Nonconvex Big-Data Optimization via System Theory: a Block-wise Incremental Gradient Method

Nonconvex Big-Data Optimization via System Theory: a Block-wise Incremental Gradient Method

Guido Carnevale, Ivano Notarnicola, Giuseppe Notarstefano

***Abstract*— In this paper, we propose and analyze Block-wise Incremental Gradient with Averaging (BIG-A), i.e., a novel optimization algorithm tailored for large-scale and big-data nonconvex optimization problems with a composite cost function. At each iteration, the algorithm uses a block of a single function gradient to properly update auxiliary variables providing a proxy of a descent direction. We interpret BIG-A as a dynamical system arising from the interconnection between a fast, time-varying scheme and a slow, time-invariant one. This interpretation allows us to prove the convergence properties by using system theory results relying on the LaSalle-Yoshizawa invariance principle and singular perturbations. The solution estimate sequence generated by BIG-A is shown to converge toward the set of stationary points of the problem, which is not assumed to be convex nor satisfying the Polyak-Łojasiewicz condition. If strong convexity is also assumed, linear convergence toward the unique optimal solution is established. Finally, numerical simulations confirm the theoretical findings.**

I. Introduction

A wide range of machine learning problems including, but not limited to, logistic regression, neural networks, multi-kernel learning, etc., can be modeled via optimization problems in the form

$$\min_{x\in\mathbb{R}^n} \sum_{s=1}^{S} f_s(x), \tag{1}$$

where $n \in \mathbb{N}$ is the size of the decision variable and each $f_s : \mathbb{R}^n \to \mathbb{R}$ represents a component of the overall cost function $f(x) := \sum_{s=1}^{S} f_s(x)$. In this paper, we focus on challenging instances of problem (1) in which we simultaneously deal with a very large number S and a very large size n of the decision variable. Moreover, general nonconvex problems are considered since they typically arise in many complex learning applications.

The literature review is organized into three main parts: (i) algorithms tailored for a large-scale problem (S large), (ii) block-wise methods to deal with big-data problems (n large), and (iii) analysis carried out in nonconvex settings.

When the number S and/or the size n are very large, a classical Gradient Descent (GD) method is inefficient since it would require the full cost gradient computation at every iteration of the algorithm. The key tool to lower the computational workload in a large-scale problem is based on the so-called Incremental First-order Oracle (IFO) approach [1]. Namely, at each iteration, only the gradient of one cost function (rather than the entire sum forming f) is evaluated and used. In a deterministic setting, the incremental gradient methods for differentiable (unconstrained) problems have a long tradition in IFO. The most notable application is the training of neural networks, where they are known as back-propagation methods. A concise survey of incremental gradient methods is [2]. In a stochastic setting, the Stochastic Gradient Descent (SGD), originally proposed in the early reference [3], has proven itself in practice as viable solution to address problem (1). However, a well-known drawback of SGD is that it achieves exact convergence at the expense of a diminishing step size, which deteriorates the algorithmic performance. In recent years, a large number of variations of stochastic-gradient-like algorithms have been proposed in the literature to improve the performance of the plain SGD. Such variations can be roughly (and, inevitably, non-exhaustively) grouped in the following classes: SAG/SAGA [4], [5], SDCA [6], MISO [7], SVRG/S2GD [8], [9], SARAH [10].

Considerable research efforts have been also made in developing big-data optimization methods based on block-wise updates, also known as (block-)coordinate updates. Namely, problem (1) is solved by iteratively optimizing only a portion of the entire decision variable at each iteration of the algorithm. Such an approach has been applied to convex optimization problems in [11], [12]. The extensions to a parallel computational paradigm are investigated in [13]. The majority of the works in the literature consider nonconvex instances of problem (1) satisfying the so-called gradient-dominated condition on f, see [14] for a definition. In [15], a fast incremental aggregated gradient method for smooth nonconvex (gradient-dominated) problems is investigated. A SVRG algorithm for the same class of problems is studied in [16] with extensions also to mini-batch variants. Recently, a cyclic block-coordinate variance reduction algorithm has been proposed in [17]. A globally convergent block-coordinate algorithm for nonconvex problems (enjoying Kurdyka-Łojasiewicz property) has been developed in [18]. Since it may be unrealistic in practical applications, we remove such a condition on the cost function f, so that most of the previous analyses do not apply.

