A FAST ACTIVE SET BLOCK COORDINATE DESCENT ALGORITHM FOR ℓ_1-REGULARIZED LEAST SQUARES

MARIANNA DE SANTIS*, STEFANO LUCIDI†, AND FRANCESCO RINALDI‡

Abstract. The problem of finding sparse solutions to underdetermined systems of linear equations arises in several applications (e.g. signal and image processing, compressive sensing, statistical inference). A standard tool for dealing with sparse recovery is the ℓ_1-regularized least-squares approach that has been recently attracting the attention of many researchers.

In this paper, we describe an active set estimate (i.e. an estimate of the indices of the zero variables in the optimal solution) for the considered problem that tries to quickly identify as many active variables as possible at a given point, while guaranteeing that some approximate optimality conditions are satisfied. A relevant feature of the estimate is that it gives a significant reduction of the objective function when setting to zero all those variables estimated active. This enables to easily embed it into a given globally converging algorithmic framework.

In particular, we include our estimate into a block coordinate descent algorithm for ℓ_1-regularized least squares, analyze the convergence properties of this new active set method, and prove that its basic version converges with linear rate.

Finally, we report some numerical results showing the effectiveness of the approach.

Key words. ℓ_1-regularized least squares, active set, sparse optimization

AMS subject classifications. 65K05, 90C25, 90C06

1. Introduction. The problem of finding sparse solutions to large underdetermined linear systems of equations has received a lot of attention in the last decades. This is due to the fact that several real-world applications can be formulated as linear inverse problems. A standard approach is the so called ℓ_2-ℓ_1 unconstrained optimization problem:

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}\|Ax-b\|^2+\tau\|x\|_1, \tag{1.1}$$

where $A\in\mathbb{R}^{m\times n}$, $b\in\mathbb{R}^m$, $x\in\mathbb{R}^n$ $(m<n)$ and $\tau\in\mathbb{R}^+$. We denote by $\|\cdot\|$ the standard ℓ_2 norm and by $\|\cdot\|_1$ the ℓ_1 norm defined as $\|x\|_1=\sum_{i=1}^n|x_i|$.

Several classes of algorithms have been proposed for the solution of Problem (1.1). Among the others, we would like to remind Iterative Shrinkage/Thresholding (IST) methods (see e.g. [3, 4, 9, 11, 34]), Augmented Lagrangian Approaches (see e.g. [2]), Second Order Methods (see e.g. [5, 18]), Sequential Deterministic (see e.g. [32, 33, 39]) and Stochastic (see e.g. [16, 28] and references therein) Block Coordinate Approaches, Parallel Deterministic (see e.g. [15] and references therein) and Stochastic (see e.g. [10, 29] and references therein) Block Coordinate Approaches, and Active-set strategies (see e.g. [20, 35, 36]).

The main feature of this class of problems is the fact that the optimal solution is usually very sparse (i.e. it has many zero components). Then, quickly building and/or correctly identifying the active set (i.e. the subset of zero components in an optimal solution) for Problem (1.1) is becoming a crucial task in the context of Big

*Fakultät für Mathematik, Technische Universität Dortmund, Vogelpothsweg 87, 44227 Dortmund, Germany marianna.de.santis@tu-dortmund.de

†Dipartimento di Ingegneria Informatica Automatica e Gestionale, Sapienza Università di Roma, Via Ariosto, 25, 00185 Roma, Italy stefano.lucidi@dis.uniroma1.it

‡Dipartimento di Matematica, Università di Padova, Via Trieste, 63, 35121 Padova, Italy rinaldi@math.unipd.it

Data Optimization, since it can guarantee relevant savings in terms of CPU time. As a very straightforward example, we can consider a huge scale problem having a solution with just a few nonzero components. In this case, both the fast construction and the correct identification of the active set can considerably reduce the complexity of the problem, thus also giving us the chance to use more sophisticated optimization methods than the ones usually adopted.Various attempts have been made in order to use active set technique in the context of ℓ_1-regularized problems.

In [35, 36], Wen et al. proposed a two-stage algorithm, FPC-AS, where an estimate of the active variables set is driven by using a first-order iterative shrinkage method.

In [37], a block-coordinate relaxation approach with proximal linearized subproblems yields convergence to critical points, while identification of the optimal manifold (under a nondegeneracy condition) allows acceleration techniques to be applied on a reduced space.

In [23], the authors solve an ℓ_1-regularized log determinant program related to the problem of sparse inverse covariance matrix estimation combining a second-order approach with a technique to correctly identifying the active set.

An efficient version of the two-block nonlinear constrained Gauss-Seidel algorithm that at each iteration fixes some variables to zero according to a simple active set rule has been proposed in [27] for solving ℓ_1-regularized least squares.

In a recent paper [5], Nocedal et al. described an interesting family of second order methods for ℓ_1-regularized convex problems. Those methods combine a semismooth Newton approach with a mechanism to identify the active manifold in the given problem.

In the case one wants to solve very large problems, Block Coordinate Descent Algorithms (both Sequential and Parallel) represent a very good alternative and, sometimes, the best possible answer [33]. An interesting Coordinate Descent algorithm combining a Newton steps with a line search technique was described by Yuan et al. in [38]. In this context, the authors also proposed a shrinking technique (i.e. a heuristic strategy that tries to fix to zero a subset of variables according to a certain rule), which can be seen as a way to identify the active variables. In [33], some ideas on how to speed up their Block Coordinate Descent Algorithm by including an active set identification strategy are described, but no theoretical analysis is given for the resulting approach.

What we want to highlight here is that all the approaches listed above, but the one described in [5], estimate the final active set by using the current active set and perform subspace minimization on the remaining variables. In [5], the authors define an estimate that performs multiple changes in the active manifold by also including variables that are nonzero at a given point and satisfy some specific condition. Since this active set mechanism, due to the aggressive changes in the index set, can cause cycling, including the estimate into a globally converging algorithmic framework is not always straightforward.

In this work, we adapt the active set estimate proposed in [14] for constrained optimization problems to the ℓ_1-regularized least squares case. Our estimate, similarly to the one proposed in [5], does not only focus on the zero variables of a given point. Instead it tries to quickly identify as many active variables as possible (including the nonzero variables of the point), while guaranteeing that some approximate optimality conditions are satisfied.

The main feature of the proposed active set strategy is that a significant reduction

of the objective function is obtained when setting to zero all those variables estimated active. This global property, which is strongly related to the fact that the components estimated active satisfy an approximate optimality condition, makes easy to use the estimate into a given globally converging algorithmic framework.

Furthermore, inspired by the papers [33, 38, 39], we describe a new Block Coordinate Descent Algorithm that embeds the considered active set estimate. At each iteration, the method first sets to zero the active variables, then uses a decomposition strategy for updating a bunch of the non-active ones. On the one hand, decomposing the non-active variables enables to handle huge scale problems that other active set approaches cannot solve in reasonable time. On the other hand, since the subproblems analyzed at every iteration explicitly take into account the ℓ_1-norm, the proposed algorithmic framework does not require a sign identification strategy (for the non-active variables), which is tipically needed when using other active set methods from the literature.

The paper is organized as follows. In Section 3, we introduce our active set strategy. In Section 4, we describe the active set coordinate descent algorithm, and prove its convergence. We further analyze the convergence rate of the algorithm. In Section 5, we report some numerical results showing the effectiveness of the approach. Finally, we draw some conclusions in Section 6.

2. Notation and Preliminary Results. Throughout the paper we denote by $f(x)$, $q(x)$, $g(x)$ and H the original function in Problem (1.1), the quadratic term of the objective function in Problem (1.1), the n gradient vector and the $n \times n$ Hessian matrix of $\frac{1}{2}\|Ax-b\|^2$ respectively. Explicitly

$$q(x) = \frac{1}{2}\|Ax-b\|^2, \quad g(x) = A^\top(Ax-b), \quad H = A^\top A.$$

Given a matrix $Q \in \mathbb{R}^{n\times n}$, we further denote by $\lambda_{max}(Q)$ and $\lambda_{min}(Q)$ the maximum and the minimum eigenvalue of the matrix Q, respectively. Furthermore, with I we indicate the set of indices $I = \{1, \ldots, n\}$, and with $Q_{I_j I_j}$ we indicate the submatrix of Q whose rows and columns indices are in $I_j \subseteq I$. We also report the optimality conditions for Problem (1.1):

PROPOSITION 2.1. *$x^\star \in \mathbb{R}^n$ is an optimal solution of Problem* (1.1) *if and only if*

$$\begin{cases} x_i^\star > 0, & g_i(x^\star) + \tau = 0 \\ x_i^\star < 0, & g_i(x^\star) - \tau = 0 \\ x_i^\star = 0, & -\tau \le g_i(x^\star) \le \tau. \end{cases} \tag{2.1}$$

Furthermore, we define a continuous function $\Phi_i(x)$ that measures the violation of the optimality conditions in x_i (and is connected to the Gauss-Southwell-r rule proposed in [33]), that is

$$\Phi_i(x) = -\text{mid}\left\{\frac{g_i(x)-\tau}{H_{ii}}, x_i, \frac{g_i(x)+\tau}{H_{ii}}\right\}, \tag{2.2}$$

where $\text{mid}\{a,b,c\}$ indicates the median of $a,\ b,\ c$.

Finally, we recall the concept of strict complementarity.

DEFINITION 2.2. *Strict complementarity holds if, for any $x_i^\star = 0$, we have*

$$-\tau < g_i(x^\star) < \tau. \tag{2.3}$$

3. Active set estimate. All the algorithms that adopt active set strategies need to estimate a particular subset of components of the optimal solution $x^\star$. In nonlinear constrained minimization problems, for example, using an active set strategy usually means correctly identifying the set of active constraints at the solution. In our context, we deal with Problem (1.1) and the active set is considered as the subset of zero-components of $x^\star$.

DEFINITION 3.1. *Let $x^\star \in \mathbb{R}^n$ be an optimal solution for Problem (1.1). We define the active set as follows:*

$$\bar{\mathcal{A}}(x^\star) = \big\{ i \in I : x_i^\star = 0 \big\}. \tag{3.1}$$

We further define as non-active set the complementary set of $\bar{\mathcal{A}}(x^\star)$:

$$\bar{\mathcal{N}}(x^\star) = I \setminus \bar{\mathcal{A}}(x^\star) = \big\{ i \in \{1, \ldots, n\} : x_i^\star \neq 0 \big\}. \tag{3.2}$$

In order to get an estimate of the active set we rewrite Problem (1.1) as a box constrained programming problem and we use similar ideas to those proposed in [12].

Problem (1.1) can be equivalently rewritten as follows:

$$\begin{array}{ll} \min & \frac{1}{2}\|A(u-v)-b\|^2 + \tau \sum_{i=1}^n (u_i + v_i) \\ & u \geq 0 \\ & v \geq 0, \end{array} \tag{3.3}$$

where $u, v \in \mathbb{R}^n$. Indeed, we can transform a solution $x^\star \in \mathbb{R}^n$ of Problem (1.1) into a solution $(u^\star, v^\star) \in \mathbb{R}^n \times \mathbb{R}^n$ of (3.3) by using the following transformation:

$$u^\star = \max(0, x^\star), \qquad v^\star = \max(0, -x^\star).$$

Equivalently, we can transform a solution $(u^\star, v^\star) \in \mathbb{R}^n \times \mathbb{R}^n$ of (3.3) into a solution $x^\star \in \mathbb{R}^n$ of Problem (1.1) by using the following transformation:

$$x^\star = u^\star - v^\star.$$

The Lagrangian function associated to (3.3) is

$$\mathcal{L}(u, v, \lambda, \mu) = \frac{1}{2}\|A(u-v)-b\|^2 + \tau \sum_{i=1}^n (u_i + v_i) - \lambda^\top u - \mu^\top v,$$

with $\lambda, \mu \in \mathbb{R}^n$ vectors of Lagrangian multipliers. Let $(u^\star, v^\star, \lambda^\star, \mu^\star)$ be an optimal solution of Problem (3.3). Then, from necessary optimality conditions, we have

$$\begin{array}{l} \lambda_i^\star = g_i(u^\star - v^\star) + \tau \; = g_i(x^\star) + \tau \, ; \\ \mu_i^\star = \tau \; - g_i(u^\star - v^\star) = \tau \; - g_i(x^\star). \end{array} \tag{3.4}$$

From (3.4), we can introduce the following two multiplier functions

$$\begin{array}{l} \lambda_i(u, v) = g_i(u - v) + \tau \, ; \\ \mu_i(u, v) = \tau \; - g_i(u - v). \end{array} \tag{3.5}$$

By means of the multiplier functions, we can recall the non-active set estimate $\mathcal{N}(u,v)$ and active set estimate $\mathcal{A}(u,v)$ proposed in the field of constrained smooth optimization (see [14] and references therein):

$$\mathcal{N}(u,v) = \{i : u_i > \epsilon\, \lambda_i(u,v)\} \cup \{i : v_i > \epsilon\, \mu_i(u,v)\}, \tag{3.6}$$

$$\mathcal{A}(u,v) = I \setminus \mathcal{N}(u,v), \tag{3.7}$$

where ϵ is a positive scalar.

