), we conclude that $f(x^k) - f^* \leq C/2$, implying that (7.10) holds. □

A question that can naturally arise is whether the constant C in Theorem 7.4 is tight. To answer this challenging question, we can look in detail at the steps of the above proofs, from which it seems that C may in fact be loose. For example, in the proof of Theorem 7.4 we got a lower bound for C by decomposing the last term in (7.11) as the sum of $\hat{L}^{\max}(R^0)^2 + 2\sqrt{n-1}\,G^* R^0$ and

$$(2\sqrt{n-1}-1)\hat{L}^{\max}(R^0)^2 + \sqrt{n-1}\left[\frac{L^{\max}(R^0)^2}{4} + 4\frac{(\hat{L}^{\max}R^0)^2 + (G^*)^2 + 2\hat{L}^{\max}G^* R^0}{L^{\max}}\right].$$

We then obtained (7.12) by lower bounding the above quantity by 0. But the above quantity may be much larger than 0 and, for large values of n and G^*, even dominant over $\hat{L}^{\max}(R^0)^2 + 2\sqrt{n-1}\,G^* R^0$, observing that $\hat{L}^{\max} = \xi L^{\max}$, with $\xi \in [1, n-1]$, as we see from (2.4), (2.5) and (2.6).

In the literature, a non-asymptotic convergence rate was also shown for other coordinate descent methods on different settings with one or more linear constraints, where the working set is chosen by random selection [37, 38, 40, 46, 48] or by rules based on first-order optimality violation [1, 28]. In particular, just like AC2CD, random coordinate descent do not use ∇f to choose the working set. A sublinear rate (in expectation) with respect to the objective values was shown for random coordinate descent in [40] under convexity of f, and a linear rate (in expectation) was shown in [48] under the additional assumption of proximal-PL inequality. We note that the sublinear rate $f(x^k) - f^* \leq n^2 L(\bar{R}^0)^2/[k + n^2 L(\bar{R}^0)^2/(f(x^0) - f^*)]$ obtained for random coordinate descent in [40], where $\bar{R}^0 = \max_x\{\max_{x^* \in X^*}\|x - x^*\| \colon x \in \mathcal{L}^0\}$, holds with respect to the inner iterations, so k should be multiplied by a factor $\mathcal{O}(n)$ to have a fair comparison with AC2CD, for which the rate was computed with respect to the outer iterations. With this adjustment, the rate of random coordinate descent is however better than $f(x^k) - f^* \leq C/k$ obtained for AC2CD, with the constant C from Theorem 7.4 being $\mathcal{O}(n(nL^{\max}R^0 + G^*)^2)$ if we reasonably assume $\sqrt{n} \gg 1/L^{\max}$ and consider $\hat{L}^{\max} = \mathcal{O}(nL^{\max})$ (since $\hat{L}^{\max} = \xi L^{\max}$, with $\xi \in [1, n-1]$, as observed above), where $L^{\max} \leq 2L$ from (2.2).

These results seem in agreement with the unconstrained case, where cyclic coordinate selection achieves worse convergence rate than random selection and Gauss-Southwell-type rules [2, 44], even if practical performances of the algorithms usually depend on the specific features of the problems.

7.3. Computation of the active-set complexity. Using all the previous results, we can now compute the active-set complexity of AC2CD, that is, the maximum number of iterations required by the algorithm to identify the active set. In particular, we give an upper bound for $\bar{k}$ appearing in Theorem 6.4 under convexity of f and a quadratic growth condition, which is now described.

We assume that there exists $\mu > 0$ such that

$$f(x) - f^* \geq \frac{\mu}{2}\|x - x^*\|^2, \quad \forall\, x \in \mathcal{L}^0, \tag{7.13}$$

where $x^* \in X^*$. Note that (7.13) is automatically satisfied if f is μ-strongly convex over $\mathcal{L}^0$ [41]. However, (7.13) is a weaker condition than strong convexity of f over $\mathcal{L}^0$, since there exist convex functions that satisfy (7.13) even if they are non-strongly convex. This can be seen in the following example, obtained from [39] with proper adjustments. Note that, in the provided example, f is not even strictly convex, there is a unique optimal solution x^* (so that $\{x^k\} \to x^*$) and Assumption 1 is satisfied.

