On the convergence of a regularized Jacobi algorithm for convex optimization

Goran Banjac, Kostas Margellos, and Paul J. Goulart

***Abstract*—In this paper we consider the regularized version of the Jacobi algorithm, a block coordinate descent method for convex optimization with an objective function consisting of the sum of a differentiable function and a block-separable function. Under certain regularity assumptions on the objective function, this algorithm has been shown to satisfy the so-called sufficient decrease condition, and consequently to converge in objective function value. In this paper we revisit the convergence analysis of the regularized Jacobi algorithm and show that it also converges in iterates under very mild conditions on the objective function. Moreover, we establish conditions under which the algorithm achieves a linear convergence rate.**

***Index Terms*—Decentralized optimization, Jacobi algorithm, block coordinate descent methods, linear convergence.**

I. Introduction

In this paper we consider large-scale optimization problems in which a collection of individual actors (or *agents*) cooperate to minimize some common objective function while incorporating local constraints or additional local utility functions. We consider a decentralized optimization method based on block coordinate descent, an iterative coordinating procedure which has attracted significant attention for solving large-scale optimization problems [1]–[3].

Solving large-scale optimization problems via an iterative procedure that coordinates among blocks of variables enables the solution of very large problem instances by parallelizing computation across agents. This enables one to overcome computational challenges that would be prohibitive otherwise, without requiring agents to reveal their local utility functions and constraints to other agents. Due to its pricing mechanism implications, decentralized optimization is also a natural choice for many applications, including demand side management in smart grids, charging coordination for plug-in electric vehicles, coordination of multiple agents in robotic systems etc. [4]–[6].

Based on the algorithms outlined in [2], two classes of iterative methods have been employed recently for solving such optimization problems in a decentralized way. The first covers block coordinate gradient descent (BCGD) methods and it requires each agent to perform, at every iteration, a local (proximal) gradient descent step [1], [6]. Under certain regularity assumptions (differentiability of the objective function and Lipschitz continuity of its gradient), and for an appropriately chosen gradient step size, this method converges to a minimizer of the centralized problem. This class of algorithms includes both sequential [7] and parallel [8], [9] implementations.

The second covers block coordinate minimization (BCM) methods, does not assume differentiability of the objective and is based on minimizing the common objective function in each block by fixing variables associated with other agents to their previously computed values. Although BCM methods have a larger per iteration cost than the BCGD methods in the case when there are no local utility functions (constraints) in the problem, or when their proximal operators (projections) have closed-form solutions, in the general case both approaches require solutions of ancillary optimization problems. On the other hand, iterations of BCM methods are numerically more stable than gradient iterations, as observed in [10].

If the block-wise minimizations are done in a cyclic fashion across agents, then the algorithm is known as the Gauss-Seidel algorithm [3], [7], [11]. An alternative implementation, known as the Jacobi algorithm, involves performing the block-wise minimizations in parallel. However, convergence of the Jacobi algorithm is not guaranteed in general, even in the case when the objective function is smooth and convex, unless certain contractiveness properties are satisfied [2, Prop. 2.6 in Sec 3.2 & Prop. 3.10 in Sec 3.3].

The authors in [12] have proposed a regularized Jacobi algorithm wherein, at each iteration, each agent minimizes the weighted sum of the common objective function and a quadratic regularization term penalizing the distance to the previous iterate of the algorithm. A similar regularization has been used in Gauss-Seidel methods [7], [11] which are however not parallelizable. Under certain regularity assumptions, and for an appropriately selected regularization weight, the algorithm converges *in objective value* to the optimal value of the centralized problem [12]. Recently, the authors in [13] have quantified the regularization weighting required to ensure convergence in objective value as a function of the number of agents and other problem parameters. However, convergence of the algorithm *in its iterates* to an optimizer of the centralized problem counterpart was not established, apart from the particular case where the objective function is quadratic.

In this paper we revisit the algorithm proposed in [12] and enhance its convergence properties under milder conditions. By adopting an analysis based on a power growth property, which is in turn sufficient for the satisfaction of the so-called Kurdyka-Łojasiewicz condition [11], [14], we show that the algorithm's iterates converge under much milder assumptions

The authors are with the Department of Engineering Science, University of Oxford, Parks Road, Oxford, OX1 3PJ, UK. Email: {goran.banjac, kostas.margellos, paul.goulart}@eng.ox.ac.uk

This work was supported by the European Commission research project FP7-PEOPLE-2013-ITN under grant agreement no. 607957 [*Training in Embedded Optimization and Predictive Control (TEMPO)*].

on the objective function than those used in [2] and [13]. A similar approach was used in [3], [11] to establish convergence of iterates generated by Gauss-Seidel type methods. We also show that the algorithm achieves a linear convergence rate without imposing restrictive strong convexity assumptions on the objective function, in contrast to typical methods in the literature. Our analysis is based on the quadratic growth condition, which is closely related to the so-called error bound property [15], [16] that is used in [8] to establish linear convergence of parallel BCGD methods *in objective value*.

The remainder of the paper is organized as follows. In Section II we introduce the class of problems under study, outline the regularized Jacobi algorithm for solving such problems in a decentralized fashion, and state the main convergence result of the paper. Section III provides the proof of the main result. Section IV provides a convergence rate analysis, while Section V concludes the paper.