We draw inspiration from (3.6) and (3.7) to propose the new estimates of active and non-active set for Problem (1.1). Indeed, by using the relations

$$u = \max(0,x) \qquad \text{and} \qquad v = \max(0,-x),$$

we can give the following definitions.

DEFINITION 3.2. *Let $x \in \mathbb{R}^n$. We define the following sets as estimate of the non-active and active variables sets:*

$$\mathcal{N}(x) = \{i : \max(0,x_i) > \epsilon\,(\tau + g_i(x))\} \cup \{i : \max(0,-x_i) > \epsilon\,(\tau - g_i(x))\}, \tag{3.8}$$

$$\mathcal{A}(x) = I \setminus \mathcal{N}(x). \tag{3.9}$$

In the next Subsections, we first discuss local and global properties of our estimate, then we compare it with other active set estimates.

3.1. Local properties of the active set estimate. Now, we describe some local properties (in the sense that those properties only hold into a neighborhood of a given point) of our active set estimate. In particular, the following theoretical result states that when the point is sufficiently close to an optimal solution the related active set estimate is a subset of the active set calculated in the optimal point (and it includes the optimal active variables that satisfy strict complementarity). Furthermore, when strict complementarity holds the active set estimate is actually equal to the optimal active set.

THEOREM 3.3. *Let $x^\star \in \mathbb{R}^n$ be an optimal solution of Problem* (1.1)*. Then, there exists a neighborhood of $x^\star$ such that, for each x in this neighborhood, we have*

$$\bar{\mathcal{A}}^+(x^\star) \subseteq \mathcal{A}(x) \subseteq \bar{\mathcal{A}}(x^\star), \tag{3.10}$$

with $\bar{\mathcal{A}}^+(x^\star) = \bar{\mathcal{A}}(x^\star) \cap \{i : -\tau < g_i(x^\star) < \tau\}$.
Furthermore, if strict complementarity (2.2) *holds in $x^\star$, then there exists a neighborhood of $x^\star$ such that, for each x in this neighborhood, we have*

$$\mathcal{A}(x) = \bar{\mathcal{A}}(x^\star). \tag{3.11}$$

Proof. The proof follows from Theorem 2.1 in [14]. ∎

3.2. A global property of the active set estimate. Here, we analyze a global property of the active set estimate. In particular, we show that, for a suitably chosen value of the parameter ϵ appearing in Definition 3.2, by starting from a point $z \in \mathbb{R}^n$ and fixing to zero all variables whose indices belong to the active set estimate $\mathcal{A}(z)$, it is possible to obtain a significant decrease of the objective function. This property, which strongly depends on the specific structure of the problem under analysis, represents a new interesting theoretical result, since it enables to easily embed the active set estimate into any globally converging algorithmic framework (in the next section, we will show how to include it into a specific Block Coordinate Descent method). Furthermore, the global property cannot be deduced from the theoretical results already reported in [14].

ASSUMPTION 1. *Parameter ϵ appearing in Definition 3.2 satisfies the following condition:*

$$0 < \epsilon < \frac{1}{\lambda_{max}(A^\top A)}. \tag{3.12}$$

PROPOSITION 3.4. *Let Assumption 1 hold. Given a point $z \in \mathbb{R}^n$ and the related sets $\mathcal{A}(z)$ and $\mathcal{N}(z)$, let y be the point defined as*

$$y_{\mathcal{A}(z)} = 0, \qquad y_{\mathcal{N}(z)} = z_{\mathcal{N}(z)}.$$

Then,

$$f(y) - f(z) \leq -\frac{1}{2\epsilon}\|y - z\|^2.$$

Proof. see Appendix A. ∎

3.3. Comparison with other active set strategies. Our active set estimate is somehow related to those proposed respectively by Byrd et al. in [5] and by Yuan et al. in [38]. It is also connected in some way to the IST Algorithm (ISTA), see e.g. [3, 11]. Indeed, an ISTA step can be seen as a simple way to set to zero the variables in the context of ℓ_1-regularized least-squares problems.

Here, we would like to point out the similarities and the differences between those strategies and the one we propose in the present paper.

First of all, we notice that, at a generic iteration k of a given algorithm, if x^k is the related iterate and $i \in I$ is an index estimated active by our estimate, that is,

$$i \in \mathcal{A}(x^k) = \{i : \max(0, x_i^k) \leq \epsilon\,(\tau + g_i(x^k))\} \cap \{i : \max(0, -x_i^k) \leq \epsilon\,(\tau - g_i(x^k))\},$$

this is equivalent to write

$$x_i^k \in [\epsilon(g_i(x^k) - \tau), \epsilon(g_i(x^k) + \tau)] \quad \text{and} \quad -\tau \leq g_i(x^k) \leq \tau, \tag{3.13}$$

which means that x_i^k is sufficiently small and satisfies the optimality condition associated with a zero component (see (2.1)). As we will see, the estimate, due to the way it is defined, tends to be more conservative than other active set strategies (i.e. it might set to zero slightly smaller sets of variables). On the other hand, the global property analyzed in the previous section (i.e. decrease of the objective function when

setting to zero the active variables) seems to indicate that the estimate truly contains indices related to variables that will be active in the optimal solution. As we will see later on, this important property does not hold when considering the other active set strategies analyzed here.

In the block active set algorithm for quadratic ℓ_1-regularized problems proposed in [5], the active set estimate, at a generic iteration k, can be rewritten in the following way:

$$\mathcal{A}^k_{Byrd} = \{i : x^k_i = 0;\, g_i(x^k) \in (-\tau, \tau)\} \cup \{i : x^k_i < 0;\, g_i(x^k) = -\tau\} \cup \{i : x^k_i > 0;\, g_i(x^k) = \tau\}.$$

Let $x^k \in \mathbb{R}^n$ and $i \in \{1, \ldots, n\}$ be an index estimated active by our estimate, from (3.13), we get $g_i(x^k) \in [-\tau, \tau]$.

Then, in the case $x^k_i = 0$, $i \in \mathcal{A}^k_{Byrd}$ implies $i \in \mathcal{A}(x^k)$. In fact, let $i \in \mathcal{A}^k_{Byrd}$. If $x^k_i = 0$ we have $g_i(x^k) \in (-\tau, \tau)$ so that $i \in \mathcal{A}(x^k)$. It is easy to see that the other way around is not true.

Other differences between the two estimates come out when considering indices i such that $x^k_i \neq 0$. Let $i \in \mathcal{A}^k_{Byrd}$ and, in particular, $i \in \{i : x^k_i < 0;\, g_i(x^k) = -\tau\}$. If $|x^k_i| > \epsilon\, 2\tau$, then we get

$$max(0, -x^k_i) = -x^k_i > \epsilon\, 2\tau = \epsilon\, (\tau - g_i(x^k)),$$

so that $i \notin \mathcal{A}(x^k)$. Using the same reasoning we can see that, in the case $i \in \mathcal{A}^k_{Byrd}$ and, in particular, $i \in \{i : x^k_i > 0;\, g_i(x^k) = \tau\}$, it can happen

$$max(0, x^k_i) = x^k_i > \epsilon\, 2\tau = \epsilon\, (\tau + g_i(x^k)),$$

so that $i \notin \mathcal{A}(x^k)$.

In [38], the active set estimate is defined as follows

$$\mathcal{A}^k_{Yuan} = \left\{i : x^k_i = 0;\, g_i(x^k) \in (-\tau + M^{k-1}, \tau - M^{k-1})\right\}, \tag{3.14}$$

where M^{k-1} is a positive scalar that measures the violation of the optimality conditions. It is easy to see that our active set contains the one proposed in [38]. Furthermore, we have that variables contained in our estimate are not necessarily contained in the estimate (3.14). In particular, a big difference between our estimate and the one proposed in [38] is that we can also include variables that are non-zero at the current iterate.

As a final comparison, we would like to point out the differences between the ISTA strategy and our estimate. Consider the generic iteration of ISTA with the same ϵ used in our active set strategy:

$$x^{k+1} = \arg\min_x \left\{ q(x^k) + g(x^k)^\top (x - x^k) + \epsilon \|x - x^k\|^2 + \tau \|x\|_1 \right\}. \tag{3.15}$$

From the optimality conditions of the inner problem in (3.15), we have that the zero variables at x^{k+1} belong to the following set:

$$\mathcal{A}^k_{ISTA} = \{i : \epsilon(-\tau + g_i(x^k)) \leq x^k_i \leq \epsilon(\tau + g_i(x^k))\}. \tag{3.16}$$

We can easily see that $\mathcal{A}(x^k) \subseteq \mathcal{A}^k_{ISTA}$. The opposite is not always true, apart from the variables $x^k_i = 0$. As a matter of fact, let us consider $x^k_i > 0$ and $i \in \mathcal{A}^k_{ISTA}$. Then, we have that

$$x^k_i \leq \epsilon(\tau + g_i(x^k)) \quad \Rightarrow \quad i \in \{i : \ \max(0, x^k_i) \leq \epsilon(\tau + g_i(x^k))\}$$

$$x^k_i \geq \epsilon(-\tau + g_i(x^k)) \quad \Rightarrow \quad -x^k_i \leq \epsilon(\tau - g_i(x^k))$$

In order to have $i \in \mathcal{A}(x^k)$ it should be

$$\epsilon(\tau - g_i) \geq \max\{0, -x^k_i\} = 0$$

that is a tighter requirement with respect to the one within $\mathcal{A}^k_{ISTA}$. A similar reasoning applies also to variables $x^k_i < 0$ with $i \in \mathcal{A}^k_{ISTA}$. We would also like to notice that the ISTA step might generate unnecessary projections of variables to zero, thus being not always effective as a tool for identifying the active set.

In this final remark, we show that, when using the active set strategies analyzed above, a sufficient decrease of the objective function cannot be guaranteed by setting to zero the variables in the active set (i.e. Proposition 3.4 does not hold). This fact makes hard, in some cases, to include those active set strategies into a globally convergent algorithmic framework.

REMARK 1. *Proposition 3.4 does not hold for the active set strategies described above. This can be easily seen in the following case.*

Let us assume that, at some iteration k, it exists only one index $\hat{\imath} \in \mathcal{A}^k_{Byrd}$, with $x^k_{\hat{\imath}} > 0$, $H_{\hat{\imath}\hat{\imath}} > 0$ and $g_{\hat{\imath}}(x^k) = \tau$. Let $z = x^k$ and y be the point defined as $y_i = x^k_i$ for all $i \neq \hat{\imath}$, and $y_{\hat{\imath}} = 0$. Then,

$$f(y) = f(x^k) + (g_{\hat{\imath}}(x^k) - \tau)(y_{\hat{\imath}} - x^k_{\hat{\imath}}) + \frac{1}{2}(y_{\hat{\imath}} - x^k_{\hat{\imath}})^2 H_{\hat{\imath}\hat{\imath}}.$$

Since $H_{\hat{\imath}\hat{\imath}} > 0$ and $g_{\hat{\imath}}(x^k) = \tau$, we have $f(y) - f(x^k) > 0$, so that by setting to zero the active variable we get an increase of the objective function value.