Example 1. *Consider the following convex problem:*

$$\begin{aligned} & \min\ f(x) = \frac{1}{2}x_1^2 + \sum_{i=2}^{n} x_i \\ & e^T x = 0 \\ & x_1 \geq -1 \\ & x_i \geq 0, \quad i = 2, \dots, n, \end{aligned}$$

with arbitrary dimension $n \geq 3$. Since the smallest eigenvalue of the Hessian matrix of f is equal to 0, then f is not strongly convex. Actually, f is not even strictly convex, since $f(\omega x' + (1-\omega)x'') = \omega f(x') + (1-\omega) f(x'')$ for all $\omega \in [0,1]$ and any distinct feasible points x', x'' such that $x'_1 = x''_1$. We also have that $x^* = 0$ is the unique optimal solution and $f^* = 0$. We conclude that (7.13) is satisfied with $\mu = 1$, since $f(x) - f^* = \frac{1}{2}x_1^2 + \sum_{i=2}^n x_i \geq \frac{1}{2}\sum_{i=1}^n x_i^2 = \frac{1}{2}||x - x^*||^2$ for all feasible x.

Theorem 7.5. *The following upper bound holds for $\bar{k}$ appearing in Theorem 6.4 if f is convex over $\mathbb{R}^n$ and statisfies (7.13):*

$$\bar{k} \leq \left\lfloor \frac{2C}{\mu} \max\left\{ \left(\frac{\tau}{\tau+1} D^{max}(x^*)\right)^{-2}, \left(\frac{\zeta(x^*)}{2L + \max\left\{\dfrac{1}{A_l}, \dfrac{L^{max}}{2(1-\gamma)}\right\}}\right)^{-2} \right\} \right\rfloor + 1,$$

where $C \geq 0$ is the constant of the sublinear convergence rate defined in Theorem 7.4, $D^{max}(x^*) > 0$ is the maximum distance from the bounds at x^*, defined as in Proposition 6.2, $\zeta(x^*) > 0$ is the minimum strict complementarity measure at x^*, defined as in (6.5), L is the Lipschitz constant of ∇f, $A_l > 0$ is the lower bound on the parameter $A^{k,i}$ used to compute the initial stepsize $\Delta^{k,i}$ in the Armijo line search (see (3.6) and (3.8)), $L^{max} > 0$ is the maximum among the local Lipschitz constants $L_{i,j}$, defined as in (2.4), and $\tau \in (0,1]$ is the parameter used to choose $j(k)$, satisfying (3.3).

Proof. By the definition of $\bar{k}$ given in Theorem 6.4, it holds that $\bar{k} \geq \max\{k^j, k^z\}$ and (6.6) is satisfied. Recalling the definition of k^z given in Proposition 6.3 and the fact that (7.1) holds for all $k \geq 0$, we have $k^z = 0$, and then $\bar{k} \geq \max\{k^j, k^z\} = k^j$. So, from (6.6) and the definition of k^j given in Proposition 6.2, it follows that $\bar{k}$ is the first outer iteration such that

$$\|x^k - x^*\|_\infty < \frac{\tau}{\tau+1} D^{\max}(x^*), \qquad \forall\, k \ge \bar{k}. \tag{7.14a}$$

$$\|z^{k,i} - x^*\| < \frac{\zeta(x^*)}{2L + \max\left\{\dfrac{1}{A_l}, \dfrac{L^{\max}}{2(1-\gamma)}\right\}}, \quad i = 1,\dots,n, \qquad \forall\, k \ge \bar{k}. \tag{7.14b}$$

By Theorem 7.4 and (7.13), for all $k \ge 1$ we hence have that

$$\|x^k - x^*\|_\infty^2 \le \|x^k - x^*\|^2 \le \frac{2}{\mu}[f(x^k) - f^*] \le \frac{2C}{\mu k},$$
$$\|z^{k,i} - x^*\|^2 \le \frac{2}{\mu}[f(z^{k,i}) - f^*] \le \frac{2}{\mu}[f(x^k) - f^*] \le \frac{2C}{\mu k}, \quad i = 1,\dots,n,$$

where, in the last chain of inequalities, we used the fact that $f(z^{k,i+1}) \le f(z^{k,i}) \le f(x^k)$, $i = 1,\dots,n$. Therefore, (7.14) holds for all k such that $\sqrt{2C/(\mu k)}$ is less than both the right-hand side of (7.14a) and the right-hand side of (7.14b), yielding to the upper bound for $\bar{k}$ given in the assertion. □