Notation

Let $\mathbb{N}$ denote the set of nonnegative integers, $\mathbb{R}$ the set of real numbers, $\mathbb{R}_+$ the set of nonnegative real numbers, $\bar{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$ the extended real line, and $\mathbb{R}^n$ the n-dimensional real space equipped with inner product $\langle x, y\rangle$ and induced norm $\|x\|$. Consider a vector $x = (x^1, \ldots, x^m)$ where $x^i \in \mathbb{R}^{n_i}$, $i = 1, \ldots, m$. We denote by x^{-i} the remainder of vector x when component x^i is removed. Denote the effective domain of $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ as $\operatorname{dom} f := \{x \in \mathbb{R}^n \,|\, f(x) < \infty\}$. The directional derivative of f at $x \in \operatorname{dom} f$ in the direction $d \in \mathbb{R}^n$ is denoted by $f'(x, d)$. The subdifferential of f at x is denoted by $\partial f(x)$. If f is continuously differentiable, then $\nabla f(x)$ denotes the gradient of f evaluated at x. We denote by $[a \le f \le b] := \{x \in \mathbb{R}^n \,|\, a \le f(x) \le b\}$ a set of points whose value of function f is between a and b; similar notation will be used for strict inequalities and for one-sided bounds. The set of minimizers of f is denoted by $\operatorname{argmin} f := \{x \in \operatorname{dom} f \,|\, f(x) = \min f\}$, where $\min f$ is the minimum value of f. We say that a differentiable function f is strongly convex with convexity parameter $\sigma > 0$ if

$$f(y) \ge f(x) + \langle \nabla f(x), y - x\rangle + \frac{\sigma}{2}\|y - x\|^2,$$

holds for all x and y. The distance of a point x to a closed convex set $\mathcal{C}$ is denoted by $\operatorname{dist}(x, \mathcal{C}) := \inf_{c\in\mathcal{C}} \|x - c\|$, and the projection of x onto $\mathcal{C}$ is denoted by $\operatorname{proj}_{\mathcal{C}}(x) := \{y \in \mathcal{C} \,|\, \|x - y\| = \operatorname{dist}(x, \mathcal{C})\}$.

II. Problem description and main result

A. Regularized Jacobi algorithm

We consider the following optimization problem:

$$\min_{\{x^i\}_{i=1}^m} \Big\{ f(x^1, \ldots, x^m) + \sum_{i=1}^m g_i(x^i) \Big\}, \qquad (\mathcal{P})$$

where $x := (x^1, \ldots, x^m) \in \mathbb{R}^n$, $x^i \in \mathbb{R}^{n_i}$ and $n = \sum_{i=1}^m n_i$. To simplify subsequent derivations we define $f(x) := f(x^1, \ldots, x^m)$, $g(z) := \sum_{i=1}^m g_i(z^i)$ with $\operatorname{dom} g = \operatorname{dom} g_1 \times \ldots \times \operatorname{dom} g_m$, and the combined objective function in $\mathcal{P}$ as

$$h(x) := f(x) + g(x). \qquad (1)$$

Algorithm 1 Regularized Jacobi algorithm

1: **Initialization**
2: $k = 0$.
3: Set $x_0^i \in \operatorname{dom} g_i$, for all $i = 1, \ldots, m$.
4: **For** $i = 1, \ldots, m$ **repeat until convergence**
5: $x_{k+1}^i = \operatorname{argmin}_{z^i} \{ f_i(z^i; x_k^{-i}) + g_i(z^i) + c\|z^i - x_k^i\|^2 \}$.
6: $k \leftarrow k + 1$.

Problems in the form $\mathcal{P}$ can be viewed as multi-agent optimization programs wherein each agent has its own local decision vector x^i and agents cooperate to determine a minimizer of h, which couples the local decision vectors of all agents through the common objective function f. Since the number of agents can be large, solving the problem in a centralized fashion may be computationally intensive. Moreover, even if this were possible from a computational point of view, agents may not be willing to share their local objectives g_i, $i = 1, \ldots, m$, with other agents, since this may encode information about their local utility functions or constraint sets.

For each $i = 1, \ldots, m$, we let $f_i(\,\cdot\,; x^{-i}) : \mathbb{R}^{n_i} \mapsto \mathbb{R}$ be a function of the decision vector of the i-th block of variables, with the remaining variables $x^{-i} \in \mathbb{R}^{n-n_i}$ treated as a fixed set of parameters, i.e.,

$$f_i(z^i; x^{-i}) := f(x^1, \cdots, x^{i-1}, z^i, x^{i+1}, \cdots, x^m).$$

We wish to solve $\mathcal{P}$ in a decentralized fashion using Algorithm 1. At the $(k+1)^{\text{th}}$ iteration of Algorithm 1, agent i solves a local optimization problem accounting for its local function g_i and the function f_i with the parameter vector set to the decisions x_k^{-i} of the other agents from the previous iteration. Moreover in the local cost function an additional term penalizes the squared distance between the optimization variables and their values at the previous iteration x_k^i. The relative importance of the original cost function and the penalty term is regulated by the weight $c > 0$, which should be selected large enough to guarantee convergence [12], [13]. We show in the Appendix that the fixed points of Algorithm 1 coincide with optimal solutions of problem $\mathcal{P}$.

A problem structure equivalent to $\mathcal{P}$ was considered in [13], with the difference that a collection of convex constraints $x^i \in \mathcal{X}_i$ for each $i = 1, \ldots, m$ were introduced instead of the functions g_i. We can rewrite this problem in the form of $\mathcal{P}$ by selecting g_i to be an indicator function of a given convex set. On the other hand, problem $\mathcal{P}$ can be written in epigraph form, and thus reformulated in the framework of [13]. The reason that we use the problem structure of $\mathcal{P}$ is twofold. First, some widely used problems such as ℓ_1-regularized least squares are typically posed in the form $\mathcal{P}$. Second, the absence of constraints will ease the convergence analysis of Section III since many results in the relevant literature use the same problem structure.

B. Statement of the main result

Before stating the main result we provide some necessary definitions and assumptions. Let $h^\star$ denote the minimum value of $\mathcal{P}$. We then have the following definition.