The same reasoning applies also to the ISTA step, assuming that at some iteration k, there exists only one index $\hat{\imath}$ such that

$$\epsilon(-\tau + g_{\hat{\imath}}(x^k)) < x^k_{\hat{\imath}} < \epsilon(\tau + g_{\hat{\imath}}(x^k))$$

and $g_{\hat{\imath}}(x^k) = \tau$.

Finally, it is easy to notice that, at each iteration k, the active set estimate $\mathcal{A}^k_{Yuan}$ defined in [38] only keeps fixed to zero, at iteration k, some of the variables that are already zero in x^k, thus not changing the objective function value.

4. A Fast Active Set Block Coordinate Descent Algorithm. In this section, we describe our Fast Active SeT Block Coordinate Descent Algorithm (`FAST-BCDA`) and analyze its theoretical properties. The main idea behind the algorithm is that of exploiting as much as possible the good properties of our active set estimate, more specifically:

- the ability to identify, for k sufficiently large, the "strong" active variables (namely, those variables satisfying the strict complementarity, see Theorem 3.3);
- the ability to obtain, at each iteration, a sufficient decrease of the objective function, by fixing to zero those variables belonging to the active set estimate (see Proposition 3.4 of the previous section).

As we have seen in the previous section, the estimate, due to the way it is defined, tends to be more conservative than other active set strategies (i.e. it might set to zero a slightly smaller set of variables at each iteration). Anyway, since for each block we exactly solve an ℓ_1-regularized subproblem, we can eventually force to zero some other variables in the non-active set. Another important consequence of including the ℓ_1-norm in the subproblems is that we do not need any sign identification strategy for the non-active variables.

At each iteration k, the algorithm defines two sets $\mathcal{N}^k = \mathcal{N}(x^k)$, $\mathcal{A}^k = \mathcal{A}(x^k)$ and executes two steps:

1) it sets to zero all of the active variables;
2) it minimizes only over a subset of the non-active variables, i.e. those which violate the optimality conditions the most.

More specifically, we consider the measure related to the violation of the optimality conditions reported in (2.2). We then sort in decreasing order the indices of non-active variables (i.e. the set of indices $\mathcal{N}^k$) with respect to this measure and define the subset $\bar{\mathcal{N}}^k_{ord} \subseteq \mathcal{N}^k$ containing the first s sorted indices.

The set $\bar{\mathcal{N}}^k_{ord}$ is then partitioned into q subsets $I_1, \ldots, I_q$ of cardinality r, such that $s = qr$. Then the algorithm performs q subiterations. At the j-th subiteration the algorithm considers the set $I_j \subseteq \bar{\mathcal{N}}^k_{ord}$ and solves to optimality the subproblem we get from (1.1), by fixing all the variables but the ones whose indices belong to I_j. Below we report the scheme of the proposed algorithm (see Algorithm 1).

Algorithm 1 `Fast Active SeT Block Coordinate Descent Algorithm (FAST-BCDA)`

Choose $x^0 \in \mathbb{R}^n$, **Set** $k = 0$.
For $k = 0, 1 \ldots$
  Compute $\mathcal{A}^k$, $\mathcal{N}^k$, $\bar{\mathcal{N}}^k_{ord}$;
  Set $y^{0,k}_{\mathcal{A}^k} = 0$ and $y^{0,k}_{\mathcal{N}^k} = x^k_{\mathcal{N}^k}$;
  For $j = 1, \ldots, q$
    Compute $y^{j,k}_{I_j}$, with $I_j \subseteq \bar{\mathcal{N}}^k_{ord}$, solution of problem

$$\min_{w \in R^r} g_{I_j}(y^{j-1,k})^\top (w - y^{j-1,k}_{I_j}) + \frac{1}{2}(w - y^{j-1,k}_{I_j})^\top H_{I_j I_j}(w - y^{j-1,k}_{I_j}) + \tau \|w\|_1$$

    Set $y^{j,k}_i = y^{j-1,k}_i$ if $i \notin I_j$;
  End For
  Set $x^{k+1} = y^{q,k}$;
End For

The convergence of `FAST-BCDA` is based on two important results. The first one is Proposition 3.4, which guarantees a sufficient decrease of the objective function by setting to zero the variables in the active set. The second one is reported in the proposition below. It shows that, despite the presence of the nonsmooth term, by exactly minimizing Problem (1.1) with respect to a subset J of the variables (keeping all the other variables fixed), it is possible to get a sufficient decrease of the objective function in case $\lambda_{min}(H_{JJ}) > 0$.

PROPOSITION 4.1. *Given a point $z \in \mathbb{R}^n$ and a set $J \subseteq I$, let $w^* \in \mathbb{R}^{|J|}$ be the solution of Problem* (1.1)*, where all variables but the ones whose indices belong to J*

are fixed to $z_{I\setminus J}$. *Let* $y \in \mathbb{R}^n$ *be defined as*

$$y_J = w^*, \qquad y_{I\setminus J} = z_{I\setminus J}.$$

Then we have

$$f(y) - f(z) \leq -\frac{1}{2}\lambda_{min}(H_{JJ})\|y - z\|^2. \tag{4.1}$$

Proof. See Appendix B. ∎

Now, we introduce an assumption that will enable us to prove global convergence of our algorithm.

ASSUMPTION 2. *The matrix* $A \in \mathbb{R}^{m\times n}$ *satisfies the following condition*

$$\min_J \lambda_{min}((A^\top A)_{JJ}) \geq \sigma > 0, \tag{4.2}$$

where J *is any subset of* $\{1, \ldots, n\}$ *such that* $|J| = r$, *with* r *cardinality of the blocks used in* FAST-BCDA.

REMARK 2. *We notice that even though there are some similarities between Condition* (4.2) *and the well-known Restricted Isometry Property (RIP) condition with fixed order* r *(see e.g. [6] for further details), Condition* (4.2) *is weaker than the RIP condition.*

Finally, we are ready to state the main result concerning the global convergence of FAST-BCDA.

THEOREM 4.2. *Let Assumption 1 and Assumption 2 hold. Let* $\{x^k\}$ *be the sequence produced by Algorithm* FAST-BCDA.

Then, either an integer $\bar{k} \geq 0$ *exists such that* $x^{\bar{k}}$ *is an optimal solution for Problem* (1.1), *or the sequence* $\{x^k\}$ *is infinite and every limit point* $x^\star$ *of the sequence is an optimal point for Problem* (1.1).

Proof. see Appendix B. ∎

Now, we discuss Assumptions 1 and 2 that are needed to guarantee convergence of FAST-BCDA.

4.1. Comments on the assumptions. Assumption 1 requires the evaluation of $\lambda_{max}(A^\top A)$, which is not always easily computable for large scale problems. Hence, we describe an updating rule for the parameter ϵ, that enables to avoid any "a priori" assumption on ϵ.

In practice, at each iteration k we need to find the smallest $h \in \mathbb{N}$ such that the value $\epsilon = \theta^h\tilde{\epsilon}$ and the corresponding sets $\mathcal{A}^k$, $\mathcal{N}^k$ give a point

$$y^{0,k}_{\mathcal{A}^k} = 0 \quad \text{and} \quad y^{0,k}_{\mathcal{N}^k} = x^k_{\mathcal{N}^k}$$

satisfying

$$f(y^{0,k}) \leq f(x^k) - \gamma\|y^{0,k} - x^k\|^2, \tag{4.3}$$

with $\gamma > 0$. Then, we can introduce a variation of FAST-BCDA, namely FAST-BCDA-ϵ, that includes the updating rule for the parameter ϵ in its scheme, and prove its convergence.

THEOREM 4.3. *Let Assumption 2 hold. Let $\{x^k\}$ be the sequence produced by Algorithm* FAST-BCDA-ϵ.

Then, either an integer $\bar{k} \geq 0$ exists such that $x^{\bar{k}}$ is an optimal solution for Problem (1.1)*, or the sequence $\{x^k\}$ is infinite and every limit point $x^\star$ of the sequence is an optimal point for Problem* (1.1).

Proof. The proof follows by repeating the same arguments of the proof of Theorem 4.2 by replacing the relation (B.10) with (4.3). ∎

Assumption 2, which we need to satisfy in order to guarantee convergence of both FAST-BCDA and FAST-BCDA-ϵ, is often met in practice if we consider blocks of 1 or 2 variables (i.e. r equal to 1 or 2). Indeed, when solving blocks of 1 variable, we need to guarantee that any column A_j of matrix A is such that

$$\|A_j\|^2 \geq \sigma > 0.$$

This is often the case when dealing with overcomplete dictionaries for signal/image reconstruction (as the columns of matrix A are usually normalized, see e.g. [1]). When using 2-dimensional blocks, we want no parallel columns in the matrix A. This is a quite common requirement in the context of overcomplete dictionaries (as it corresponds to ask that mutual coherence is lower than 1, see e.g. [1]). Furthermore, the solution of 1-dimensional block subproblems can be determined in closed form by means of the well-known scalar soft-threshold function (see e.g. [3, 34]). Similarly, we can express in closed form the solution of 2-dimensional block subproblems.

Summarizing, thanks to the possibility to use an updating rule for ϵ, and due to the fact that we only use blocks of dimensions 1 or 2 in our algorithm, we have that Assumptions 1 and 2 are quite reasonable in practice.

4.2. Convergence rate analysis. Here, we report a result related to the convergence rate of FAST-BCDA with 1-dimensional blocks (namely FAST-1CDA). In particular, we show that it converges at a linear rate. In order to prove the result, we make an assumption that is common when analyzing the convergence rate of both algorithms for ℓ_1-regularized problems (see e.g. [21]) and algorithms for general problems (see e.g. [25]):

ASSUMPTION 3. *Let $\{x^k\}$ be the sequence generated by* FAST-1CDA*. We have that*

$$\lim_{k\to\infty} x^k = x^\star, \tag{4.4}$$

where $x^\star$ is an optimal point of problem (1.1).

Now, we state the theoretical result related to the linear convergence.

THEOREM 4.4. *Let Assumptions 1, 2 and 3 hold. Let $\{x^k\}$ be the sequence generated by* FAST-1CDA.

Then $\{f(x^k)\}$ converges at least Q-linearly to $f^\star$, where $f^\star = f(x^\star)$. Furthermore, $\{x^k\}$ converges at least R-linearly to $x^\star$.

Proof. See Appendix C. ∎

5. Numerical Results. In this section, we report the numerical experiments related to FAST-BCDA. We implemented our method in MATLAB, and considered four different versions of it in the experiments:

- FAST-1CDA and FAST-2CDA, basic versions of FAST-BCDA where blocks of dimension 1 and 2 are respectively considered;

- `FAST-1CDA-E` and `FAST-2CDA-E`, "enhanced" versions of `FAST-BCDA` where again blocks of dimension 1 and 2 are respectively considered (see subsection 5.1 for further details).

We first analyzed the performance of these four versions of our algorithm. Then, we compared the best one with other algorithms for ℓ_1-regularized least squares problems. Namely, we compared `FAST-2CDA-E` with ISTA [3, 11], FISTA [3], PSSgb [30], SpaRSA [34] and FPC_AS [35].
All the tests were performed on an Intel Xeon(R) CPU E5-1650 v2 3.50 GHz using MATLAB R2011b.

We considered two different testing problems of the form (1.1), commonly used for software benchmarking (see e.g. [35, 18]). In particular, we generated artificial signals of dimension $n = 2^{14}, 2^{15}, 2^{16}, 2^{17}$, with a number of observations $m = n/4$ and we set the number of nonzeros $T = round(\rho\, m)$, with $\rho = \{0.01, 0.03, 0.05, 0.07, 0.1\}$. The two test problems (P1 and P2) differ in the way matrix A is generated:

P1: Considering $\bar{A}$ as the Gaussian matrix whose elements are generated independently and identically distributed from the normal distribution $\mathcal{N}(0, 1)$, the matrix A was generated by scaling the columns of $\bar{A}$.

P2: Considering $\bar{A}$ as the matrix generated by using the MATLAB command

$$A = \texttt{sprand}(m, n, \text{density}),$$

with density $= 0.5$, the matrix A was generated by scaling the columns of $\bar{A}$.

We would like to notice that the Hessian matrices $A^\top A$ related to instances of problem P1 have most of the mass on the diagonal. Then, those instances are in general easier to solve than the ones of problem P2.