We remark that Theorem 7.5 requires convexity and quadratic growth, but it uses the convergence rate result stated in Theorem 7.4, holding for general convex objective functions. As a consequence, we expect the upper bound provided for $\bar{k}$ in Theorem 7.5 to be loose. Improving the convergence rate of the algorithm under the additional quadratic growth condition may hence be a challenging question, since it affects the active-set complexity.

8. Additional results. So far we have shown that AC2CD identifies $\mathscr{X}^+(x^*)$ in a finite number $\bar{k}$ of outer iterations (provided $\{x^k\} \to x^*$), also giving an upper bound for $\bar{k}$ when f is convex and satisfies a quadratic growth condition.

Now, we want to show that the counterparts of these results hold as well, in the sense that AC2CD is able to identify the complement of $\mathscr{X}(x^*)$, the so called *non-active set*, in a finite number $\hat{k}$ of outer iterations, where an upper bound for $\hat{k}$ can be computed when f is convex and satisfies (7.13). More specifically, still considering a sequence $\{x^k\} \to x^*$, we want to show that, for all $k > \hat{k}$,

$$x_h^k \in (l_h, u_h), \quad \forall\, h \notin \mathscr{X}(x^*). \tag{8.1}$$

Actually, (8.1) is quite obvious (it follows from the properties of the limit), but obtaining an upper bound for $\hat{k}$ can be of interest. In particular, if (6.1) and (8.1) hold for $k > \bar{k}$ and $k > \hat{k}$, respectively, for all $k > \max\{\bar{k}, \hat{k}\}$ we have that

$$\mathscr{X}^+(x^*) \subseteq \big\{i \colon x_i^k \in \{l_i, u_i\}\big\} \subseteq \mathscr{X}(x^*).$$

As a consequence, if x^* is non-degenerate, for all $k > \max\{\bar{k}, \hat{k}\}$ it holds

$$x_h^k \in \{l_h, u_h\} \;\Leftrightarrow\; h \in \mathscr{X}(x^*), \tag{8.2}$$

that is, the active set is exactly identified after $\max\{\bar{k}, \hat{k}\}$ outer iterations.

First we show that (8.1) holds for all sufficiently k, without any assumption on the convexity of f, provided the whole sequence $\{x^k\}$ converges.

THEOREM 8.1. *Let $\{x^k\}$ be a sequence of points produced by AC2CD and assume that $\lim_{k\to\infty} x^k = x^*$. Define the minimum non-zero distance from the bounds at x^* as*

$$D^{min}(x^*) = \min_{i\notin \mathscr{Z}(x^*)} D_i(x^*),$$

which is well defined and positive by Assumption 1, and let $\hat{k}$ be the first outer iteration such that

$$\|x^k - x^*\|_\infty < D^{min}(x^*), \quad \forall\, k > \hat{k}.$$

Then, for all $k > \hat{k}$ we have that

$$x_h^k \in (l_h, u_h), \quad \forall\, h \notin \mathscr{Z}(x^*).$$

Proof. Consider an outer iteration $k > \hat{k}$ and any index $h \notin \mathscr{Z}(x^*)$. We have $|x_h^k - x_h^*| \le \|x^k - x^*\|_\infty < D^{\min}(x^*) \le D_h(x^*)$, implying that

$$x_h^k - l_h > x_h^* - l_h - D_h(x^*) \quad \text{and} \quad u_h - x_h^k > u_h - x_h^* - D_h(x^*). \tag{8.3}$$

Therefore, we can write

$$\begin{aligned} D_h(x^k) = \min\{x_h^k - l_h, u_h - x_h^k\} &\overset{(8.3)}{>} \min\{x_h^* - l_h, u_h - x_h^*\} - D_h(x^*) \\ &= D_h(x^*) - D_h(x^*) = 0, \end{aligned}$$

that is, $x_h^k \in (l_h, u_h)$. ∎

We finally give an upper bound for $\hat{k}$ under the same assumptions used in Theorem 7.5. As in the previous section, also here we assume the parameter A_i^k in the Armijo line search to be computed in order to satisfy, together with (3.8), condition (7.1) for all $k \ge 0$, as explained in Subsection 7.1.