Definition 1 (Power-type growth condition): A function $h : \mathbb{R}^n \mapsto \tilde{\mathbb{R}}$ satisfies a *power-type growth condition* on $[h^\star < h < h^\star + r]$ if there exist $r > 0$, $\gamma > 0$ and $p \geq 1$ such that, for all $x \in [h^\star < h < h^\star + r]$,

$$h(x) - h^\star \geq \gamma \operatorname{dist}(x, \operatorname{argmin} h)^p. \tag{2}$$

It should be noted that (2) is a very mild condition, since it requires only that the function h is not excessively 'flat' in the neighborhood of the set $\operatorname{argmin} h$. For instance, all polynomial, real-analytic and semi-algebraic functions satisfy this condition [14], [17].

We impose the following standing assumptions on problem $\mathcal{P}$:

Assumption 1:

a) The function f is convex and differentiable.
b) The gradient ∇f is Lipschitz continuous on $\operatorname{dom} g$ with Lipschitz constant L, i.e.,
$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \quad \forall x, y \in \operatorname{dom} g.$$
c) The functions g_i are all convex, lower semicontinuous and proper.
d) The function h is coercive, i.e.
$$\lim_{\|x\| \to +\infty} h(x) = +\infty.$$
e) The function h exhibits the power-type growth condition of Definition 1.

Notice that we do not require differentiability of the functions g_i. Coerciveness of h implies the existence of some $\zeta \in \mathbb{R}$ for which the sublevel set $[h \leq \zeta]$ is nonempty and bounded, which is sufficient to prove existence of a minimizer of h [18, Prop. 11.12 & Thm. 11.9].

We are now in a position to state the main result of the paper.

Theorem 1: Under Assumption 1, if

$$c > \frac{m-1}{2m-1}\sqrt{m-1}L, \tag{3}$$

then the iterates $\{x_k\}_{k\in\mathbb{N}}$ generated by Algorithm 1 converge to a minimizer of problem $\mathcal{P}$, i.e., $\lim_{k\to\infty} x_k = x^*$, where x^* is a minimizer of $\mathcal{P}$.

The proof of Theorem 1 involves several intermediate statements and is provided in the next section.

III. Proof of the main result

Many results on convergence of optimization algorithms establish only convergence in function value [2], [13], [19], without guaranteeing convergence of the iterates $\{x_k\}_{k\in\mathbb{N}}$ as well. Convergence of iterates is straightforward to show when h is strongly convex, or when $\{x_k\}_{k\in\mathbb{N}}$ is Fejér monotone with respect to $\operatorname{argmin} h$, which is true whenever the operator underlying the iteration update is nonexpansive [18]. The latter condition was used in [13] to establish convergence of the sequence $\{x_k\}_{k\in\mathbb{N}}$ in the special case that f is a convex quadratic function.

In the single-agent case, i.e. when $m = 1$, Algorithm 1 reduces to the proximal minimization algorithm whose associated fixed-point operator is nonexpansive for any convex, lower semicontinuous and proper function h. However, in the multi-agent setting the resulting fixed-point operator is not necessarily nonexpansive, which implies that the Fejér monotonicity based analysis can not be employed to establish convergence of the sequence $\{x_k\}_{k\in\mathbb{N}}$. To achieve this and prove Theorem 1 we exploit the following result, which follows directly from Theorem 14 in [14].

Theorem 2 ([14, Thm. 14]): Consider Assumption 1, with $\operatorname{argmin} h \neq \emptyset$ and $h^\star := \min h$. Assume that the initial iterate x_0 of Algorithm 1 satisfies $h(x_0) < h^\star + r$, where r is as in Definition 1. Finally, assume that subsequent iterates $\{x_k\}_{k\in\mathbb{N}}$ generated by Algorithm 1 possess the following properties:

1) Sufficient decrease condition:
$$h(x_{k+1}) \leq h(x_k) - a\|x_{k+1} - x_k\|^2, \tag{4}$$
where $a > 0$.
2) Relative error condition: There exists $w_{k+1} \in \partial h(x_{k+1})$ such that
$$\|w_{k+1}\| \leq b\|x_{k+1} - x_k\|, \tag{5}$$
where $b > 0$.

Then the sequence $\{x_k\}_{k\in\mathbb{N}}$ converges to some $x^\star \in \operatorname{argmin} h$, i.e. $\lim_{k\to\infty} x_k = x^*$, and for all $k \geq 1$

$$\begin{aligned}\|x_k - x^\star\| \leq \frac{b}{a}\frac{p}{\gamma^{\frac{1}{p}}}\left(h(x_k) - h^\star\right)^{\frac{1}{p}} \\ + \sqrt{\frac{1}{a}\left(h(x_{k-1}) - h^\star\right)}.\end{aligned} \tag{6}$$

It should be noted that Theorem 2 constitutes a relaxed version of Theorem 14 in [14]. This is due to the fact that we impose the power-type growth property as an assumption, which is in turn a sufficient condition for the satisfaction of the so-called Kurdyka-Łojasiewicz (KL) property[1] [11], [17]. Specifically, we could replace the last part of Assumption 1 with the KL property and the conclusion of Theorem 2 would remain valid.

Notice that, under the assumptions of Theorem 2, $\{x_k\}_{k\in\mathbb{N}}$ converges to some $x^\star \in \operatorname{argmin} h$ even if $h(x_0) \geq h^\star + r$. Since $\{h(x_k)\}_{k\in\mathbb{N}}$ converges to $h^\star$ (as a consequence of the sufficient decrease condition (4)), there exists some $k_0 \in \mathbb{N}$ such that $h(x_{k_0}) < h^\star + r$, and hence Theorem 2 remains valid if x_k is replaced by x_{k+k_0}.

To prove Theorem 1 it suffices to show that, given Assumption 1, the iterates generated by Algorithm 1 satisfy the *sufficient decrease condition* and the *relative error condition*. To show this we first provide an auxiliary lemma.