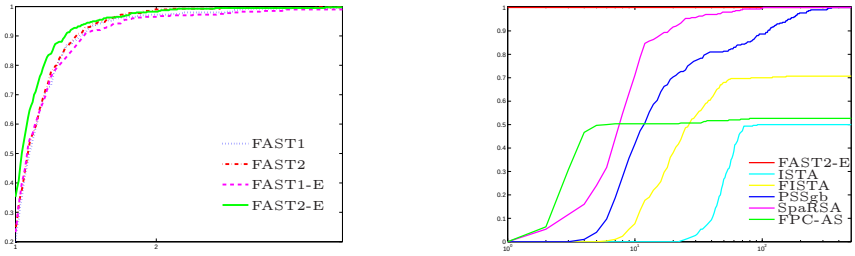
Once the matrix A was generated, the true signal $x^\star$ was built as a vector with T randomly placed ± 1 spikes, with zero in the other components. Finally, for all problems, the vector of observations b was chosen as $b = A\, x^\star + \eta$, where η is a Gaussian white noise vector, with variance 10^{-3}. We set $\tau = 0.1\|A^\top b\|_\infty$ as in [2, 34]. We produced ten different random instances for each problem, for a total of 400 instances. The comparison of the overall computational effort is carried out by using the performance profiles proposed by Dolan and Moré in [13], plotting graphs in a logarithmic scale.

For the value of s (number of non-active variables to be used in $\bar{\mathcal{N}}_{ord}$) we set $s = round(0.8\, T)$ for `FAST-1CDA` and $s = round(0.65\, T)$ for `FAST-2CDA` (these s values are the ones that guarantee the best performances among the ones we tried). For what concerns the choice of the ϵ parameter used in the active set estimate, the easiest choice is that of setting ϵ to a fixed value. We tested several values and obtained the best results with $\epsilon = 10^{-4}$ and $\epsilon = 10^{-5}$ for `FAST-1CDA` and `FAST-2CDA` respectively. We further tested an implementation of both `FAST-1CDA-`ϵ and `FAST-2CDA-`ϵ. Since there were no significant improvements in the performance, we decided to keep the ϵ value fixed.

We would also like to spend a few words about the criterion for choosing the variables in $\bar{\mathcal{N}}^k_{ord}$. In some cases, we found more efficient using the following measure:

$$\begin{cases} |g_i(x^k) + \tau| & \text{if } x^k_i > 0; \\ |g_i(x^k) - \tau| & \text{if } x^k_i < 0; \\ \max\{0, -(g_i(x^k) + \tau), g_i(x^k) - \tau\} & \text{if } x^k_i = 0, \end{cases} \tag{5.1}$$

in place of the one reported in (2.2), which we considered for proving the theoretical results. The main feature of this new measure is that it only takes into account



(a) Preliminary experience. (b) Comparison with other solvers.

Fig. 1: Performance profiles on all instances (CPU time)

first order information (while (2.2) considers proximity of the component value to zero too). Anyway, replacing (2.2) with the new measure is not a big deal, since convergence can still be proved using (5.1). Furthermore, linear rate can be easily obtained assuming that strict complementarity holds. Intuitively, considering only first order information in the choice of the variables should make more sense in our context, since proximity to zero is already taken into account when using the estimate to select the active variables.

5.1. Enhanced version of `FAST-BCDA`. By running our codes, we noticed that the cardinality of the set related to the non-active variables decreases quickly as the iterations go by. In general, very few iterations are needed to obtain the real non-active set. By this evidence, and keeping in mind the theoretical result reported in Section 3, we decided to develop an "enhanced" version of our algorithms, taking inspiration by the second stage of FPC-AS algorithm [35]. Once a "good" estimate $\mathcal{N}^k$ of $\mathcal{N}(x^\star)$ was obtained, we solved the following smooth optimization subproblem

$$\begin{array}{ll} \min & \frac{1}{2}\|Ax-b\|^2 + \tau sign(x_{\mathcal{N}^k})^\top x_{\mathcal{N}^k} \\ \text{s.t.} & x_i = 0 \quad i \in \mathcal{A}^k. \end{array}$$

In practice, we considered an estimate $\mathcal{N}^k$ "good" if both there are no changes in the cardinality of the set with respect to the last two iterations, and $|\mathcal{N}^k|$ is lower or equal than a certain threshold ξ (we fixed $\xi = 0.05n$ in our experiments).

5.2. Preliminary experiments. In order to pick the best version among the four we developed, we preliminary compared the performance of `FAST-1CDA` (FAST1), `FAST-2CDA` (FAST2), `FAST-1CDA-E` (FAST1-E) and `FAST-2CDA-E` (FAST2-E). In Figure 1a, we report the performance profiles with respect to the CPU time.

As we can see, even if the four version of `FAST-BCDA` have similar behaviour, `FAST-2CDA-E` is the one that gives the overall best performance. We then choose `FAST-2CDA-E` as the algorithm to be compared with the other state-of-the art algorithms for ℓ_1-regularized problems.

5.3. Comparison with other algorithms. In this section, we report the numerical experience related to the comparison of `FAST-2CDA-E` with ISTA [3, 11], FISTA [3], PSSgb [30], SpaRSA [34] and FPC_AS [35].

In our tests, we first ran `FAST-2CDA` to obtain a target objective function value, then ran the other algorithms until each of them reached the given target (see e.g.

[34]). Any run exceeding the limit of 1000 iterations is considered failure. Default values were used for all parameters in SpaRSA [34] and FPC_AS [35]. For PSSgb [30] we considered the two-metric projection method and we set the parameter `options.quadraticInit` to 1, since this setting can achieve better performance for problems where backtracking steps are required on each iteration (see `http://www.cs.ubc.ca/~schmidtm/Software/thesis.html`). In all codes, we considered the null vector as starting point and all matrices were stored explicitly. In Figure 1b, we report the plot of the performance profiles related to the CPU time for all instances. From these profiles it is clear that `FAST-2CDA-E` outperforms all the other algorithms and that SpaRSA and PSSgb are the two best competitors. We then further compare, in Figure 2, `FAST-2CDA-E`, SpaRSA and PSSgb reporting the box plots related to the distribution of the CPU time. On each box, the central mark is the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually. In particular, Figure 2 shows the plots related to the distribution of

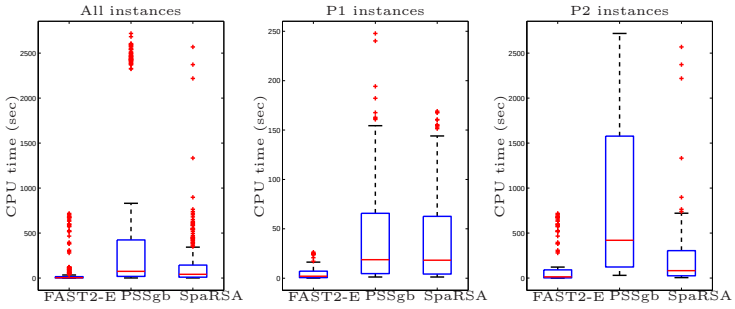


Fig. 2: Box plots (CPU time).

the CPU time for all instances, for P1 instances and for P2 instances, respectively. For what concerns P1 instances, SpaRSA and PSSgb show a similar behavior, while observing the plot related to P2 instances SpaRSA shows a better performance. For both classes, `FAST-2CDA-E` shows the lowest median. As a further comparison among `FAST-2CDA-E`, SpaRSA and PSSgb, we report in Figure 3 and in Figure 4, the plots of the relative error vs. the CPU time for the P1 and the P2 instances respectively. In each plot, the curves are averaged over the ten runs for fixed ρ and n. Observing these plots, we notice that `FAST-2CDA-E` is able to reach better solutions with lower CPU time.

5.4. Real Examples. In this subsection, we test the efficiency of our algorithm on realistic image reconstruction problems. We considered six images: a SheppLogan phantom available through the MATLAB Image Processing Toolbox and five widely used images downloaded from `http://dsp.rice.edu/cscamera` (the letter R, the mandrill, the dice, the ball, the mug). Each image has 128×128 pixels. We followed the procedure described in [35] to generate the instances (i.e. matrix A and vector b). What we want to highlight here is that the optimal solutions are unknown. Hence the reconstructed images can only be compared by visual inspection. Also in this case, we first ran `FAST-2CDA` to obtain a target objective function value, then ran the other algorithms until each of them reached the given target. The CPU-time needed for reconstructing the images is reported in Table 1. In Figure 5, we report the images

of the dice and of the mandrill reconstructed by `FAST-2CDA-E`, PSSgb and SpaRSA. It is interesting to notice that the quality of the reconstructed images can depend on the algorithm used. In Table 1, we can easily see that `FAST-BCDA` was faster in all problems.

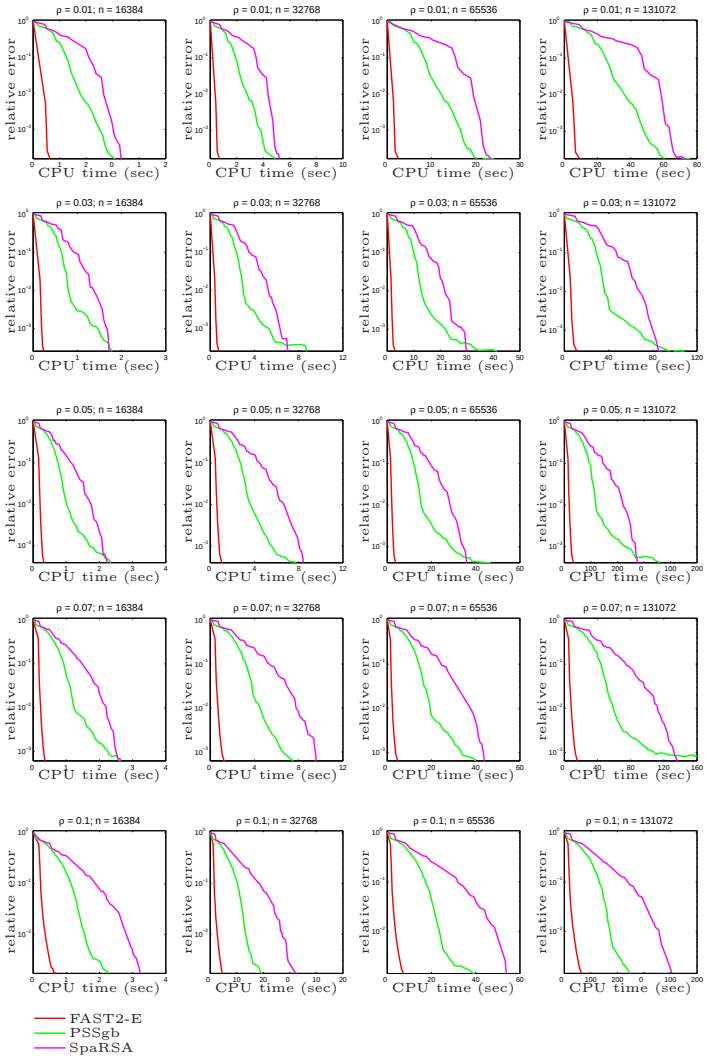


Fig. 3: Relative error vs. CPU time - P1 instances

6. Conclusions. In this paper, we devised an active set-block coordinate descent method (`FAST-BCDA`) for solving ℓ_1-regularized least squares problems. The way the active set estimate is calculated guarantees a sufficient decrease in the objective function at every iteration when setting to zero the variables estimated active. Furthermore, since the subproblems related to the blocks explicitly take into account the ℓ_1-norm, the proposed algorithmic framework does not require a sign identification strategy for the non-active variables.