THEOREM 8.2. *The following upper bound holds for $\hat{k}$ appearing in Theorem 8.1 if f is convex over $\mathbb{R}^n$ and statisfies (7.13):*

$$\hat{k} \le \left\lfloor \frac{2C}{\mu} \left(D^{min}(x^*)\right)^{-2} \right\rfloor + 1,$$

where $C \ge 0$ is the constant of the sublinear convergence rate defined in Theorem 7.4, and $D^{min}(x^*) > 0$ is the minimum non-zero distance from the bounds at x^*, defined as in Theorem 8.1.

Proof. Reasoning as in the proof of Theorem 7.5, the desired result follows from Theorem 8.1 and the fact that $\|x^k - x^*\|_\infty^2 \le \|x^k - x^*\|^2 \le \frac{2C}{\mu k}$ for all $k \ge 1$. ∎

The same remarks stated after Theorem 7.5 hold for Theorem 8.2 as well. Namely, we expect the upper bound provided for $\hat{k}$ to be loose, since it requires convexity and quadratic growth, but it uses the convergence rate result of Theorem 7.4, holding for general convex objective functions.

Appendix A. Proofs of the technical results of Section 5.

Proof of Lemma 5.2. For all $x \in \mathcal{F}$ we have $x_j = b - \sum_{i\neq j} x_i$, $j = 1, \ldots, n$. So,

$$\begin{aligned} v^T(x' - x'') &= \sum_{i\neq j} v_i(x_i' - x_i'') + v_j(x_j' - x_j'') \\ &= \sum_{i\neq j} v_i(x_i' - x_i'') - v_j\left(\sum_{i\neq j} x_i' - \sum_{i\neq j} x_i''\right) = \sum_{i\neq j}(v_i - v_j)(x_i' - x_j''). \end{aligned}$$ ∎

Proof of Lemma 5.3. Fix $j \in \{1,\dots,n\}$ and $x', x'' \in \mathcal{F}$. For all $i = 1,\dots,n$ and $x \in \mathcal{F}$, let $\phi_{i,j,x}$ be the functions appearing in (2.3). Pick any $h \neq j$ and, from known results on functions with Lipschitz continuous derivatives [41], we can write

$$\begin{aligned} f(x + t(e_h - e_j)) = \phi_{h,j,x}(t) &\le \phi_{h,j,x}(0) + t\dot{\phi}_{h,j,x}(0) + \frac{L_{h,j}}{2}t^2 \\ &= f(x) + t\nabla f(x)^T(e_h - e_j) + \frac{L_{h,j}}{2}t^2, \quad \forall\, t \in \mathbb{R}. \end{aligned}$$

Using $t = \dfrac{1}{L_{h,j}}(\nabla_j f(x) - \nabla_h f(x))$, we get

$$f(x) - f\Big(x + \frac{1}{L_{h,j}}(\nabla_j f(x) - \nabla_h f(x))(e_h - e_j)\Big) \ge \frac{1}{2L_{h,j}}(\nabla_h f(x) - \nabla_j f(x))^2. \tag{A.1}$$

Let $\bar f = \inf_{x\in\mathbb{R}^n} f(x)$. For all $x \in \mathbb{R}^n$ we can write

$$\begin{aligned} f(x) - \bar f \;&\ge\; f(x) - f\Big(x + \frac{1}{L_{h,j}}(\nabla_j f(x) - \nabla_h f(x))(e_h - e_j)\Big) \\ &\ge \frac{1}{2}\max_{i\neq j}\frac{1}{L_{i,j}}(\nabla_i f(x) - \nabla_j f(x))^2 \\ &\overset{(*)}{\ge} \frac{1}{2\sum_{i\neq j} L_{i,j}}\sum_{i=1}^{n}(\nabla_i f(x) - \nabla_j f(x))^2 \\ &= \frac{1}{2L_j}\sum_{i\neq j}(\nabla_i f(x) - \nabla_j f(x))^2 = \frac{1}{2L_j}\big\|\nabla f(x) - \nabla_j f(x)e\big\|^2_{\langle j\rangle}, \end{aligned} \tag{A.2}$$