In this paper, we propose and analyze Block-wise Incremental Gradient with Averaging (BIG-A), a novel algorithm tailored for big-data optimization problems. The peculiarity of BIG-A is that at each iteration, it only requires to compute a block of the gradient vector of a single function of the

The authors are with the Department of Electrical, Electronic and Information Engineering, Alma Mater Studiorum - Università di Bologna, Bologna, 40136, Italy. This work was supported in part by the Italian Ministry of Foreign Affairs and International Cooperation, grant number BR22GR01. The corresponding author is G. Carnevale guido.carnevale@unibo.it.

composite cost. Therefore, BIG-A requires low computational effort for execution, making it well-suited for big-data problems. These appealing features lead to a complex algorithm that we model as a time-varying, interconnected dynamical system. This system theory framework for the design and analysis represents one of the contributions of the paper along the lines of a recent trend in the literature. More in detail, the convergence analysis is performed by exploiting system theory tools for stability, i.e., the LaSalle-Yoshizawa invariance principle and singular perturbations. We prove that the solution estimates generated by BIG-A converge to the set of the stationary points of the optimization problem, which is not assumed to be convex nor satisfying the Polyak-Łojasiewicz (PL) condition. If strong convexity is considered, the analysis of the general case guarantees the existence of a globally exponentially stable equilibrium point for BIG-A, corresponding to the unique optimal solution.

As a side contribution of the paper, we also provide Theorem A.3. To the best of our knowledge, it improves the state of the art in two ways: it extends [19, Thm. 1] (focused only on time-invariant systems) and [20, Thm. II] (focused only on singularly-perturbed system with exponential stability properties) to time-varying, singularly perturbed discrete-time systems in which the slow-dynamics does not have an exponentially stable equilibrium. To prove Theorem A.3, we also derived Theorem A.2, which represents a LaSalle-Yoshizawa invariance principle for generic discrete-time systems (a counter-part of the continuous time result in [21, Thm. 4.7]).

The paper is organized as follows. Section II presents BIG-A with its convergence properties while Section III sketches the proof of the convergence result. Section IV provides numerical simulations confirming the effectiveness of the approach. Finally, the appendices report useful results about LaSalle-Yoshizawa stability and singular perturbations for time-varying discrete-time systems.

Notation: The identity matrix of order n is I_n. The all-zero vector in $\mathbb{R}^n$ is denoted as 0_n, while the all-one vector in $\mathbb{R}^N$ is denoted as 1_N. The vertical concatenation of column vectors $v_1, \dots, v_N$ is $\mathrm{col}(v_1, \dots, v_N)$. A diagonal matrix with diagonal entries $d_1, \dots, d_N \in \mathbb{R}$ is denoted as $\mathrm{diag}(d_1, \dots, d_N)$. The symbol $\otimes$ denotes the Kronecker product. Given $r > 0$ and $x \in \mathbb{R}^n$, we use $\mathcal{B}_r(x)$ to denote the ball of radius $r > 0$ centered in x, namely $\mathcal{B}_r(x) := \{y \in \mathbb{R}^n \mid \|y\| \leq r\}$. Given $f : \mathbb{R}^n \to \mathbb{R}$ and $x = \mathrm{col}(x_1, \dots, x_N) \in \mathbb{R}^n$ with $x_i \in \mathbb{R}^{n_i}$ for all $i \in \{1, \dots, N\}$, the symbol $\nabla_i f(x) \in \mathbb{R}^{n_i}$ denotes the i-th block of the full gradient $\nabla f(x)$, corresponding to the block variable x_i of x.