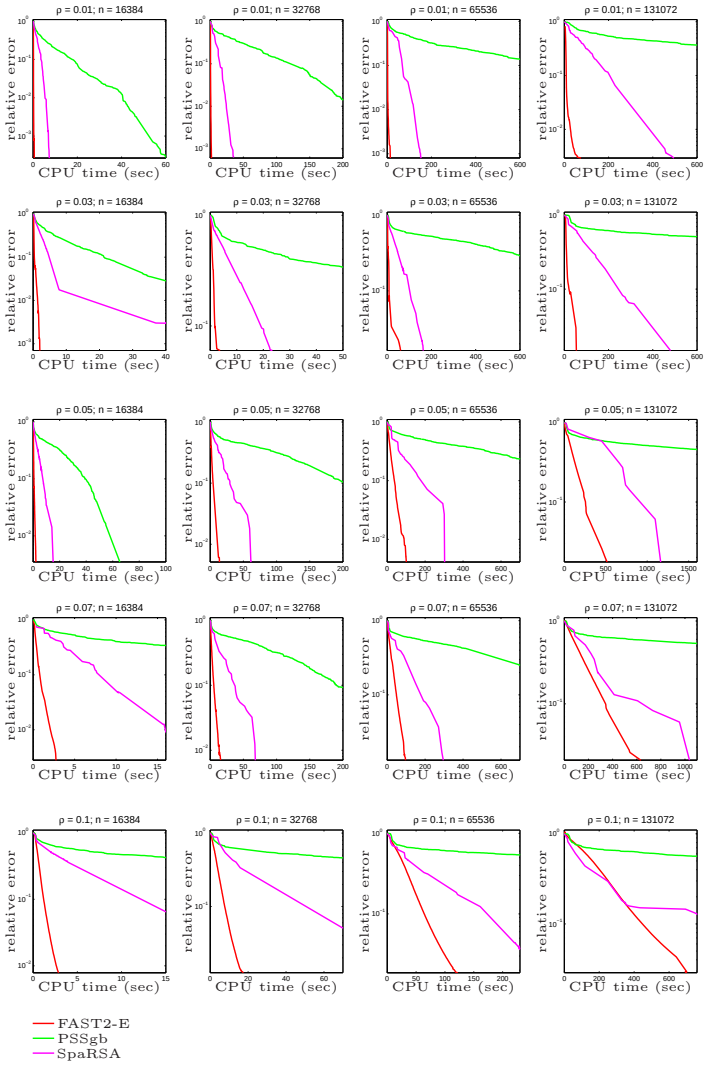


Fig. 4: Relative error vs. CPU time - P2 instances

Global convergence of the method is established. A linear convergence result is also proved. Numerical results are presented to verify the practical efficiency of the method, and they indicate that `FAST-BCDA` compares favorably with other state-of-the-art techniques.

We further would like to remark that the proposed active set strategy is independent from the specific algorithm we have designed and can be easily included into other algorithms for ℓ_1-regularized least squares, both sequential and parallel, to improve their performance. We finally highlight that the algorithmic scheme we described can be easily modified in order to work in a parallel fashion. Future work will be devoted to adapt the presented approach to handle convex ℓ_1-regularized problems.

Acknowledgments The authors would like to thank the Associate Editor and the

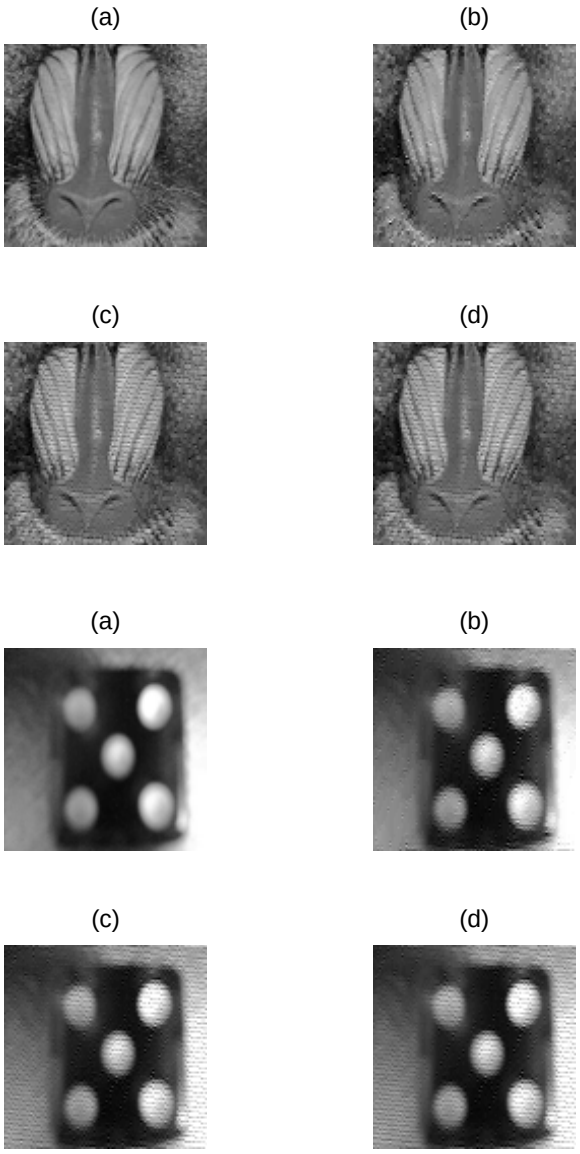


Fig. 5: Real Examples Experiment. (a) original image - (b) `FAST-2CDA-E` reconstruction - (c) PSSgb reconstruction - (d) SpaRSA reconstruction

anonymous Reviewers for their thorough and useful comments that significantly helped to improve the paper.

Appendix A. Main theoretical result related to the active set estimate. Here, we prove the main theoretical result related to the active set estimate.

Proof of Proposition 3.4. We first define the sets $\mathcal{N} = \mathcal{N}(z)$ and $\mathcal{A}{=}\mathcal{A}(z)$. By taking

Fast1	Fast2	Fast1-E	Fast2-E	ISTA	FISTA	PSSgb	SpaRSA	FPC_AS
2.18	2.46	2.54	2.02	34.65	9.26	5.01	5.08	10.46
1.65	1.67	1.51	1.95	73.01	16.31	7.77	14.48	12.87
1.86	1.97	1.91	1.65	78.05	18.41	7.90	15.56	14.69
3.52	2.05	2.13	2.32	63.13	12.69	6.51	7.53	9.77
2.29	2.12	1.79	2.16	51.12	13.79	6.36	11.13	9.33
4.12	4.21	4.16	2.41	56.66	12.09	6.69	7.12	9.72

Table 1: Real Examples Experiment - CPU time.

into account the definitions of the sets $\mathcal{A}$ and $\mathcal{N}$ and the points y and z, we have:

$$f(y) = q(y) + \tau \sum_{i=1}^{n} sign(y_i)\, y_i = q(y) + \tau \sum_{i \in \mathcal{N}} sign(y_i)\, y_i + \tau \sum_{i \in \mathcal{A}} sign(z_i)\, y_i. \tag{A.1}$$

from which

$$f(y) = f(z) + (g_{\mathcal{A}}(z) + \tau S_{\mathcal{A}} e)^\top (y - z)_{\mathcal{A}} + \frac{1}{2}(y - z)_{\mathcal{A}}^T H_{\mathcal{A}\mathcal{A}} (y - z)_{\mathcal{A}},$$

where $e \in \mathbb{R}^{|\mathcal{A}|}$ is the unit vector, and $S_{\mathcal{A}}$ is the diagonal matrix defined as

$$S_{\mathcal{A}} = Diag(sign(z_{\mathcal{A}})),$$

with the function $sign(\cdot)$ intended componentwise.
Since $H = A^\top A$ we have that the following inequality holds

$$f(y) \leq f(z) + (g_{\mathcal{A}}(z) + \tau S_{\mathcal{A}} e)^\top (y - z)_{\mathcal{A}} + \frac{\lambda_{max}(A^\top A)}{2} \|(y - z)_{\mathcal{A}}\|^2.$$

Recalling (3.12) we obtain:

$$f(y) \leq f(z) + (g_{\mathcal{A}}(z) + \tau S_{\mathcal{A}} e)^\top (y - z)_{\mathcal{A}} + \frac{1}{2\epsilon} \|(y - z)_{\mathcal{A}}\|^2. \tag{A.2}$$

Then, we can write

$$f(y) \leq f(z) + \left(g_{\mathcal{A}}(z) + \tau S_{\mathcal{A}} e + \frac{1}{\epsilon}(y - z)_{\mathcal{A}}\right)^\top (y - z)_{\mathcal{A}} - \frac{1}{2\epsilon} \|(y - z)_{\mathcal{A}}\|^2.$$

In order to prove the proposition, we need to show that

$$\left(g_{\mathcal{A}}(z) + \tau S_{\mathcal{A}} e + \frac{1}{\epsilon}(y - z)_{\mathcal{A}}\right)^\top (y - z)_{\mathcal{A}} \leq 0. \tag{A.3}$$

Inequality (A.3) follows from the fact that $\forall i \in \mathcal{A}$:

$$\left(g_i(z) + \tau sign(z_i) + \frac{1}{\epsilon}(y_i - z_i)\right)^\top (y_i - z_i) \leq 0. \tag{A.4}$$

We distinguish two cases:

a) If $z_i > 0$, we have that $sign(z_i) = 1$ and, since $y_i = 0$, $(y_i - z_i) \leq 0$.
Then, from the fact that $i \in \mathcal{A}$, we have

$$\begin{aligned} y_i &= 0 \\ z_i &\leq \epsilon\,(g_i(z) + \tau) \\ (z_i - y_i) &\leq \epsilon\,(g_i(z) + \tau) \\ \frac{1}{\epsilon}\,(z_i - y_i) &\leq g_i(z) + \tau \end{aligned}$$

so that

$$g_i(z) + \tau + \frac{1}{\epsilon}\,(y_i - z_i) \geq 0.$$

and (A.4) is satisfied.

b) If $z_i < 0$, we have that $sign(z_i) = -1$ and, since $y_i = 0$, $(y_i - z_i) \geq 0$. Then, by reasoning as in case a), from the fact that $i \in \mathcal{A}$, we can write

$$\begin{aligned} y_i &= 0 \\ -z_i &\leq \epsilon\,(\tau - g_i(z)) \\ (y_i - z_i) &\leq \epsilon\,(\tau - g_i(z)) \\ \frac{1}{\epsilon}\,(y_i - z_i) &\leq \tau - g_i(z) \end{aligned}$$

from which we have:

$$g_i(z) - \tau + \frac{1}{\epsilon}\,(y_i - z_i) \leq 0.$$

Again, we have that (A.4) is satisfied.

□

Appendix B. Theoretical results related to the convergence analysis. First, we prove the result that guarantees a sufficent decrease when minimizing with respect to a given block.

Proof of Proposition (4.1). Let us consider the subproblem obtained by fixing all variables in I but the ones whose indices belong to J to $z_{I\setminus J}$. Let $w^* \in \mathbb{R}^{|J|}$ be a solution of this subproblem.

We consider the set $J = \{j_1, \ldots, j_{|J|}\}$ as the union of two sets

$$J = J_E \cup J_D,$$

where

$$J_E = J_{E^+} \cup J_{E^-}, \qquad J_D = J_{D^+} \cup J_{D^-}$$

and

$$J_{D^+} = \{j_i \in J : sign(w_i^*) > 0\}; \qquad J_{D^-} = \{j_i \in J : sign(w_i^*) < 0\};$$

$$J_{E^+} = \{j_i \in J : w_i^* = 0;\, sign(z_{j_i}) > 0\}; \qquad J_{E^-} = \{j_i \in J : w_i^* = 0;\, sign(z_{j_i}) < 0\}.$$

Let $\tilde{f} : \mathbb{R}^{|J|} \to \mathbb{R}$, with $w \in \mathbb{R}^{|J|}$, be the following function:

$$\begin{aligned} \tilde{f}(w) = \quad & q(z) + \tau \textstyle\sum_{j \in I\setminus J} sign(z_j)\, z_j + g_J(z)^\top (w - z_J) + \tfrac{1}{2}(w - z_J)^\top H_{JJ}(w - z_J) \\ & + \tau \textstyle\sum_{j_i \in J_E} sign(z_{j_i})\, w_i + \tau \sum_{j_i \in J_D} sign(w_i^*)\, w_i. \end{aligned}$$

Then, w^* can be equivalently seen as the solution of the following problem

$$\begin{aligned} \min \quad & \tilde{f}(w) \\ s.t. \quad & w_i \geq 0 \quad \text{for } j_i \in J_{D^+} \cup J_{E^+}, \\ & w_i \leq 0 \quad \text{for } j_i \in J_{D^-} \cup J_{E^-}, \end{aligned} \tag{B.1}$$

By introducing the diagonal matrix $S = Diag(s) \in \mathbb{R}^{|J|\times|J|}$, where $s \in \{-1, 0, 1\}^{|J|}$ is the vector defined as

$$s_i = \begin{cases} sign(w_i^*) & \text{if } j_i \in J_D \\ sign(z_{j_i}) & \text{if } j_i \in J_E, \end{cases}$$

Problem (B.1) can be written in a more compact form as

$$\begin{aligned} \min \quad & \tilde{f}(w) \\ s.t. \quad & Sw \geq 0. \end{aligned} \tag{B.2}$$

From the KKT condition for Problem (B.2) at w^* we have:

$$g_J(z) + H_{JJ}(w^* - z_J) + \tau s - S\lambda = 0; \tag{B.3}$$

where $\lambda \in \mathbb{R}^{|J|}$ is the vector of multipliers with respect to the constraints $Sw \geq 0$.