where the second inequality follows (A.1), whereas the inequality $(*)$ follows from the fact that

$$\max_{i=1,\dots,r}\frac{a_i}{b_i} \ge \frac{1}{b_1 + \ldots + b_r}\sum_{i=1}^{n} a_i,$$

for all $a_1,\dots,a_r \in \mathbb{R}$ and $b_1,\dots,b_r > 0$.

Now, define the convex function $\psi_1(x) = f(x) - f(x') - \nabla f(x')^T(x - x')$. Since $\nabla\psi_1(x) = \nabla f(x) - \nabla f(x')$, for all $x \in \mathcal{F}$, $i \in \{1,\dots,n\}$ and $t, s \in \mathbb{R}$, we can write

$$\begin{aligned} &|\nabla\psi_1(x + t(e_i - e_j))^T(e_i - e_j) - \nabla\psi_1(x + s(e_i - e_j))^T(e_i - e_j)| \\ =&|\nabla f(x + t(e_i - e_j))^T(e_i - e_j) - \nabla f(x + s(e_i - e_j))^T(e_i - e_j)| \le L_{i,j}|t - s|, \end{aligned}$$

where the last inequality follows from the fact that $L_{i,j}$ are local Lipschitz constants for $\nabla f(x)$. Therefore, $L_{i,j}$ are also local Lipschitz constants for $\nabla\psi_1$. Consequently, we can use (A.2) with f replaced by ψ_1. Observing that $\min_{x\in\mathbb{R}^n}\psi_1(x) = 0$, we obtain

$$\begin{aligned} \psi_1(x) &\ge \frac{1}{2L_j}\big\|\nabla\psi_1(x) - \nabla_j\psi_1(x)e\big\|^2_{\langle j\rangle} \\ &= \frac{1}{2L_j}\big\|(\nabla f(x) - \nabla_j f(x)e) - (\nabla f(x') - \nabla_j f(x')e)\big\|^2_{\langle j\rangle}, \quad \forall\, x \in \mathbb{R}^n. \end{aligned}$$

Using $x = x''$ in the above relation, we get

$$\psi_1(x'') \ge \frac{1}{2L_j}\big\|(\nabla f(x'') - \nabla_j f(x'')e) - (\nabla f(x') - \nabla_j f(x')e)\big\|^2_{\langle j\rangle}. \tag{A.3}$$

Defining the function $\psi_2(x) = f(x) - f(x'') - \nabla f(x'')^T(x - x'')$, we can reason as above and we obtain

$$\psi_2(x') \geq \frac{1}{2L_j}\left\|(\nabla f(x') - \nabla_j f(x')e) - (\nabla f(x'') - \nabla_j f(x'')e)\right\|^2_{\langle j \rangle}. \tag{A.4}$$

Summing (A.3) and (A.4), we get

$$\left\|[\nabla f(x') - \nabla_j f(x')e] - [\nabla f(x'') - \nabla_j f(x'')e]\right\|^2_{\langle j \rangle} \leq L_j[\nabla f(x') - \nabla f(x'')]^T(x' - x'').$$

So, to obtain the desired result we have to show that $[\nabla f(x') - \nabla f(x'')]^T(x' - x'')$ is less than or equal to

$$\left\|[\nabla f(x') - \nabla_j f(x')e] - [\nabla f(x'') - \nabla_j f(x'')e]\right\|_{\langle j \rangle} \|x' - x''\|_{\langle j \rangle}. \tag{A.5}$$

This can be achieved by using Lemma 5.2 first with $v = \nabla f(x')$ and then with $v = \nabla f(x'')$, in order to rewrite $[\nabla f(x') - \nabla f(x'')]^T(x' - x'')$ as

$$\langle [\nabla f(x') - \nabla_j f(x')e] - [\nabla f(x'') - \nabla_j f(x'')e], x' - x'' \rangle_j.$$