II. BIG-A Algorithm: Description and Convergence Properties

Consider problem (1), let the decision variable x be partitioned in $B \in \mathbb{N}$ blocks as $x := \mathrm{col}(x_1, \dots, x_B)$, with each block $x_b \in \mathbb{R}^{n_b}$, $n_b \in \mathbb{N}$, for all $b \in \{1, \dots, B\}$. Clearly, $\sum_{b=1}^{B} n_b = n$. Then, we denote as $\mathrm{x}^t \in \mathbb{R}^n$, the estimate of the optimal solution of (1) at iteration $t \in \mathbb{N}$. The value of x^t is updated by resorting to auxiliary variables $\mathrm{d}_{b,s}^t \in \mathbb{R}^{n_b}$, with $b \in \{1, \dots, B\}$ and $s \in \{1, \dots, S\}$. Each $\mathrm{d}_{b,s}^t$ stores the gradient portion $\nabla_b f_s(\mathrm{x}^t)$, by elaborating its most updated value whenever it is chosen during the algorithmic evolution. Formally, $\mathrm{d}_{b,s}^t$ is arbitrarily initialized and updated, for all $b \in \{1, \dots, B\}$ and $s \in \{1, \dots, S\}$, according to

$$\mathrm{d}_{b,s}^{t+1} = \begin{cases} \nabla_b f_s(\mathrm{x}^t) - \mathrm{d}_{b,s}^t & \text{if } c_{b,s}^t = 1 \\ \mathrm{d}_{b,s}^t & \text{if } c_{b,s}^t = 0, \end{cases} \tag{2}$$

where $c_{b,s}^t \in \{0, 1\}$ is a (possibly unknown) deterministic binary signal that is used to model whether $\nabla_b f_s(\mathrm{x}^t)$ is effectively chosen (and evaluated) at iteration t or not. Concurrently, the auxiliary variables $\mathrm{d}_{b,s}^t$ are used to improve the solution estimate x^t. By partitioning the solution estimate x^t in blocks as $\mathrm{x}^t := \mathrm{col}(\mathrm{x}_1^t, \dots, \mathrm{x}_B^t)$, with $\mathrm{x}_b^t \in \mathbb{R}^{n_b}$ for all $b \in \{1, \dots, B\}$, each x_b^t is arbitrarily initialized and updated in a gradient-like fashion as

$$\mathrm{x}_b^{t+1} = \mathrm{x}_b^t - \gamma \sum_{s=1}^{S} \mathrm{d}_{b,s}^t, \tag{3}$$

for all $b \in \{1, \dots, B\}$, where $\gamma > 0$ is the conventional step size. Let $\mathrm{d}_b := \mathrm{col}(\mathrm{d}_{b,1}, \dots, \mathrm{d}_{b,S}) \in \mathbb{R}^{Sn_b}$,

$$\mathcal{C}_b^t := \mathrm{diag}(c_{b,1}^t, \dots, c_{b,S}^t) \otimes I_{n_b} \in \mathbb{R}^{Sn_b \times Sn_b},$$

and define $G_b : \mathbb{R}^n \to \mathbb{R}^{Sn_b}$ as

$$G_b(x) = \begin{bmatrix} \nabla_b f_1(x) \\ \vdots \\ \nabla_b f_S(x) \end{bmatrix}.$$

Then, by collecting (2)-(3), and embedding the just introduced notation, the algorithmic updates can be rephrased as

$$\mathrm{x}_b^{t+1} = \mathrm{x}_b^t - \gamma \sum_{s=1}^{S} \mathrm{d}_{b,s}^t \tag{4a}$$

$$\mathrm{d}_b^{t+1} = \mathrm{d}_b^t - \mathcal{C}_b^t (G_b(\mathrm{x}^t) - \mathrm{d}_b^t), \tag{4b}$$

for all $b \in \{1, \dots, B\}$.

We formally present the assumptions for our setup.