We now analyze (B.3) for each index $i \in J$. We distinguish two cases:

- $j_i \in J_D$. In this case we have that $s_i = sign(w_i^*)$ and $\lambda_i = 0$. Then, from (B.3) we have

$$g_{j_i}(z) + H_{j_i j_i}(w_i^* - z_{j_i}) + \tau s_i = 0. \tag{B.4}$$

- $j_i \in J_E$. In this case we have that $s_i = sign(z_{j_i})$ and $\lambda_i \geq 0$.

Therefore,

$$\begin{aligned} g_{j_i}(z) + H_{j_i j_i}(w_i^* - z_{j_i}) + \tau s_i \geq 0 \quad & \text{if } s_i = sign(z_{j_i}) \geq 0, \\ g_{j_i}(z) + H_{j_i j_i}(w_i^* - z_{j_i}) + \tau s_i \leq 0 \quad & \text{if } s_i = sign(z_{j_i}) \leq 0. \end{aligned}$$

The previous inequalities and the fact that $w_i^* = 0$ for all $j_i \in J_E$ imply that, whatever is the sign of z_i, we have

$$\Big(g_{j_i}(z) + H_{j_i j_i}(w_i^* - z_{j_i}) + \tau s_i\Big)(w_i^* - z_{j_i}) \leq 0. \tag{B.5}$$

Taking into account (B.4) and (B.5), we have that

$$\Big(g_J(z) + H_{JJ}(w^* - z_J) + \tau s\Big)^\top (w^* - z_J) \leq 0. \tag{B.6}$$

Now, consider the difference between $\tilde{f}(w^*)$ and $\tilde{f}(z_J)$. We have that

$$\begin{aligned} \tilde{f}(w^*) - \tilde{f}(z_J) = g_J(z)^\top (w^* - z_J) + \frac{1}{2}(w^* - z_J)^\top H_{JJ}(w^* - z_J) \\ + \tau \sum_{j_i \in J_E} sign(z_{j_i})(w_i^* - z_{j_i}) + \tau \sum_{j_i \in J_D} sign(w_i^*)(w_i^* - z_{j_i}), \end{aligned}$$

which can be rewritten as

$$\tilde{f}(w^*) - \tilde{f}(z_J) = \Big(g_J(z) + H_{JJ}(w^* - z_J) + \tau s\Big)^\top (w^* - z_J) - \frac{1}{2}(w^* - z_J)^\top H_{JJ}(w^* - z_J).$$

Recalling (B.6) and the fact that $y_J = w^*$ we have

$$\tilde{f}(w^*) - \tilde{f}(z_J) \leq -\frac{1}{2}(w^* - z_J)^\top H_{JJ}(w^* - z_J) \leq -\frac{1}{2}\lambda_{min}(H_{JJ})\|y - z\|^2. \tag{B.7}$$

Since

$$q(y) = q(z) + g_J(z)^\top (y - z)_J + \frac{1}{2}(y - z)_J^\top H_{JJ}(y - z)_J,$$

by definition of $\tilde{f}$ we have that

$$\begin{aligned} f(y) = q(y) + \tau \sum_{j=1}^{n} sign(y_j) y_j = q(y) + \tau \sum_{j \in I \setminus J} sign(z_j)\, z_j + \\ + \ \tau \sum_{j_i \in J_E} sign(z_{j_i})\, w_i^* + \tau \sum_{j_i \in J_D} sign(w_i^*)\, w_i^* = \tilde{f}(w^*) \end{aligned} \tag{B.8}$$

and

$$\tilde{f}(z_J) = q(z) + \tau \sum_{j_i \in I \setminus J_D} sign(z_{j_i}) z_{j_i} + \tau \sum_{j_i \in J_D} sign(w_i^*) z_{j_i} \leq q(z) + \tau\|z\|_1 = f(z). \tag{B.9}$$

Now (B.7), (B.8) and (B.9) prove the Proposition. □

Then, we prove the main convergence result related to `FAST-BCDA`.

Proof of Theorem 4.2. We first prove that `FAST-BCDA` is well defined (in the sense that $x^{k+1} \neq x^k$ iff the point x^k is not an optimum). Let x^k not be optimum, then by contradiction we assume that $x^{k+1} = x^k$. Thus we have that either $\mathcal{N}(x^k) = \emptyset$, or forall $i \in \mathcal{A}(x^k)$, $x_i^k = 0$. This, in turns, implies that $x^k = 0$ and, by taking into account the definition of $\mathcal{A}(x^k)$, we have that x^k is optimal, thus getting a contradiction. The proof of the other implication easily follows from Propositions 3.4 and 4.1.

Let $\{y^{h,k}\}$, with $h = 0, \ldots, q$ be the sequence of points produced by Algorithm `FAST-BCDA`. By setting $y = y^{0,k}$ and $z = x^k$ in Proposition 3.4, we have:

$$f(y^{0,k}) \leq f(x^k) - \frac{1}{2\epsilon}\|y^{0,k} - x^k\|^2. \tag{B.10}$$

By setting $y = y^{h+1,k}$ and $z = y^{h,k}$, for $h = 0, \ldots, q-1$ in Proposition 4.1, we have:

$$f(y^{h+1,k}) \leq f(y^{h,k}) - \frac{\sigma}{2}\|y^{h+1,k} - y^{h,k}\|^2. \tag{B.11}$$

By using (B.10) and (B.11), we can write

$$f(x^{k+1}) \leq f(y^{q-1,k}) \leq \cdots \leq f(y^{0,k}) \leq f(x^k), \tag{B.12}$$

from which we have:

$$x^k \in \mathcal{L}^0 = \{x \in \mathbb{R}^n : \ f(x) \leq f(x^0)\}.$$

From the coercivity of the objective function of Problem (1.1) we have that the level set $\mathcal{L}^0$ is compact. Hence, the sequence $\{x^k\}$ has at least a limit point and

$$\lim_{k\to\infty} \big(f(x^{k+1}) - f(x^k)\big) = 0. \tag{B.13}$$

Now, let $x^\star$ be any limit point of the sequence $\{x^k\}$ and $\{x^k\}_K$ be the subsequence such that

$$\lim_{k\to\infty, k\in K} x^k = x^\star. \tag{B.14}$$

Let us assume, by contradiction, that $x^\star$ is not an optimal point of Problem (1.1). By taking into account that inequality $\|\sum_{i=1}^l a_i\| \leq l \ \sum_{i=1}^l \|a_i\|^2$ holds for the squared norm of sums of l vectors a_i, and by recalling (B.10), (B.11) and (B.12), we have

$$f(x^{k+1}) \leq f(y^{0,k}) \leq f(x^k) - \frac{1}{2\epsilon}\|y^{0,k} - x^k\|^2, \tag{B.15}$$

$$f(x^{k+1}) \leq f(y^{h,k}) \leq f(x^k) - \frac{\sigma}{2}\|y^{h,k} - x^k\|^2. \tag{B.16}$$

with $h = 1, \ldots, q$.

Now, (B.13), (B.14), (B.15) and (B.16) imply

$$\lim_{k\to\infty, k\in K} y^{h,k} = x^\star, \tag{B.17}$$

for $h = 0, \ldots, q$.

For every index $j \in \mathcal{A}^k$, we can define the point $\tilde{y}^{j,k}$ as follows:

$$\tilde{y}_i^{j,k} = \begin{cases} 0 & \text{if } i = j \\ x_i^k & \text{otherwise} \end{cases} \tag{B.18}$$

Recalling the definition of points $\tilde{y}^{j,k}$ and $y^{0,k}$, we have

$$\|\tilde{y}^{j,k} - x^k\|^2 = (\tilde{y}^{j,k} - x^k)_j^2 = (y^{0,k} - x^k)_j^2 \leq (y^{0,k} - x^k)_j^2 + \sum_{i\in\mathcal{A}^k, i\neq j} (x_i^k)^2 = \|y^{0,k} - x^k\|^2.$$

From the last inequality and (B.17) we obtain

$$\lim_{k\to\infty, k\in K} \tilde{y}^{j,k} = x^\star, \tag{B.19}$$

for all $j \in \mathcal{A}^k$.

To conclude the proof, we consider the function $\Phi_i(x)$, defined in (2.2), that measures the violation of the optimality conditions for a variable x_i.

Since, by contradiction, we assume that $x^\star$ is not an optimal point there must exists an index $\hat{\imath}$ such that

$$|\Phi_{\hat{\imath}}(x^\star)| > 0. \tag{B.20}$$

Taking into account that the number of possible different choices of $\mathcal{A}^k$ and $\mathcal{N}^k$ is finite, we can find a subset $\hat{K} \subseteq K \subseteq \{1, 2, 3, \dots\}$ such that $\mathcal{A}^k = \hat{\mathcal{A}}$ and $\mathcal{N}^k = \hat{\mathcal{N}}$ for all $k \in \hat{K}$. We can have two different cases: either $\hat{\imath} \in \hat{\mathcal{A}}$ or $\hat{\imath} \in \hat{\mathcal{N}}$ for k sufficiently large.

Suppose first that $\hat{\imath} \in \hat{\mathcal{A}}$ for k sufficiently large. Then, by Definition 3.1, we have for all $k \in \hat{K}$:

$$\max\{0, x_{\hat{\imath}}^k\} \le \epsilon\,(g_{\hat{\imath}}(x^k) + \tau) \quad \text{and} \quad \max\{0, -x_{\hat{\imath}}^k\} \le \epsilon\,(\tau - g_{\hat{\imath}}(x^k)).$$

For all $k \in \hat{K}$, let $\tilde{y}^{\hat{\imath},k}$ be the point defined as in (B.18). By construction we have that

$$\tilde{y}_{\hat{\imath}}^{\hat{\imath},k} = 0. \tag{B.21}$$

Now we consider three different subcases:

i) $x_{\hat{\imath}}^k > 0$. In this case, (B.18) and (B.21) imply

$$(\tilde{y}_{\hat{\imath}}^{\hat{\imath},k} - x_{\hat{\imath}}^k) \le 0. \tag{B.22}$$

Recalling (3.12) in Assumption 1, there exists $\rho \ge 0$, such that

$$\epsilon \le \frac{1}{H_{\hat{\imath}\hat{\imath}} + \rho}.$$

Furthermore, since $\hat{\imath} \in \hat{\mathcal{A}}$, we can write

$$\begin{aligned} x_{\hat{\imath}}^k &\le \epsilon\,(g_{\hat{\imath}}(x^k) + \tau) \\ x_{\hat{\imath}}^k - \tilde{y}_{\hat{\imath}}^{\hat{\imath},k} &\le \epsilon\,(g_{\hat{\imath}}(x^k) + \tau) \\ x_{\hat{\imath}}^k - \tilde{y}_{\hat{\imath}}^{\hat{\imath},k} &\le \frac{1}{H_{\hat{\imath}\hat{\imath}} + \rho}(g_{\hat{\imath}}(x^k) + \tau) \end{aligned}$$