Hence, by using inequality (2.8) we obtain that the above quantity is less than or equal to (A.5). ∎

Proof of Corollary 5.4. From (2.9) and the definition of $g^{k,i}$ given in (3.4), for all $v \in \mathbb{R}^n$ we have that

$$|\nabla_{p_i^k} f(v) - \nabla_{j(k)} f(v) + g^{k,i}| \leq \left\|[\nabla f(v) - \nabla_{j(k)} f(v)e] - [\nabla f(z^{k,i}) - \nabla_{j(k)} f(z^{k,i})e]\right\|_{\langle j(k) \rangle}.$$

Using Lemma 5.3, the desired result is obtained. ∎

Proof of Lemma 5.5. Fix $j \in \{1, \dots, n\}$ and $x', x'' \in \mathcal{F}$. From the mean value theorem and using Lemma 5.2 with $v = \nabla f(x' + t(x'' - x'))$, we have

$$\begin{aligned} f(x'') - f(x') &= \int_0^1 \nabla f(x' + t(x'' - x'))^T(x'' - x')\,dt \\ &= \int_0^1 \langle \nabla f(x' + t(x'' - x')) - \nabla_j f(x' + t(x'' - x'))e, x'' - x' \rangle_j\,dt. \end{aligned}$$

The integrand in the last term of the above chain of equalities can be rewritten as the sum of $\langle \nabla f(x') - \nabla_j f(x')e, x'' - x' \rangle_j$ and

$$\langle [\nabla f(x' + t(x'' - x')) - \nabla_j f(x' + t(x'' - x'))e] - [\nabla f(x') - \nabla_j f(x')e], x'' - x' \rangle_j,$$

and the latter, by using inequality (2.8) and Lemma 5.3, is less than or equal to $tL_j\|x' - x''\|^2_{\langle j \rangle}$. Therefore,

$$\begin{aligned} f(x'') &\leq f(x') + \langle \nabla f(x') - \nabla_j f(x')e, x'' - x' \rangle_j + L_j\|x' - x''\|^2_{\langle j \rangle} \int_0^1 t\,dt \\ &= f(x') + \langle \nabla f(x') - \nabla_j f(x')e, x'' - x' \rangle_j + \frac{L_j}{2}\|x' - x''\|^2_{\langle j \rangle}. \end{aligned}$$

Using Lemma 5.2 with $v = \nabla f(x')$, the desired result is obtained. ∎

Acknowledgments. The author would like to thank the two anonymous reviewers for their comments and suggestions.

REFERENCES

[1] A. Beck, *The 2-coordinate descent method for solving double-sided simplex constrained minimization problems*, J. Optim. Theory Appl., 162 (2014), pp. 892–919.

[2] A. Beck and L. Tetruashvili, *On the convergence of block coordinate descent type methods*, SIAM J. Optim., 23 (2013), pp. 2037–2060.

[3] D. P. Bertsekas, *On the Goldstein-Levitin-Polyak gradient projection method*, IEEE Trans. Automat. Control, 21 (1976), pp. 174–184.

[4] E. G. Birgin and J. M. Martínez, *Large-scale active-set box-constrained optimization method with spectral projected gradients*, Comput. Optim. Appl., 23 (2002), pp. 101–125.

[5] I. M. Bomze, F. Rinaldi, and S. R. Bulò, *First-order Methods for the Impatient: Support Identification in Finite Time with Convergent Frank–Wolfe Variants*, SIAM J. Optim., 29 (2019), pp. 2211–2226.

[6] I. M. Bomze, F. Rinaldi, and D. Zeffiro, *Active Set Complexity of the Away-Step Frank-Wolfe Algorithm*, SIAM J. Optim., 30 (2020), pp. 2470–2500.

[7] J. Burke, *On the identification of active constraints II: The nonconvex case*, SIAM J. Numer. Anal., 27 (1990), pp. 1081–1102.

[8] J. V. Burke and J. J. Moré, *On the identification of active constraints*, SIAM J. Numer. Anal., 25 (1988), pp. 1197–1211.