Assumption 2.1: For all $s \in \{1, \dots, M\}$, the function f_s has L-Lipschitz continuous gradients for some $L > 0$. Moreover, the whole function f is radially unbounded. ■

We stress that Assumption 2.1 does not impose that f is convex nor that it satisfies the PL inequality. Also, the radial unboundedness (or coercivity) of f ensures the existence of $f^\star := \min_{x \in \mathbb{R}^n} f(x)$, without necessarily enforcing the uniqueness of either minimizers or stationary points. By virtue of such a (weak) requirement, the convergence properties of (4) will be provided in terms of convergence

toward the set $X \subset \mathbb{R}^n$ of stationary points of f, namely

$$X := \{x \in \mathbb{R}^n \mid \nabla f(x) = 0\}. \tag{5}$$

We point out that the exact knowledge of the $c_{b,s}^t$ is not required, provided that it satisfies the following condition.

Assumption 2.2: For all $b \in \{1, \dots, B\}$ and $s \in \{1, \dots, S\}$, there exists a finite $T_{b,s} > 0$ such that $c_{b,s}^t = 1$ at least once in the interval $[t_0, t_0 + T_{b,s}]$ for all $t_0 \geq 0$. ■

Assumption 2.2 imposes an "essentially cyclic" rule on the selection of the gradients $\nabla_b f_s$, representing the weakest requirement that can reasonably be considered.

In order to state the convergence result for (4), we stack all the variables $\mathrm{d}_{b,s}$ as $\mathrm{d} := \mathrm{col}(\mathrm{d}_{1,1}, \dots, \mathrm{d}_{Sn}) \in \mathbb{R}^{Sn}$ and introduce the operator $G : \mathbb{R}^n \to \mathbb{R}^{Sn}$ defined as

$$G(x) := \begin{bmatrix} \nabla f_1(x) \\ \vdots \\ \nabla f_S(x) \end{bmatrix}, \tag{6}$$

for all $x \in \mathbb{R}^{Sn}$. We are ready to provide the convergence properties of BIG-A.

Theorem 2.3: Let Assumptions 2.1 and 2.2 hold. Then, there exists $\bar{\gamma} > 0$ such that, for any $\gamma \in (0, \bar{\gamma})$ and for all $(\mathrm{x}^0, \mathrm{d}^0) \in \mathbb{R}^n \times \mathbb{R}^{Sn}$, the trajectories $\{\mathrm{x}^t, \mathrm{d}^t\}_{t \in \mathbb{N}}$ of (4) satisfy

$$\lim_{t \to \infty} \inf_{\xi \in X} \left\| \begin{bmatrix} \mathrm{x}^t - \xi \\ \mathrm{d}^t - G(\mathrm{x}^t) \end{bmatrix} \right\| = 0, \tag{7}$$

with X defined in (5). ■

A sketch of the proof is postponed to the dedicated Section III by employing a system-theoretic perspective to system (4). The whole proof will be given in a forthcoming document.

When strong convexity of the cost f is further assumed, the convergence properties of BIG-A specializes as follows.

Corollary 2.4: Let Assumptions 2.1 and 2.2 hold. Moreover, let f be μ-strongly convex for some $\mu > 0$. Then, there exist $\bar{\gamma} > 0$ such that, for any $\gamma \in (0, \bar{\gamma})$ and for all $(\mathrm{x}^0, \mathrm{d}^0) \in \mathbb{R}^n \times \mathbb{R}^{Sn}$, the trajectories $\{\mathrm{x}^t, \mathrm{d}^t\}_{t \in \mathbb{N}}$ of (4) satisfy

$$\left\| \begin{bmatrix} \mathrm{x}^t - x^\star \\ \mathrm{d}^t - G(\mathrm{x}^t) \end{bmatrix} \right\| \leq a_1 \left\| \begin{bmatrix} \mathrm{x}^0 - x^\star \\ \mathrm{d}^0 - G(\mathrm{x}^0) \end{bmatrix} \right\| \exp(-a_2 t), \tag{8}$$

for some $a_1, a_2 > 0$. ■

The proof of Corollary 2.4 will be given in a forthcoming document.