Lemma 1: Under Assumption 1, for all $x, y, z \in \operatorname{dom} g$

$$\begin{aligned}\Big\|\sum_{i=1}^m \nabla f_i(z^i; x^{-i}) - \sum_{i=1}^m \nabla f_i(z^i; y^{-i})\Big\| \\ \leq \sqrt{m-1}L\|x - y\|.\end{aligned}$$

[1] This can be seen by choosing the so-called *desingularizing function* φ that appears in the definition of the KL property [11], [17] such that

$$\varphi(s) = p\,(s/\gamma)^{\frac{1}{p}}\,.$$

Proof: The statement follows from [13, Lemma 1]. However, by noticing that $\sum_{i=1}^m \|x^{-i} - y^{-i}\|^2 = (m-1)\|x-y\|^2$ instead of $m\|x-y\|^2$, we obtain an improvement in the bound of [13, Lemma 1]. ∎

We can then show that the sufficient decrease condition is satisfied.

Proposition 1 (Sufficient decrease condition): Under Assumption 1, if c is chosen according to (3), then Algorithm 1 converges to the minimum of problem $\mathcal{P}$ in value, i.e. $h(x_k) \to \min h$, and for all k the sufficient decrease condition (4) is satisfied with

$$a = \left(c - (m-1)(\sqrt{m-1}L - 2c)\right)/m > 0. \tag{7}$$

Proof: The result follows from [13, Theorem 2], with the Lipschitz constant established in Lemma 1. ∎

Note that the proofs of Lemma 1 and Proposition 1 do not require the last part of Assumption 1 related to the power-type growth condition of h.

If c is chosen according to Theorem 1, then (4) implies that $x_{k+1} - x_k \to 0$. To show this, suppose that $x_0 \in \operatorname{dom} h$ and thus $h(x_0)$ is finite. Iterating the inequality (4) gives

$$a\sum_{k=0}^{\infty} \|x_{k+1} - x_k\|^2 \leq h(x_0) - h^\star < +\infty,$$

which means that $\|x_{k+1} - x_k\|$ converges to zero. Note however that this does not necessarily imply convergence of the sequence $\{x_k\}_{k\in\mathbb{N}}$.

Proposition 2 (Relative error condition): Consider Algorithm 1. Under Assumption 1, there exists $w_{k+1} \in \partial h(x_{k+1})$ such that the relative error condition (5) is satisfied with

$$b = 2c + \sqrt{m-1}L > 0. \tag{8}$$

Proof: Iterate x_{k+1} in Algorithm 1 can be characterized via the subdifferential of the associated objective function, i.e.,

$$0 \in \sum_{i=1}^m \nabla f_i(x_{k+1}^i; x_k^{-i}) + \partial g(x_{k+1}) + 2c(x_{k+1} - x_k),$$

which ensures the existence of some $v_{k+1} \in \partial g(x_{k+1})$ such that

$$\begin{aligned}
0 &= \sum_{i=1}^m \nabla f_i(x_{k+1}^i; x_k^{-i}) + v_{k+1} + 2c(x_{k+1} - x_k) \\
&= \Big[\sum_{i=1}^m \nabla f_i(x_{k+1}^i; x_k^{-i}) - \sum_{i=1}^m \nabla f_i(x_{k+1}^i; x_{k+1}^{-i})\Big] \\
&\quad + \sum_{i=1}^m \nabla f_i(x_{k+1}^i; x_{k+1}^{-i}) + v_{k+1} + 2c(x_{k+1} - x_k) \\
&= \Big[\sum_{i=1}^m \nabla f_i(x_{k+1}^i; x_k^{-i}) - \sum_{i=1}^m \nabla f_i(x_{k+1}^i; x_{k+1}^{-i})\Big] \\
&\quad + \nabla f(x_{k+1}) + v_{k+1} + 2c(x_{k+1} - x_k).
\end{aligned}$$

Notice that in the last equality we used the identity $\sum_{i=1}^m \nabla f_i(x^i; x^{-i}) = \nabla f(x)$.

Let us now define $w_{k+1} \coloneqq \nabla f(x_{k+1}) + v_{k+1} \in \partial h(x_{k+1})$. From the above equality we can bound the norm of w_{k+1} as

$$\begin{aligned}
\|w_{k+1}\| &= \Big\| \sum_{i=1}^m \nabla f_i(x_{k+1}^i; x_k^{-i}) - \sum_{i=1}^m \nabla f_i(x_{k+1}^i; x_{k+1}^{-i}) \\
&\qquad + 2c(x_{k+1} - x_k)\Big\| \\
&\leq \Big\| \sum_{i=1}^m \nabla f_i(x_{k+1}^i; x_k^{-i}) - \sum_{i=1}^m \nabla f_i(x_{k+1}^i; x_{k+1}^{-i})\Big\| \\
&\quad + 2c\|x_{k+1} - x_k\|.
\end{aligned}$$

The last step follows from the triangle inequality, and due to Lemma 1, we obtain

$$\|w_{k+1}\| \leq (2c + \sqrt{m-1}L)\|x_{k+1} - x_k\|.$$

∎

Propositions 1 and 2 show that the conditions of Theorem 2 are satisfied. As a direct consequence the iterates generated by Algorithm 1 converge to some minimizer of $\mathcal{P}$, thus concluding the proof of Theorem 1.

IV. Convergence rate analysis

It is shown in [13] that if f is a strongly convex quadratic function and g_i are indicator functions of convex compact sets, then Algorithm 1 converges linearly. We show in this section that Algorithm 1 converges linearly under much milder assumptions. In particular, if h has the *quadratic growth property*, i.e., if p in (2) is equal to 2, then Algorithm 1 admits a linear convergence rate. This property is employed in [20] to establish linear convergence of some first-order methods in a single-agent setting, and is, according to [15], [16] closely related to the *error bound*, which was used in [21], [22] to establish linear convergence of feasible descent methods. Note that the feasible descent methods are not applicable to problem $\mathcal{P}$ since we allow for nondifferentiable objective functions.