[9] J. V. Burke and J. J. Moré, *Exposing constraints*, SIAM J. Optim., 4 (1994), pp. 573–595.

[10] P. H. Calamai and J. J. Moré, *Projected gradient methods for linearly constrained problems*, Math. Program., 39 (1987), pp. 93–116.

[11] K. L. Clarkson, *Coresets, sparse greedy approximation, and the Frank-Wolfe algorithm*, ACM Trans. Algorithms, 6 (2010), pp. 1–30.

[12] A. Cristofari, *An almost cyclic 2-coordinate descent method for singly linearly constrained problems*, Comput. Optim. Appl., 73 (2019), pp. 411–452.

[13] A. Cristofari, M. De Santis, S. Lucidi, and F. Rinaldi, *A Two-Stage Active-Set Algorithm for Bound-Constrained Optimization*, J. Optim. Theory Appl., 172 (2017), pp. 369–401.

[14] A. Cristofari, M. De Santis, S. Lucidi, and F. Rinaldi, *An active-set algorithmic framework for non-convex optimization problems over the simplex*, Comput. Optim. Appl., 77 (2020), pp. 57–89.

[15] A. Cristofari, F. Rinaldi, and F. Tudisco, *Total variation based community detection using a nonlinear optimization approach*, SIAM J. Appl. Math., 80 (2020), pp. 1392–1419.

[16] A. Daniilidis, C. Sagastizábal, and M. Solodov, *Identifying structure of nonsmooth convex functions by the bundle technique*, SIAM J. Optim., 20 (2009), pp. 820–840.

[17] M. De Santis, S. Lucidi, and F. Rinaldi, *A Fast Active Set Block Coordinate Descent Algorithm for ℓ_1-Regularized Least Squares*, SIAM J. Optim., 26 (2016), pp. 781–809.

[18] J. C. Duchi, F. Ruan, et al., *Asymptotic optimality in stochastic optimization*, Ann. Statist., 49 (2021), pp. 21–48.

[19] J. C. Dunn, *On the convergence of projected gradient processes to singular critical points*, J. Optim. Theory Appl., 55 (1987), pp. 203–216.

[20] F. Facchinei, J. Júdice, and J. Soares, *An active set Newton algorithm for large-scale nonlinear programs with box constraints*, SIAM J. Optim., 8 (1998), pp. 158–186.

[21] F. Facchinei and J.-S. Pang, *Finite-dimensional variational inequalities and complementarity problems*, Springer Science & Business Media, 2003.

[22] E. M. Gafni and D. P. Bertsekas, *Two-metric projection methods for constrained optimization*, SIAM J. Control Optim., 22 (1984), pp. 936–964.

[23] W. W. Hager and H. Zhang, *A new active set algorithm for box constrained optimization*, SIAM J. Optim., 17 (2006), pp. 526–557.

[24] W. Hare, *Identifying active manifolds in regularization problems*, in Fixed-Point Algorithms for Inverse Problems in Science and Engineering, Springer, 2011, pp. 261–271.

[25] W. L. Hare, *A proximal method for identifying active manifolds*, Comput. Optim. Appl., 43 (2009), pp. 295–306.

[26] W. L. Hare and A. S. Lewis, *Identifying active constraints via partial smoothness and prox-regularity*, J. Convex Anal., 11 (2004), pp. 251–266.

[27] C.-J. Hsieh, K.-W. Chang, C.-J. Lin, S. S. Keerthi, and S. Sundararajan, *A dual coordinate descent method for large-scale linear SVM*, in Proceedings of the 25th international conference on Machine learning, 2008, pp. 408–415.

[28] M. Jaggi and S. Lacoste-Julien, *On the global linear convergence of frank-wolfe optimization variants*, Advances in Neural Information Processing Systems, 28 (2015).

[29] S. Lee and S. J. Wright, *Manifold identification in dual averaging for regularized stochastic online learning*, J. Mach. Learn. Res., 13 (2012), pp. 1705–1744.