We point out that in the case of strongly convex cost, the result in (8) establishes the linear convergence rate toward the (unique) optimal solution $x^\star \in \mathbb{R}^n$ of the problem.

III. Sketch of Proof of Theorem 2.3

This section develops the proof of Theorem 2.3 in a constructive way combining singular perturbations and the LaSalle-Yoshizawa invariance principle. Firstly, in Section III-A, we interpret system (4) as a singularly perturbed system, i.e., as an instance of the interconnected system (20) analyzed in Theorem A.3. As is customary in this context, we consider two auxiliary systems termed (i) the boundary-layer system (cf. system (22) in Theorem A.3) and (ii) the reduced system (cf. system (21) in Theorem A.3). Their stability properties are separately characterized in Sections III-B and III-C, respectively. Finally, in Section III-D, we exploit the previous results to invoke the LaSalle-Yoshizawa invariance principle for time-varying discrete-time singularly perturbed systems (cf. Theorem A.3 in Appendix B) to transfer the stability result back to the original system, completing the proof.

A. BIG-A as a Singularly Perturbed System

To begin with, let us introduce a compact, equivalent reformulation of (4) given by

$$\mathrm{x}^{t+1} = \mathrm{x}^t - \gamma \mathbf{1}_{S,n}^\top \mathrm{d}^t \tag{9a}$$

$$\mathrm{d}^{t+1} = \mathrm{d}^t + \mathcal{C}^t(G(\mathrm{x}^t) - \mathrm{d}^t), \tag{9b}$$

with $\mathbf{1}_{S,n} := 1_S \otimes I_n$ and $\mathcal{C}^t := \mathrm{diag}(\mathcal{C}_1^t, \dots, \mathcal{C}_B^t)$ for all $t \in \mathbb{N}$.

System (9) has the structure of a singularly-perturbed system, i.e., it is the feedback interconnection of a *fast dynamics* (9b) with a *slow dynamics* (9a) (cf. system (20) in Theorem A.3). As it will become clearer in the next, we decided to adopt a singularly-perturbed perspective because (i) the fast scheme (9b) has an equilibrium $\mathrm{d}_{\mathrm{eq}}(\mathrm{x}^t) \in \mathbb{R}^{Sn}$ for every fixed value of the slow state, and (ii) variations of the slow scheme (9a) can be made arbitrarily small by reducing the magnitude of γ. Building upon this observation, we analyze the interconnected system (9) by independently characterizing the stability properties of the boundary-layer system and the reduced system.

B. Boundary-Layer System

The boundary-layer system associated to (9) is obtained from the fast dynamics (9b) by replacing, for all $t \in \mathbb{N}$, the slow state x^t with an arbitrary (fixed) value $\mathrm{x} \in \mathbb{R}^n$. Moreover, error coordinates with respect to the parametrized equilibrium point $G(\mathrm{x})$ must be considered. Namely, let $\tilde{\mathrm{d}}_{\mathrm{bl}} \in \mathbb{R}^{Sn}$ be defined as $\tilde{\mathrm{d}}_{\mathrm{bl}} := \mathrm{d} - G(\mathrm{x})$. Then, its dynamics is

$$\tilde{\mathrm{d}}_{\mathrm{bl}}^{t+1} = \tilde{\mathrm{d}}_{\mathrm{bl}}^t - \mathcal{C}^t \tilde{\mathrm{d}}_{\mathrm{bl}}^t. \tag{10}$$

The next lemma provides a (time-varying) Lyapunov function to assert the globally exponentially stability of the origin for (the time-varying) system (10).