Theorem 3: Consider Assumption 1, and further assume that power-type growth property is satisfied with $p = 2$. Let the initial iterate of Algorithm 1 be selected such that $h(x_0) < h^\star + r$, where r appears in Definition 1. Then the iterates $\{x_k\}_{k\in\mathbb{N}}$ converge to some $x^\star \in \operatorname{argmin} h$, and for all $k \geq 1$

$$h(x_k) - h^\star \leq \left(\frac{1}{1+\gamma a b^{-2}}\right)^k (h(x_0) - h^\star), \tag{9}$$

$$\|x_k - x^\star\| \leq M\left(\frac{1}{\sqrt{1+\gamma a b^{-2}}}\right)^k, \tag{10}$$

where

$$M = \left(\frac{2b}{\gamma a} + \frac{1}{\sqrt{a\,(1+\gamma a b^{-2})}}\right)\sqrt{h(x_0) - h^\star}.$$

Proof: The quadratic growth property and convexity of h, together with the relative error condition (5) imply that for $x_{k+1} \notin \operatorname{argmin} h$ and $\bar{x}_{k+1} = \operatorname{proj}_{\operatorname{argmin} h}(x_{k+1})$

$$\begin{aligned}\gamma \operatorname{dist}(x_{k+1}, \operatorname{argmin} h)^2 &\leq h(x_{k+1}) - h^\star \\ &\leq \langle w_{k+1},\, x_{k+1} - \bar{x}_{k+1} \rangle \\ &\leq \|w_{k+1}\| \, \| x_{k+1} - \bar{x}_{k+1}\| \\ &= \|w_{k+1}\| \operatorname{dist}(x_{k+1}, \operatorname{argmin} h) \\ &\leq b\|x_{k+1} - x_k\| \operatorname{dist}(x_{k+1}, \operatorname{argmin} h),\end{aligned} \tag{11}$$

where $w_{k+1} \in \partial h(x_{k+1})$. Note that since h is lower semi-continuous, the set $\operatorname{argmin} h$ is closed and thus the projection onto $\operatorname{argmin} h$ is well defined. From the right-hand sides of the first and last inequality in (11) we have

$$h(x_{k+1}) - h^\star \leq b\|x_{k+1} - x_k\| \operatorname{dist}(x_{k+1}, \operatorname{argmin} h).$$

Dividing the left-hand side of the first inequality and the right-hand side of the last inequality in (11) by $\gamma \operatorname{dist}(x_{k+1}, \operatorname{argmin} h) > 0$, we obtain

$$\operatorname{dist}(x_{k+1}, \operatorname{argmin} h) \leq \frac{b}{\gamma}\|x_{k+1} - x_k\|.$$

Substituting this inequality into the preceding one, we obtain

$$\begin{aligned}h(x_{k+1}) - h^\star &\leq \frac{b^2}{\gamma}\|x_{k+1} - x_k\|^2 \\ &\leq \frac{b^2}{\gamma a}\left(h(x_k) - h(x_{k+1})\right) \\ &= \frac{b^2}{\gamma a}\left((h(x_k) - h^\star) - (h(x_{k+1}) - h^\star)\right),\end{aligned}$$

where the second inequality follows from the sufficient decrease condition (4). Rearranging the terms, we have that

$$h(x_{k+1}) - h^\star \leq \frac{1}{1 + \gamma a b^{-2}}\left(h(x_k) - h^\star\right),$$

for all $k \geq 0$, or equivalently

$$h(x_k) - h^\star \leq \left(\frac{1}{1 + \gamma a b^{-2}}\right)^k \left(h(x_0) - h^\star\right),$$

which proves (9). Substituting the above inequality into (6) we obtain (10), which concludes the proof. ∎

A direct consequence of Theorem 3 is that Algorithm 1, with c selected as in Theorem 1, converges linearly when h satisfies the quadratic growth condition

$$h(x) - h^\star \geq \gamma \operatorname{dist}(x, \operatorname{argmin} h)^2. \tag{12}$$

This is the case when f is strongly convex with convexity parameter σ_f, implying that $\operatorname{argmin} h$ is a singleton and h has the quadratic growth property with $\gamma = \sigma_f/2$ for any $x \in \operatorname{dom} h$. It is shown in [22], [23] that if $f(x) = v(Ex) + \langle b, x\rangle$ has a Lipschitz continuous gradient, with v being strongly convex, and g being an indicator function of a convex polyhedral set, then the problem exhibits the quadratic growth property.

Note that if E does not have full column rank, then f is not strongly convex. In [14], [23] it is shown that a similar bound can be established for the ℓ_1-regularized least-squares problem. Here, we adopt an approach from [14] and show that a similar result can be provided for more general problems in which g can be any polyhedral function. The core idea is to rewrite the problem in epigraph form for which such a property is shown to hold.

We impose the following assumption.

Assumption 2:

a) The function f is defined as $f(x) = v(Ex) + \langle b, x\rangle$, with $v(\cdot)$ being a strongly convex function with convexity parameter σ_v.

b) The component functions g_i are all globally non-negative convex polyhedral functions whose composite epigraph can be represented as

$$\begin{aligned}\operatorname{epi} g &:= \left\{(x,t) \in \mathbb{R}^{n+1} \mid g(x) \leq t\right\} \\ &= \left\{(x,t) \in \mathbb{R}^{n+1} \mid Cx + ct \leq d\right\},\end{aligned}$$

where $C \in \mathbb{R}^{p\times n}$, $c \in \mathbb{R}^p$ and $d \in \mathbb{R}^p$. Note that the inequality $Cx + ct \leq d$ should be taken component-wise.