Then we have:

$$(H_{\hat{\imath}\hat{\imath}} + \rho)(x_{\hat{\imath}}^k - \tilde{y}_{\hat{\imath}}^{\hat{\imath},k}) \le g_{\hat{\imath}}(x^k) + \tau,$$

which can be rewritten as follows

$$g_{\hat{\imath}}(x^k) + H_{\hat{\imath}\hat{\imath}}(\tilde{y}_{\hat{\imath}}^{\hat{\imath},k} - x_{\hat{\imath}}^k) + \tau \ge \rho(x_{\hat{\imath}}^k - \tilde{y}_{\hat{\imath}}^{\hat{\imath},k}) \ge 0,$$

that is

$$g_{\hat{\imath}}(\tilde{y}^{\hat{\imath},k}) + \tau \ge 0. \tag{B.23}$$

On the other hand, since

$$0 \le \max\{0, -x_{\hat{\imath}}^k\} \le \epsilon\,(\tau - g_{\hat{\imath}}(x^k))$$

we have that $g_{\hat{\imath}}(x^k) - \tau \le 0$ and, as $H_{\hat{\imath}\hat{\imath}} \ge 0$ and (B.22) holds, we get

$$g_{\hat{\imath}}(\tilde{y}^{\hat{\imath},k}) - \tau = g_{\hat{\imath}}(x^k) + H_{\hat{\imath}\hat{\imath}}(\tilde{y}_{\hat{\imath}}^{\hat{\imath},k} - x_{\hat{\imath}}^k) - \tau \le 0. \tag{B.24}$$

By (B.21), (B.23) and (B.24), we have that

$$|\Phi_{\hat{\imath}}(\tilde{y}^{\hat{\imath},k})| = 0.$$

Furthermore, by (B.19) and the continuity of Φ, we can write

$$|\Phi_{\hat{\imath}}(x^\star)| = 0.$$

Thus we get a contradiction with (B.20).

ii) $x_{\hat{\imath}}^k < 0$. It is a verbatim repetition of the previous case.

iii) $x_{\hat{\imath}}^k = 0$. Since $\hat{\imath} \in \hat{\mathcal{A}}$ we have

$$g_{\hat{\imath}}(x^k) + \tau \geq 0 \quad \text{and} \quad -(g_{\hat{\imath}}(x^k) - \tau) \geq 0,$$

which imply that

$$|\Phi_{\hat{\imath}}(x^k)| = 0.$$

By the continuity of $\Phi(\cdot)$ and the fact that

$$\lim_{k\to\infty, k\in\hat{K}} x^k = x^\star,$$

we get a contradiction with (B.20).

Suppose now that $\hat{\imath} \in \hat{\mathcal{N}}$ for k sufficiently large. We can choose a further subsequence $\{x^k\}_{\tilde{K}}$ with $\tilde{K} \subseteq \hat{K}$ such that

$$|\Phi_{\bar{\imath}}(x^k)| = \max_{i\in\hat{\mathcal{N}}} |\Phi_i(x^k)|, \quad \forall\ k \in \tilde{K}.$$

Hence,

$$|\Phi_{\bar{\imath}}(x^k)| \geq |\Phi_{\hat{\imath}}(x^k)|, \quad \forall\ k \in \tilde{K}, \tag{B.25}$$

which, by continuity of $\Phi(\cdot)$, implies

$$|\Phi_{\bar{\imath}}(x^\star)| \geq |\Phi_{\hat{\imath}}(x^\star)|. \tag{B.26}$$

Furthermore, the instructions of Algorithm `FAST-BCDA` guarantee that, for all $k \in \tilde{K}$, a set of indices I_{h_k} exists such that

$$\bar{\imath} \in I_{h_k} \subseteq \bar{\mathcal{N}}_{ord}^k.$$

For all $k \in \tilde{K}$, Algorithm `FAST-BCDA` produces a vector $y^{h_k,k}$ by minimizing Problem (1.1) with respect to all the variables whose indices belong to I_{h_k}. Therefore, the point $y^{h_k,k}$ satisfies

$$|\Phi_{\bar{\imath}}(y^{h_k,k})| = 0.$$

Furthermore, by (B.17), the continuity of $\Phi(\cdot)$, and taking into account (B.26), we can write

$$0 = |\Phi_{\bar{\imath}}(x^\star)| \geq |\Phi_{\hat{\imath}}(x^\star)|,$$

which contradicts (B.20). □

Appendix C. Theoretical results related to the convergence rate analysis. Here, following the ideas in [24], we prove that the convergence rate of `FAST-BCDA` with 1-dimensional blocks (namely `FAST-1CDA`) is linear. First, we try to better analyze the indices in the set $\mathcal{N}(x)$ by introducing the following two sets:

$$\mathcal{N}^+(x) = \{i \in \mathcal{N}(x) :\ g_i(x) \leq 0\}, \text{ and } \mathcal{N}^-(x) = \{i \in \mathcal{N}(x) :\ g_i(x) > 0\}. \tag{C.1}$$

We further introduce the sets:

$$\mathcal{E}^+(x^\star) = \{i : x_i^\star \geq 0,\ g_i(x^\star) = -\tau\},\ \text{and}\ \mathcal{E}^-(x^\star) = \{i :\ x_i^\star \leq 0,\ g_i(x^\star) = \tau\}, \tag{C.2}$$

which satisfy the following equality:

$$\mathcal{E}(x^\star) = \mathcal{E}^+(x^\star) \cup \mathcal{E}^-(x^\star) = \bar{\mathcal{N}}(x^\star) \cup \{i : x_i^\star = 0,\ |g_i(x^\star)| = \tau\}.$$

We further notice that

$$I = \bar{\mathcal{A}}^+(x^\star) \cup \mathcal{E}^+(x^\star) \cup \mathcal{E}^-(x^\star). \tag{C.3}$$

We can finally prove a result that will be used in the convergence analysis:

THEOREM C.1. *Let $x^\star \in \mathbb{R}^n$ be a solution of Problem* (1.1). *Then, there exists a neighborhood of $x^\star$ such that, for each x in this neighborhood, we have*

$$\mathcal{N}^+(x) \subseteq \mathcal{E}^+(x^\star), \tag{C.4}$$

$$\mathcal{N}^-(x) \subseteq \mathcal{E}^-(x^\star). \tag{C.5}$$

Proof. Let us assume there exists a sequence $\{\epsilon^k\}$, $\epsilon^k \to 0$, a related sequence of neighborhoods $\{\mathcal{B}(x^\star, \epsilon^k)\}$ and a sequence of points $\{x^k\}$ such that $x^k \in \mathcal{B}(x^\star, \epsilon^k)$ for all k, satisfying the following:

$$\mathcal{N}^+(x^k) \not\subseteq \mathcal{E}^+(x^\star).$$

Then, since the number of indices is finite, there exist subseqences $\{\epsilon^k\}_K$ and $\{\mathcal{B}(x^\star, \epsilon^k)\}_K$ such that an index $\hat{\imath}$ can be found, satisfying the following:

$$\hat{\imath} \in \mathcal{N}^+(x^k), \quad \hat{\imath} \notin \mathcal{E}^+(x^\star).$$

From Theorem 3.3, for k sufficiently large,

$$\mathcal{N}(x^k) \subseteq \mathcal{E}^+(x^\star) \cup \mathcal{E}^-(x^\star).$$

Therefore, we have that

$$\hat{\imath} \in \mathcal{E}^-(x^\star)\ \text{ and }\ g_{\hat{\imath}}(x^\star) = \tau.$$

By continuity of the gradient, $g_{\hat{\imath}}(x^k) > 0$ for k sufficiently large. On the other hand, since $\hat{\imath} \in \mathcal{N}^+(x^k)$, we have $g_{\hat{\imath}}(x^k) \leq 0$. This gives a contradiction, and proves (C.4). A similar reasoning can be used for proving (C.5). ∎

Finally, we report another theoretical result that is used in the convergence rate analysis.

PROPOSITION C.2. *Let Assumption 3 hold. Then, there exists a $\bar{k}$ such that*

a) $x_i^k = 0, \quad i \in \bar{\mathcal{A}}^+(x^\star)$;

b) $-sign(g_i(x^\star))\ x_i^k \geq 0, \quad i \in \mathcal{E}(x^\star)$;

for all $k \geq \bar{k}$.

Proof. a). Recalling (3.10), for k sufficently large, we have

$$\bar{\mathcal{A}}^+(x^\star) \subseteq \mathcal{A}^k.$$

Therefore, by taking into account the steps of `FAST-1CDA` Algorithm, we have

$$x_i^{k+1} = 0, \quad i \in \bar{\mathcal{A}}^+(x^\star). \tag{C.6}$$

Furthermore, by continuity of g and (4.4), we obtain

$$\tau + g_i(x^{k+1}) > 0, \quad \text{and} \quad \tau - g_i(x^{k+1}) > 0, \tag{C.7}$$

and we can write

$$i \in \mathcal{A}^{k+1}.$$

Hence, (C.6) and (C.7) still hold for x^{k+2}, and so on.

b). Let us consider an index $i \in \mathcal{E}(x^\star)$. By contradiction, we assume that there exists a subsequence $K = \{k_1, k_2, \dots\}$ such that

$$-sign(g_i(x^\star))\ x_i^k < 0, \tag{C.8}$$

for all $k \in K$. Without any loss of generality, we can consider another subsequence $\hat{K} = \{\bar{k}_1, \bar{k}_2, \dots\}$ related to K, such that $i \in \mathcal{N}^{\bar{k}_j}$ and

$$x_i^{k_j} = -sign\left(g_i(x^{\bar{k}_j}) - H_{ii}x_i^{\bar{k}_j}\right)\ \frac{\max\left\{\left|g_i(x^{\bar{k}_j}) - H_{ii}x_i^{\bar{k}_j}\right| - \tau, 0\right\}}{H_{ii}}, \tag{C.9}$$

for all $k_j \in K$ and $\bar{k}_j \in \hat{K}$.

If $i \in \mathcal{E}(x^\star) \setminus \bar{\mathcal{N}}(x^\star)$, when j is sufficiently large, we have by continuity of g, (4.4) and (C.9)

$$-sign(g_i(x^\star))\ x_i^{k_j} \geq 0,$$

which contradicts (C.8).

If $i \in \bar{\mathcal{N}}(x^\star)$, when j is sufficiently large, we have by continuity of g, (3.8) and (4.4)

$$-sign(g_i(x^\star))\ x_i^{k_j} \geq 0,$$

and again we get a contradiction with (C.8). ∎

Now, we prove that the algorithm converges at linear rate.

*Proof of Theorem 4.4.*First of all, for ease of notation we set $\bar{\mathcal{A}}^+ = \bar{\mathcal{A}}^+(x^\star)$, and $\mathcal{E} = \mathcal{E}(x^\star)$. Without any loss of generality, we can assume $|\bar{\mathcal{N}}_{ord}^k| = 1$ for all k. We then notice that the objective function $f(x)$ can be rewritten as follows:

$$f(x) = q(x) + \tau \sum_{i \in \bar{\mathcal{A}}^+} sign(x_i)x_i + \tau \sum_{i \in \mathcal{E}} sign(x_i)x_i.$$

We further introduce the function

$$F(x) = q(x) - \sum_{i \in \mathcal{E}} sign(g_i(x^\star))x_i.$$

By taking into account Proposition C.2, we have, for k sufficiently large,

$$f(x^k) = F(x^k), \quad \text{and} \quad f(x^\star) = F(x^\star). \tag{C.10}$$

Furthermore, when k is sufficiently large, by definition of $\bar{\mathcal{A}}^+$ and $\mathcal{E}$, and recalling again Proposition C.2, we can write

$$x_{\bar{\mathcal{A}}^+}^k = x_{\bar{\mathcal{A}}^+}^\star, \tag{C.11}$$

$$\nabla_i F(x^\star) = 0, \quad \forall\ i \in \mathcal{E}. \tag{C.12}$$

Then, by considering (C.3), (C.11) and (C.12), it follows

$$\begin{aligned} F(x^k) - F(x^\star) &= \nabla F(x^\star)^\top (x^k - x^\star) + \frac{1}{2}(x^k - x^\star)^\top \nabla^2 F(x^\star)(x^k - x^\star) \\ &= \frac{1}{2}(x^k - x^\star)^\top \nabla^2 F(x^\star)(x^k - x^\star) \leq \frac{\lambda_{max}(\nabla^2 F(x^\star))}{2}\|x^k - x^\star\|^2, \end{aligned}$$

and, taking into account (C.10), we can write

$$f(x^k) - f(x^\star) \leq \rho \|x^k - x^\star\|^2, \tag{C.13}$$

with $\rho > 0$. Then, recalling Theorem 3.3 and C.1, for k sufficiently large the problem we actually solve is

$$\begin{aligned} \min \quad & \tilde{F}(x) = \frac{1}{2}\|Ax - b\|^2 - \tau \; sign(g_{\mathcal{N}^k}(x^k))^\top x_{\mathcal{N}^k} \\ & x_{\mathcal{A}^k} = 0 \\ & -sign(g_i(x^k)) \; x_i \geq 0 \quad i \in \mathcal{N}^k. \end{aligned} \tag{C.14}$$