Lemma 3.1: Consider the boundary-layer system (10). Then, there exist a continuous function $W : \mathbb{R}^{Sn} \times \mathbb{N} \to \mathbb{R}$ and some constants $b_1, b_2, b_3, b_4 > 0$ such that it holds

$$b_1 \|\tilde{\mathrm{d}}\|^2 \leq W(\tilde{\mathrm{d}}, t) \leq b_2 \|\tilde{\mathrm{d}}\|^2 \tag{11a}$$

$$W((I_{Sn} - \gamma \mathcal{C}^t)\tilde{\mathrm{d}}, t+1) - W(\tilde{\mathrm{d}}, t) \leq -b_2 \|\tilde{\mathrm{d}}\|^2 \tag{11b}$$

$$W(\tilde{\mathrm{d}}, t) - W(\tilde{\mathrm{d}}', t) \leq b_4 \|\tilde{\mathrm{d}} - \tilde{\mathrm{d}}'\| \left(\|\tilde{\mathrm{d}}\| + \|\tilde{\mathrm{d}}'\| \right), \tag{11c}$$

for all $\tilde{\mathrm{d}}, \tilde{\mathrm{d}}' \in \mathbb{R}^{Sn}$ and $t \in \mathbb{N}$. ■

The proof of Lemma 3.1 will be given in a forthcoming document.

C. Reduced System

In the next step, we investigate the so-called reduced system associated to (9). Namely, a system obtained from the slow dynamics (9a) by considering the fast state always into its parametrized equilibrium, namely $\mathrm{d}^t := G(\mathrm{x}^t)$ for all $t \in \mathbb{N}$. Since $\mathbf{1}_{S,n}^\top G(\mathrm{x}) = \nabla f(\mathrm{x})$, the reduced system associated to (9) simplifies to a plain gradient method applied to problem (1), expressed as follows

$$\mathrm{x}_{\mathrm{rs}}^{t+1} = \mathrm{x}_{\mathrm{rs}}^t - \gamma \nabla f(\mathrm{x}_{\mathrm{rs}}^t), \tag{12}$$

with $\mathrm{x}_{\mathrm{rs}}^t \in \mathbb{R}^n$.

The next lemma characterizes the equilibrium stability for system (12).

Lemma 3.2: Consider the reduced system (12). Then, there exist $\bar{\gamma}_1, c_1, c_2, c_3 > 0$ and a continuous, radially unbounded function $U : \mathbb{R}^n \to \mathbb{R}$ such that for all $\gamma \in (0, \bar{\gamma}_1)$, the following hold

$$U(\mathrm{x} - \gamma \nabla f(\mathrm{x})) - U(\mathrm{x}) \leq -\gamma c_1 \left\|\nabla f(\mathrm{x})\right\|^2 \tag{13a}$$

$$U(\mathrm{x} + \mathrm{x}') - U(\mathrm{x} + \mathrm{x}'') \leq c_2 \left\|\nabla f(\mathrm{x})\right\| \left\|\mathrm{x}' - \mathrm{x}''\right\| + c_3 \left(\left\|\mathrm{x}'\right\|^2 + \left\|\mathrm{x}''\right\|^2\right), \tag{13b}$$

for all $\mathrm{x}, \mathrm{x}', \mathrm{x}'' \in \mathbb{R}^n$. ■

The proof of Lemma 3.2 will be given in a forthcoming document.

D. Reverting to the original system (9)

We are ready to complete the proof of Theorem 2.3 by reverting the stability results regarding the boundary-layer system (10) (cf. Lemma 3.1) and the reduced system (12) (cf. Lemma 3.2) to the original system (9), namely to BIG-A.

We first notice that being each gradient ∇f_s Lipschitz continuous (cf. Assumption 2.1), it turns out that also the vector field describing the dynamics in (9) and the operator $G(\mathrm{x})$ in (6) are Lipschitz continuous.

Then, in light of Lemma 3.1 and Lemma 3.2, we guarantee the existence of $\bar{\gamma}_1 > 0$ and the functions W and U that, for all $\gamma \in (0, \bar{\gamma}_1)$, satisfy the conditions requested to invoke Theorem A.3, i.e., the conditions in (23) and (24). Therefore, Theorem A.3 applies with the identification $\mathrm{x} = \chi$ and $\mathrm{d} = \zeta$, and readily gives the bound in (7).