The conditions of Assumption 2 are satisfied when f is quadratic and g_i are indicator functions of convex polyhedral sets or any polyhedral norms. Note that the dual of a quadratic program satisfies this assumption. The Lipschitz constant of ∇f, which is required for computing the appropriate parameter c for Algorithm 1, can be upper bounded by $\|E\|^2 L_v$, where $\|E\|$ is the spectral norm of E, and L_v is the Lipschitz constant of ∇v. We will now define the Hoffman constant which will be used in the further analysis.

Lemma 2 (Hoffman constant, see e.g., [23]): Let X and Y be two polyhedra defined as

$$X = \{x \in \mathbb{R}^n \mid Ax \leq a\}, \quad Y = \{x \in \mathbb{R}^n \mid Ex = e\},$$

where $A \in \mathbb{R}^{m\times n}$, $a \in \mathbb{R}^m$, $E \in \mathbb{R}^{p\times n}$, $e \in \mathbb{R}^p$, and assume that $X \cap Y \neq \emptyset$. Then there exists a constant $\theta = \theta(A, E)$ such that any $x \in X$ satisfies

$$\operatorname{dist}(x, X \cap Y) \leq \theta \, \|Ex - e\|.$$

We refer to θ as the *Hoffman constant* associated with matrix $[A^T, E^T]^T$.

Let x_0 be an initial iterate of the algorithm and let $r = h(x_0)$. Since h is coercive, $[h \leq r]$ is a compact set and we can thus define the following quantities:

$$\begin{aligned}D^r &:= \max_{x,y\in[h\leq r]} \|x - y\|, \\ D_E^r &:= \max_{x,y\in[h\leq r]} \|Ex - Ey\| \leq D\,\|E\|, \\ V^r &:= \max_{x\in[h\leq r]} \|\nabla v(Ex)\|.\end{aligned}$$

Since Algorithm 1 generates a non-increasing sequence $\{h(x_k)\}_{k\in\mathbb{N}}$, for all k we have $x_k \in [h \leq r]$ and

$$\begin{aligned}g(x_k) &\leq g(x_0) + f(x_0) - f(x_k) \\ &\leq g(x_0) + v(Ex_0) - v(Ex_k) + \langle b, x_0 - x_k\rangle \\ &\leq g(x_0) + \|\nabla v(Ex_0)\| \, \|Ex_0 - Ex_k\| \\ &\quad + \|b\| \, \|x_0 - x_k\| \\ &\leq g(x_0) + V^r D_E^r + \|b\| D^r.\end{aligned}$$

We conclude that $\operatorname{argmin} h \subseteq [h \leq r] \subset [g \leq R]$, for any fixed $R > g(x_0) + V^r D_E^r + \|b\| D^r$. For such a bound R, we have

$$\begin{aligned}
&\min\Big\{v(Ex) + \langle b, x\rangle + g(x) \mid x \in \mathbb{R}^n\Big\}\\
&= \min\Big\{v(Ex) + \langle b, x\rangle + t \quad \mid (x,t) \in \mathbb{R}^n \times \mathbb{R},\\
&\qquad\qquad\qquad\qquad\qquad\qquad g(x) \leq R, t = g(x)\Big\}\\
&= \min\Big\{v(Ex) + \langle b, x\rangle + t \quad \mid (x,t) \in \mathbb{R}^n \times \mathbb{R},\\
&\qquad\qquad\qquad\qquad\qquad\qquad g(x) \leq t, t \leq R\Big\}\\
&= \min\Big\{\underbrace{v(\tilde{E}\tilde{x}) + \langle \tilde{b}, \tilde{x}\rangle}_{=:\tilde{h}(\tilde{x})} \quad \mid \tilde{x} \in \mathbb{R}^{n+1}\\
&\qquad\qquad\qquad\qquad\qquad\qquad \tilde{\mathcal{X}} := \{M\tilde{x} \leq \tilde{R}\}\Big\}, \quad (13)
\end{aligned}$$

where $\tilde{x} = (x, t)$ and

$$\tilde{E} = \begin{bmatrix} E & 0 \end{bmatrix}, \quad \tilde{b} = \begin{bmatrix} b \\ 1 \end{bmatrix}, \quad M = \begin{bmatrix} C & c \\ 0 & 1 \end{bmatrix}, \quad \tilde{R} = \begin{bmatrix} d \\ R \end{bmatrix}.$$

It can be easily seen that $\tilde{x}^\star = (x^\star, t^\star)$ minimizes (13) if and only if $x^\star \in \operatorname{argmin} h$ and $t^\star = g(x^\star)$. Using [23, Lemma 2.5], we obtain

$$\operatorname{dist}(\tilde{x}, \operatorname{argmin} \tilde{h})^2 \leq \kappa_R \left(\tilde{h}(\tilde{x}) - \tilde{h}^\star\right), \quad \forall \tilde{x} \in \tilde{\mathcal{X}}, \quad (14)$$

where $\kappa_R = \theta^2 \left(\|\tilde{b}\| \tilde{D}^R + 3\tilde{V}^R \tilde{D}_{\tilde{E}}^R + \frac{2((\tilde{V}^R)^2+1)}{\sigma_v}\right)$ and θ is the Hoffman constant associated with matrix $[M^T, \tilde{E}^T, \tilde{b}]^T$. Moreover,

$$\begin{aligned}
\tilde{D}^R &:= \max_{\tilde{x},\tilde{y} \in \tilde{\mathcal{X}}} \|\tilde{x} - \tilde{y}\| \leq \max_{x,y \in [g \leq R]} \|x - y\| + \max_{t,s \in [0,R]} \|t - s\|\\
&= D^R + R,\\
\tilde{V}^R &:= \max_{\tilde{x} \in \tilde{\mathcal{X}}} \|\nabla v(\tilde{E}\tilde{x})\| = \max_{x \in [g \leq R]} \|\nabla v(Ex)\| = V^R,\\
\tilde{D}_{\tilde{E}}^R &:= \max_{\tilde{x},\tilde{y} \in \tilde{\mathcal{X}}} \|\tilde{E}\tilde{x} - \tilde{E}\tilde{y}\| = \max_{x,y \in [g \leq R]} \|Ex - Ey\| = D_E^R.
\end{aligned}$$