Now, let $y^{0,k}$ be the point obtained at Step 4 of Algorithm 1 (i.e. after fixing to zero the active variables) and $y_s^{0,k}$ the component that most violates condition (2.2) in the non-active set. We notice that finding the most violating variable according to condition (2.2) is equivalent, when considering Problem (C.14), to get the component that most violates the following condition

$$|x_i - [x_i - \nabla_i \tilde{F}(x)]_+|,$$

see [24] for further details. Thus, we can write

$$\begin{aligned} \frac{1}{\sqrt{|\mathcal{N}^k|}}\|y^{0,k} - [y^{0,k} - \nabla \tilde{F}(y^{0,k})]_+\| &\leq |y_s^{0,k} - [y_s^{0,k} - \nabla_s \tilde{F}(y^{0,k})]_+| \\ &= |y_s^{0,k} - [y_s^{0,k} - \nabla_s \tilde{F}(y^{0,k})]_+ - x_s^{k+1} + [x_s^{k+1} - \nabla_s \tilde{F}(x^{k+1})]_+| \\ &\leq 2|y_s^{0,k} - x_s^{k+1}| + |\nabla_s \tilde{F}(y^{0,k}) - \nabla_s \tilde{F}(x^{k+1})| \\ &\leq 2\|y^{0,k} - x^{k+1}\| + \|\nabla \tilde{F}(y^{0,k}) - \nabla \tilde{F}(x^{k+1})\| \\ &\leq M\|y^{0,k} - x^{k+1}\| = M\|x^k - x^{k+1}\|_{\mathcal{N}^k}, \end{aligned} \tag{C.15}$$

where $[\cdot]_+$ is the projection on the set of inequalities in Problem (C.14), and $M = \max\{2, L\}$, with L Lipschitz constant of $\nabla \tilde{F}$. By using Propositions 3.4 and 4.1, we can also write:

$$f(x^k) - f(x^{k+1}) \geq \delta \|x^{k+1} - x^k\|^2 \tag{C.16}$$

with $\delta > 0$. By taking into account inequality (C.15) and the definition of $y^{0,k}$, we can write, for k sufficiently large,

$$\begin{aligned} \|x^{k+1} - x^k\|^2 &= \|x^k - x^\star\|^2_{\mathcal{A}^k} + \|y^{0,k} - x^{k+1}\|^2_{\mathcal{N}^k} \\ &\geq \|x^k - x^\star\|^2_{\mathcal{A}^k} + \frac{1}{M\sqrt{|\mathcal{N}^k|}}\|y^{0,k} - [y^{0,k} - \nabla \tilde{F}(y^{0,k})]_+\|^2. \end{aligned} \tag{C.17}$$

Now, considering Theorem 2.1 in [24] we have, for k sufficiently large,

$$\sigma\|y^{0,k} - [y^{0,k} - \nabla \tilde{F}(y^{0,k})]_+\| \geq \|y^{0,k} - x^\star\| = \|x^k - x^\star\|_{\mathcal{N}^k},$$

with $\sigma > 0$. Therefore, by taking into account inequality (C.17), we can write

$$\|x^{k+1} - x^k\|^2 \geq \|x^k - x^\star\|^2_{\mathcal{A}^k} + \gamma\|x^k - x^\star\|^2_{\mathcal{N}^k} \geq \tilde{\gamma}\|x^k - x^\star\|^2, \tag{C.18}$$

with $\tilde{\gamma} > 0$. By combining inequalities (C.13), (C.16) and (C.18), we can write

$$f(x^k) - f(x^\star) \leq c_1 \left(f(x^k) - f(x^{k+1})\right), \tag{C.19}$$

with $c_1 > 1$. After rearranging the terms in (C.19), we obtain

$$f(x^{k+1}) - f(x^\star) \leq c_2 \left(f(x^k) - f(x^\star) \right)$$

with $c_2 = \left(1 - \frac{1}{c_1}\right) < 1$. Then, $\{f(x^k)\}$ converges at least linearly to $f^\star$.

Finally, by using (C.16) and Lemma 3.1 in [24] we get that the sequence $\{x^k\}$ converges at least linearly to $x^\star$. □

REFERENCES

[1] M. Aharon, M. Elad and A. M. Bruckstein. *On the uniqueness of overcomplete dictionaries and a practical way to retrieve them.* Linear Algebra Appl., 416, pp. 48–67, 2006.

[2] M. V. Afonso, J. M. Bioucas-Dias, and M. A. T. Figueiredo. *Fast image recovery using variable splitting and constrained optimization.* IEEE Trans. on Image Proc., 19(9), pp. 2–45, 2010.

[3] A. Beck and M. Teboulle. *A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problem.* SIAM J. Imaging Sciences, 2(1), pp. 183–202, 2009.

[4] J. M. Bioucas-Dias, and M.Figueiredo. *A New TwIST: Two-Step Iterative Shrinkage/Thresholding Algorithms for Image Restoration.* IEEE Trans. on Image Proc., 16(12), pp. 2992–3004, 2007.

[5] R. H. Byrd, G. M. Chi, J. Nocedal, and F. Oztoprak. *A Family of Second-Order Methods for Convex L1-Regularized Optimization.* Optimization Center: Northwestern University, Tech Report, 2012.

[6] E. J. Candes and T. Tao, *Decoding by Linear Programming*, IEEE Trans. Inf. Th., 51(12), pp. 4203–4215, 2005.

[7] E. Candès, J. Romberg, and T. Tao. *Stable signal recovery from incomplete and inaccurate measurements.* Comm. Pure Appl. Math., 59(8), pp. 1207–1223, 2006.

[8] K. W. Chang, C. J. Hsieh, and C. J. Lin. *Coordinate descent method for large-scale L2-loss linear SVM.* J. Mach. Learn. Res., 9, pp. 1369–1398, 2008.

[9] P. Combettes and V. Wajs. *Signal recovery by proximal forward-backward splitting.* Multiscale Model. Simul., 4(4), pp. 1168–1200, 2005.

[10] A. Daneshmand, F. Facchinei, V. Kungurtsev, G. Scutari. *Hybrid Random/Deterministic Parallel Algorithms for Nonconvex Big Data Optimization.* arXiv:1407.4504v2, 2014

[11] I. Daubechies, M. Defriese, and C. De Mol. *An iterative thresholding algorithm for linear inverse problems with a sparsity constraint.* Comm. Pure Appl. Math., 57(11), pp. 1413–1457, 2004.

[12] M. De Santis, G. Di Pillo, and S. Lucidi. *An active set feasible method for large- scale minimization problems with bound constraints.* Comput. Opt. Appl., 53(2), pp. 395–423, 2012.

[13] E. D. Dolan, and J. J. Moré. *Benchmarking optimization software with performance profiles.* Math. Program., 91, pp. 201–213, 2002.

[14] F. Facchinei and S. Lucidi. *Quadratically and Superlinearly Convergent Algorithms for the Solution of Inequality Constrained Minimization Problems.* J. Optim. Theory Appl., 85(2), pp. 265-289, 1995.

[15] F. Facchinei, S. Sagratella, and G. Scutari. *Flexible Parallel Algorithms for Big Data Optimization.* arXiv:1311.2444, 2013.

[16] K. Fountoulakis and R. Tappenden. *Robust Block Coordinate Descent.* Technical Report ERGO-14-010, 2014.

[17] M. Fukushima. *Parallel Variable Transformation in Unconstrained Optimization.* Siam J. on Optimization, 8(4), pp. 658–672, 1998.

[18] K. Fountoulakis and J. Gondzio. *A Second-Order Method for Strongly Convex L1-Regularization Problems* Technical Report ERGO-13-011, School of Mathematics, The University of Edinburgh, 2013.

[19] G. H. Golub, P. C. Hansen, and D. P. O'Leary. *Tikhonov regularization and total least squares.* SIAM J. Matrix Anal. Appl., 21, pp. 185–194, 1999.

[20] R. Griesse and D. A. Lorenz. *A semismooth Newton method for Tikhonov functionals with sparsity constraints.* Inverse Problems 24(3),pp. 1–19, 2008.

[21] E. T. Hale, W. Yin,Y. Zhang. *Fixed-point continuation for \ell_1-minimization: Methodology and convergence.* SIAM J. on Optimization, 19(3), pp. 1107–1130, 2008.

[22] P. C. Hansen and D. P. O'Leary. *The use of the L-curve in the regularization of discrete ill-posed problems.* SIAM J. Sci. Comput., 14, pp. 1487–1503, 1993.

[23] C. J. Hsieh, M. A. Sustik, I. S. Dhillon, P. Ravikumar, *Sparse Inverse Covariance Matrix Estimation Using Quadratic Approximation.* Advances in Neural Information Processing Systems, vol. 24, 2011.

[24] Z.Q. Luo, P. Tseng *On the linear convergence of descent methods for convex essentially smooth minimization.* SIAM J. on Cont. and Opt. 30(2) pp. 408–425, 1992.

[25] J. M. Ortega, W. C. Rheinboldt *Iterative solution of nonlinear equations in several variables.* SIAM, Vol. 30, 1970.

[26] Z. Peng, M. Yan, and W. Yin. *Parallel and Distributed Sparse Optimization.* preprint, 2013.

[27] M. Porcelli and F. Rinaldi. *Variable fixing version of the two-block nonlinear constrained Gauss-Seidel algorithm for l1-regularized least-squares.* Comput. Opt. Appl., 2014. DOI: 10.1007/s10589-014-9653-0.

[28] P. Richtárik and M. Takáč. *Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function.* Math. Program., 2012.

[29] P. Richtárik and M. Takáč. *Parallel Coordinate Descent Methods for Big Data Optimization.* arXiv:1212.0873, 2012.

[30] M. Schmidt. *Graphical Model Structure Learning with ℓ_1-Regularization* Phd Thesis. 2010.

[31] A. N. Tikhonov and V. Y. Arsenin. *Solution of Ill-Posed Problems.* V. H. Winston, Washington, DC, 1977.

[32] P. Tseng. *A coordinate gradient descent method for nonsmooth separable minimization.* J. Optim. Theory Appl., 109(3), pp. 475–494, 2001.

[33] P. Tseng, S. Yun. *A coordinate gradient descent method for nonsmooth separable minimization.* Math. Program., 117(1), pp. 387–423, 2009.

[34] S. Wright, R. Nowak, and M. Figueiredo. *Sparse Reconstruction by Separable Approximation.* IEEE Trans. on Signal Proc. Vol. 57(7), pp. 2479–2493, 2009.

[35] Z. Wen, W. Yin, D. Goldfarb, and Y. Zhang. *A fast algorithm for sparse reconstruction based on shrinkage, subspace optimization and continuation.* SIAM J. Sci. Comput., 32(4), pp. 1832–1857, 2010.

[36] Z. Wen, W. Yin, H. Zhang, and D. Goldfarb. *On the convergence of an active-set method for l1 minimization.* Optim. Methods Softw., 27(6), pp. 1127–1146, 2012.

[37] S. J. Wright. *Accelerated Block-coordinate Relaxation for Regularized Optimization.* SIAM J. on Optimization., 22(1), pp. 159–186, 2012.

[38] G. X. Yuan, K. W. Chang, C. J. Hsieh, and C. J. Lin. *A Comparison of Optimization Methods and Software for Large-scale L1-regularized Linear Classification.* J. Mach. Learn. Res., 11, pp. 3183–3234, 2010.

[39] S. Yun and K. Toh. *A coordinate gradient descent method for l1-regularized convex minimization.* Comput. Opt. Appl., 48(2), pp. 273–307, 2011.