IV. Numerical Simulations

In this section, we test BIG-A via numerical comparison with Gradient Descent on a logistic regression scenario. We aim to train a linear classifier for a set of points in a given feature space. We consider S points $p_1, \ldots, p_S \in \mathbb{R}^{n-1}$ with binary labels $l_s \in \{-1, 1\}$ for all $s \in \{1, \ldots, S\}$. The problem consists of building a linear classification model from these points solving the minimization problem described by

$$\min_{w \in \mathbb{R}^{n-1}, w_0 \in \mathbb{R}} \sum_{s=1}^{S} \log\left(1 + e^{-l_s(w^\top p_s + w_0)}\right) + \frac{C}{2} \left\| \begin{bmatrix} w \\ w_0 \end{bmatrix} \right\|^2,$$

where $C > 0$ is the so-called regularization parameter. We set $S = 10$, $n = 3$, and partioned x in 2 blocks with $x_1 \in \mathbb{R}^2$ and $x_2 \in \mathbb{R}$. We randomly generate the points and labels of the dataset and the deterministic sequences $\{c_{b,s}^t\}$ for all $s \in \{1, \ldots, 10\}$ and $b \in \{1, 2\}$ (by ensuring that Assumption 2.2 is satisfied). We empirically tune the parameters of BIG-A and Gradient Descent by choosing $\gamma = 0.05$. Fig. 1 shows the comparison among BIG-A and Gradient Descent in terms of $\|\mathrm{x}^t - x^\star\|$, where $x^\star$ is the unique stationary point of the logistic regression problem. As

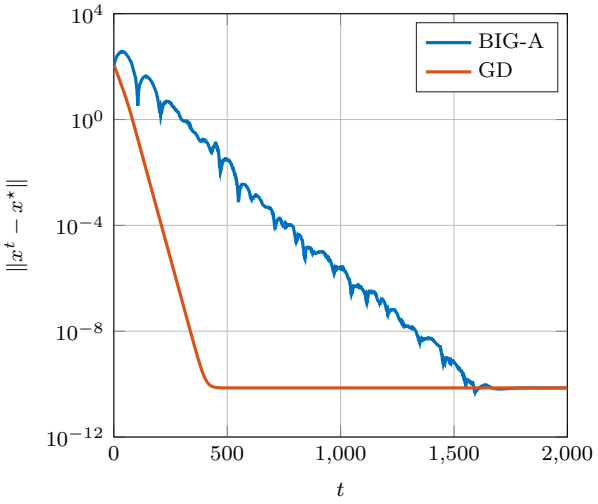

Fig. 1. Comparison among Block-wise Incremental Gradient with Averaging (BIG-A) and Gradient Descent (GD) in terms of $\|\mathrm{x}^t - x^\star\|$.

predicted by Theorem 2.3, Fig. 1 confirms the convergence properties of BIG-A. More in detail, Fig. 1 shows that BIG-A has convergence rate comparable to the one of Gradient Descent.

V. Conclusions

In this paper, we presented and analyzed BIG-A to solve big-data nonconvex optimization problems. The proposed algorithm is computationally efficient since it requires the evaluation of only one block of only one component of the cost function gradient per iteration. The analysis relied on the reformulation of the optimization algorithm as a feedback interconnected dynamical systems, which lends itself to a system-theoretic analysis based on LaSalle-Yoshizawa invariance principle and singular perturbations.

Appendix

A. LaSalle-Yoshizawa invariance principle for time-varying discrete-time systems

Next, we present a LaSalle-Yoshizawa invariance principle for time-varying discrete-time systems, which is a counter-

part of the continuous-time result in [21, Thm. 4.7]. As a preliminary step, we also provide the following instrumental theorem, which ensures partial stability for time-invariant discrete-time systems, extending [21, Thm. 13.10], by relaxing the positive-definiteness requirement on the Lyapunov function.