Inequality (14) implies that for all $x \in [g \leq R]$ and for all $t \in [0, R]$

$$\begin{aligned}
\operatorname{dist}(x, \operatorname{argmin} h)^2 &+ \|t - t^\star\|^2\\
&\leq \kappa_R \left(f(x) + t - f(x^\star) - t^\star\right).
\end{aligned}$$

Setting $t = g(x)$, we then have that

$$\begin{aligned}
\operatorname{dist}(x, \operatorname{argmin} h)^2 &\leq \operatorname{dist}(x, \operatorname{argmin} h)^2 + \|t - t^\star\|^2\\
&\leq \kappa_R \left(h(x) - h^\star\right).
\end{aligned}$$

Lemma 3: Let $r = h(x_0)$ and fix any $R > g(x_0) + V^r D_E^r + \|b\| D^r$. Under Assumptions 1 and 2, for all $x \in [h \leq r]$ we have

$$h(x) - h^\star \geq \kappa_R^{-1} \operatorname{dist}(x, \operatorname{argmin} h)^2,$$

where

$$\kappa_R = \theta^2 \Big((\|b\| + 1)(D^R + R) + 3V^R D_E^R + \frac{2\left((V^R)^2 + 1\right)}{\sigma_v}\Big).$$

V. Conclusion

In this paper we revisited the regularized Jacobi algorithm proposed in [12], and enhanced its convergence properties. It was shown that iterates generated by the algorithm converge to a minimizer of the centralized problem counterpart, provided that the objective function satisfies a power growth property. We also established linear convergence of the algorithm when the power growth condition satisfied by the objective function is quadratic.

Appendix

In this section we show that the set of fixed points of Algorithm 1 coincides with the set of minimizers of problem $\mathcal{P}$. The result follows from [13, §3]; however, the proof is modified to account for the presence of the nondifferentiable terms g_i, $i = 1, \ldots, m$. We first recall the optimality condition for a nondifferentiable convex function h.

Proposition 3 ([18, Proposition 17.3]): Let the function $h : \mathbb{R}^n \to \tilde{\mathbb{R}}$ be proper and convex and let $x^\star \in \operatorname{dom} h$. Then $x^\star \in \operatorname{argmin} h$ if and only if $h'(x^\star, d) \geq 0$ for all $d \in \mathbb{R}^n$.

Proposition 3 states that $x^\star$ is a local minimizer of h if and only if there is no local direction d along which the function attains lower value. For convex functions a local minimizer is also global.

Similarly to [13], we define an operator T such that

$$T(x) = \operatorname*{argmin}_z \Big\{\sum_{i=1}^m f_i(z^i; x^{-i}) + g(z) + c\|z - x\|^2\Big\} \quad (15)$$

and operators $T_i(y^{-i})$ such that

$$T_i(x^i; y^{-i}) = \operatorname*{argmin}_{z^i} \big\{f_i(z^i; y^{-i}) + g_i(z^i) + c\|z^i - x^i\|^2\big\},$$

where $y^{-i} \in \mathbb{R}^{n-n_i}$ is treated as a fixed parameter. Observe that we can characterize the operator $T(x)$ via the operators $T_i(x^i; x^{-i})$ as follows

$$T(x) = \big(T_1(x^1; x^{-1}), \ldots, T_m(x^m; x^{-m})\big).$$

We define the sets of fixed points for these operators as

$$\begin{aligned}
\operatorname{Fix} T &:= \{x \mid x = T(x)\},\\
\operatorname{Fix} T_i(y^{-i}) &:= \{x^i \mid x^i = T_i(x^i; y^{-i})\}, \quad i = 1, \ldots, m.
\end{aligned}$$

Note that, in the spirit of [24, §5], we treat T as a single valued function $T : \mathbb{R}^n \mapsto \mathbb{R}^n$ since the quadratic term in the right hand side of (15) ensures that the set of minimizers is always single-valued, with an identical comment applying to the operators $T_i(y^{-i})$.

We now show that the sets $\operatorname{argmin} h$ and $\operatorname{Fix} T$ coincide.

Proposition 4: If Assumption 1 holds, then

$$\operatorname{argmin} h = \operatorname{Fix} T.$$

Proof: The proof is based on [13, proofs of Propositions 1–3]. We first show that $\operatorname{argmin} h \subseteq \operatorname{Fix} T$. Fix any $x \in \operatorname{argmin} h$. If x minimizes h, then it is also a block-wise minimizer of h at x, i.e. for all $i = 1, \ldots, m$, we have

$$x^i \in \operatorname*{argmin}_{z^i} \big\{f_i(z^i; x^{-i}) + g_i(z^i)\big\}. \quad (16)$$

Since x^i minimizes both $\big\{f_i(\,\cdot\,;x^{-i})+g_i\big\}$ and $c\|(\,\cdot\,)-x^i\|^2$, it is also the unique minimizer of their sum, i.e.

$$x^i=\operatorname*{argmin}_{z^i}\big\{f_i(z^i;x^{-i})+g_i(z^i)+c\|z^i-x^i\|^2\big\},$$

implying that $x^i\in\operatorname{Fix}T_i(x^{-i})$, and thus $x=(x^1,\dots,x^m)$ is a fixed point of $T(x)=\big(T_1(x^1;x^{-1}),\dots,T_m(x^m;x^{-m})\big)$.