[30] A. S. Lewis and S. J. Wright, *Identifying activity*, SIAM J. Optim., 21 (2011), pp. 597–614.
[31] R. M. Lewis and V. Torczon, *Active set identification for linearly constrained minimization without explicit derivatives*, SIAM J. Optim., 20 (2010), pp. 1378–1405.
[32] J. Liang, J. Fadili, and G. Peyré, *Activity Identification and Local Linear Convergence of Forward–Backward-type Methods*, SIAM J. Optim., 27 (2017), pp. 408–437.
[33] C.-J. Lin, S. Lucidi, L. Palagi, A. Risi, and M. Sciandrone, *Decomposition algorithm model for singly linearly-constrained problems subject to lower and upper bounds*, J. Optim. Theory Appl., 141 (2009), pp. 107–126.
[34] Z.-Q. Luo and P. Tseng, *On the convergence of the coordinate descent method for convex differentiable minimization*, J. Optim. Theory Appl., 72 (1992), pp. 7–35.
[35] Z.-Q. Luo and P. Tseng, *On the convergence rate of dual ascent methods for linearly constrained convex minimization*, Math. Oper. Res., 18 (1993), pp. 846–867.
[36] R. Mifflin and C. Sagastizábal, *Proximal points are on the fast track*, J. Convex Anal., 9 (2002), pp. 563–580.
[37] I. Necoara, *Random coordinate descent algorithms for multi-agent convex optimization over networks*, IEEE Trans. Automat. Control, 58 (2013), pp. 2001–2012.
[38] I. Necoara, Y. Nesterov, and F. Glineur, *Random block coordinate descent methods for linearly constrained optimization over networks*, Journal of Optimization Theory and Applications, 173 (2017), pp. 227–254.
[39] I. Necoara, Y. Nesterov, and F. Glineur, *Linear convergence of first order methods for non-strongly convex optimization*, Math. Program., 175 (2019), pp. 69–107.
[40] I. Necoara and A. Patrascu, *A random coordinate descent algorithm for optimization problems with composite objective function and linear coupled constraints*, Computational Optimization and Applications, 57 (2014), pp. 307–337.
[41] Y. Nesterov, *Introductory lectures on convex optimization: A basic course*, vol. 87, Springer Science & Business Media, 2013.
[42] J. Nutini, I. Laradji, and M. Schmidt, *Let's Make Block Coordinate Descent Go Fast: Faster Greedy Rules, Message-Passing, Active-Set Complexity, and Superlinear Convergence*, preprint, https://arxiv.org/abs/1712.08859 (2017).
[43] J. Nutini, M. Schmidt, and W. Hare, *"Active-set complexity" of proximal gradient: How long does it take to find the sparsity pattern?*, Optim. Lett., 13 (2019), pp. 645–655.
[44] J. Nutini, M. Schmidt, I. Laradji, M. Friedlander, and H. Koepke, *Coordinate descent converges faster with the gauss-southwell rule than random selection*, in International Conference on Machine Learning, PMLR, 2015, pp. 1632–1641.
[45] J. M. Ortega and W. C. Rheinboldt, *Iterative solution of nonlinear equations in several variables*, vol. 30, Siam, 1970.
[46] A. Patrascu and I. Necoara, *Efficient random coordinate descent algorithms for large-scale structured nonconvex optimization*, J. Global Optim., 61 (2015), pp. 19–46.
[47] C. Poon, J. Liang, and C. Schoenlieb, *Local convergence properties of SAGA/Prox-SVRG and acceleration*, in Proc. Mach. Learn. Res. (PMLR), 2018, pp. 4124–4132.
[48] J. She and M. Schmidt, *Linear convergence and support vector identification of sequential minimal optimization*, in 10th NIPS Workshop on Optimization for Machine Learning, vol. 5, 2017.
[49] Y. Sun, H. Jeong, J. Nutini, and M. Schmidt, *Are we there yet? manifold identification of gradient-related proximal methods*, in The 22nd International Conference on Artificial Intelligence and Statistics, 2019, pp. 1110–1119.
[50] S. J. Wright, *Identifiable surfaces in constrained optimization*, SIAM J. Control Optim., 31 (1993), pp. 1063–1079.
[51] S. J. Wright, *Accelerated block-coordinate relaxation for regularized optimization*, SIAM J. Optim., 22 (2012), pp. 159–186.