We now show that $\operatorname{Fix}T\subseteq\operatorname{argmin}h$. Let $x\in\operatorname{Fix}T$, and thus for all $i=1,\dots,m$, $x^i\in\operatorname{Fix}T_i(x^{-i})$, i.e.

$$x^i=\operatorname*{argmin}_{z^i}\big\{f_i(z^i;x^{-i})+g_i(z^i)+c\|z^i-x^i\|^2\big\}.$$

According to Proposition 3 the above condition means that for all $z^i\in\mathbb{R}^{n_i}$ we have

$$\langle\nabla f_i(x^i;x^{-i}),z^i-x^i\rangle+g_i'(x^i,z^i-x^i)\\+\langle(2c\underbrace{(x^i-x^i)}_{=0}),z^i-x^i\rangle\geq 0,$$

or equivalently for all $d^i\in\mathbb{R}^{n_i}$

$$\langle\nabla f_i(x^i;x^{-i}),d^i\rangle+g_i'(x^i,d^i)\geq 0,$$

which again by Proposition 3 implies that x^i is a minimizer of $\big\{f_i(\,\cdot\,;x^{-i})+g_i\big\}$. According to [25, Lemma 3.1] differentiability of f and component-wise separability of g imply that any $x=(x^1,\dots,x^m)$ for which (16) holds for all $i=1,\dots,m$, is also a minimizer of $\big\{f+g\big\}$, i.e., $x\in\operatorname{argmin}h$, thus concluding the proof. ∎

REFERENCES

[1] A. Beck and L. Tetruashvili, "On the convergence of block coordinate descent type methods," *SIAM Journal on Optimization*, vol. 23, no. 4, pp. 2037–2060, 2013.

[2] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*. Athena Scientific, 1997.

[3] Y. Xu and W. Yin, "A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion," *SIAM Journal on Imaging Sciences*, vol. 6, no. 3, pp. 1758–1789, 2013.

[4] S. Grammatico, F. Parise, M. Colombino, and J. Lygeros, "Decentralized convergence to Nash equilibria in constrained deterministic mean field control," *IEEE Transactions on Automatic Control*, vol. 61, no. 11, pp. 3315–3329, 2016.

[5] L. Deori, K. Margellos, and M. Prandini, "On decentralized convex optimization in a multi-agent setting with separable constraints and its application to optimal charging of electric vehicles," in *IEEE Conference on Decision and Control (CDC)*, 2016, pp. 6044–6049.

[6] D. Paccagnan, M. Kamgarpour, and J. Lygeros, "On aggregative and mean field games with applications to electricity markets," in *European Control Conference (ECC)*, 2016, pp. 196–201.

[7] M. Hong, X. Wang, M. Razaviyayn, and Z.-Q. Luo, "Iteration complexity analysis of block coordinate descent methods," *Mathematical Programming*, vol. 163, no. 1-2, pp. 85–114, 2017.

[8] I. Necoara and D. Clipici, "Parallel random coordinate descent method for composite minimization: Convergence analysis and error bounds," *SIAM Journal on Optimization*, vol. 26, no. 1, pp. 197–226, 2016.

[9] P. Richtárik and M. Takáč, "Parallel coordinate descent methods for big data optimization," *Mathematical Programming*, vol. 156, no. 1–2, pp. 433–484, 2016.

[10] D. P. Bertsekas, "Incremental proximal methods for large scale convex optimization," *Mathematical Programming*, vol. 129, no. 2, pp. 163–195, 2011.

[11] H. Attouch, J. Bolte, and B. Svaiter, "Convergence of descent methods for semi-algebraic and tame problems: Proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods," *Mathematical Programming*, vol. 137, no. 1-2, pp. 91–129, 2013.

[12] G. Cohen, "Optimization by decomposition and coordination: A unified approach," *IEEE Transactions on Automatic Control*, vol. 23, no. 2, pp. 222–232, 1978.

[13] L. Deori, K. Margellos, and M. Prandini, "Regularized Jacobi iteration for decentralized convex quadratic optimization with separable constraints," *IEEE Transactions on Control Systems Technology (To appear)*, 2018.

[14] J. Bolte, T. P. Nguyen, J. Peypouquet, and B. W. Suter, "From error bounds to the complexity of first-order descent methods for convex functions," *Mathematical Programming*, vol. 165, no. 2, pp. 471–507, 2017.

[15] H. Zhang, "The restricted strong convexity revisited: analysis of equivalence to error bound and quadratic growth," *Optimization Letters*, vol. 11, no. 4, pp. 817–833, 2017.

[16] D. Drusvyatskiy and A. S. Lewis, "Error bounds, quadratic growth, and linear convergence of proximal methods," *Mathematics of Operations Research*, vol. 43, no. 3, pp. 919–948, 2018.

[17] H. Attouch, J. Bolte, P. Redont, and A. Soubeyran, "Proximal alternating minimization and projection methods for nonconvex problems: an approach based on the Kurdyka-Łojasiewicz inequality," *Mathematics of Operations Research*, vol. 35, no. 2, pp. 438–457, 2010.

[18] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, 1st ed. Springer Publishing Company, Incorporated, 2011.

[19] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[20] I. Necoara, Y. Nesterov, and F. Glineur, "Linear convergence of first order methods for non-strongly convex optimization," *Mathematical Programming (To appear)*, 2018.

[21] Z.-Q. Luo and P. Tseng, "On the linear convergence of descent methods for convex essentially smooth minimization," *SIAM Journal on Control and Optimization*, vol. 30, no. 2, pp. 408–425, 1992.

[22] P.-W. Wang and C.-J. Lin, "Iteration complexity of feasible descent methods for convex optimization," *Journal of Machine Learning Research*, vol. 15, pp. 1523–1548, 2014.

[23] A. Beck and S. Shtern, "Linearly convergent away-step conditional gradient for non-strongly convex functions," *Mathematical Programming*, vol. 164, no. 1-2, pp. 1–27, 2017.

[24] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*. Springer-Verlag, 1998.

[25] P. Tseng, "Convergence of a block coordinate descent method for nondifferentiable minimization," *Journal of Optimization Theory and Applications*, vol. 109, no. 3, pp. 475–494, 2001.