Theorem A.1: Consider the following time-invariant dynamical system

$$\mathrm{x}_1^{t+1} = \psi_1(\mathrm{x}_1^t, \mathrm{x}_2^t) \tag{14a}$$
$$\mathrm{x}_2^{t+1} = \psi_2(\mathrm{x}_1^t, \mathrm{x}_2^t), \tag{14b}$$

with states $\mathrm{x}_1^t \in \mathbb{R}^{n_1}$, $\mathrm{x}_2^t \in \mathbb{R}^{n_2}$ and continuous vector fields $\psi_1 : \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \to \mathbb{R}^{n_1}$, $\psi_2 : \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \to \mathbb{R}^{n_2}$, for some given initial conditions x_1^0 and x_2^0. Further, assume there exist a continuous, radially unbounded function $V : \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \to \mathbb{R}$, a nonnegative continuous function $\varphi : \mathbb{R}^{n_1} \to \mathbb{R}$ such that

$$V(\psi_1(\mathrm{x}_1, \mathrm{x}_2), \psi_2(\mathrm{x}_1, \mathrm{x}_2)) - V(\mathrm{x}_1, \mathrm{x}_2) \leq -\varphi(\mathrm{x}_1), \tag{15}$$

for all $\mathrm{x}_1 \in \mathbb{R}^{n_1}$ and $\mathrm{x}_2 \in \mathbb{R}^{n_2}$. Then, the trajectories $\{\mathrm{x}_1^t, \mathrm{x}_2^t\}_{t\in\mathbb{N}}$ of system (14) satisfy

$$\lim_{t\to\infty} \inf_{\xi\in\mathcal{M}} \left\|\mathrm{x}_1^t - \xi\right\| = 0, \tag{16}$$

for all $(\mathrm{x}_1^0, \mathrm{x}_2^0)$, where $\mathcal{M} := \{\mathrm{x}_1 \in \mathbb{R}^{n_1} \mid \varphi(\mathrm{x}_1) = 0\}$. ■

Proof: Being V radially unbounded, it is bounded from below by some $c \in \mathbb{R}$ and its level sets are compact, i.e., for any $\ell > 0$, the set $\Omega_\ell := \{(\mathrm{x}_1, \mathrm{x}_2) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \mid V(\mathrm{x}_1, \mathrm{x}_2) \leq \ell\}$ is compact. Moreover, in light of (15), any level set is also invariant for system (14). Since $V(\mathrm{x}_1^t, \mathrm{x}_2^t) \geq c$ is monotonically nonincreasing with respect to $t \in \mathbb{N}$, it follows that $\lim_{t\to\infty} V(\mathrm{x}_1^t, \mathrm{x}_2^t)$ exists and is finite (see, e.g., [21, Th. 2.10]). Note that, for all $t \geq 0$, it holds

$$\begin{aligned}\sum_{k=0}^{t-1} \varphi(\mathrm{x}_1^k) &= -\sum_{k=0}^{t-1} V(\psi_1(\mathrm{x}_1^k, \mathrm{x}_2^k), \psi_2(\mathrm{x}_1^k, \mathrm{x}_2^k)) - V(\mathrm{x}_1^k, \mathrm{x}_2^k) \\ &= V(\mathrm{x}_1^0, \mathrm{x}_2^0) - V(\mathrm{x}_1^t, \mathrm{x}_2^t),\end{aligned}$$

which is finite and, thus, allows us to conclude that $\lim_{t\to\infty} \varphi(\mathrm{x}_1^t)$ exists and is finite. Hence, by [21, Lemma 3], it holds $\lim_{t\to\infty} \varphi(\mathrm{x}_1^t) = 0$ and the proof follows. ■

With this result at hand, we are ready to state a LaSalle-Yoshizawa invariance principle for time-varying discrete-time systems.

Theorem A.2: Consider the following time-varying dynamical system

$$\mathrm{x}^{t+1} = \psi(\mathrm{x}^t, t), \tag{17}$$

with state $\mathrm{x}^t \in \mathbb{R}^n$ and continuous vector field $\psi : \mathbb{R}^n \times \mathbb{N} \to \mathbb{R}^n$, for a given initial condition x^0. Further, assume there exist a continuous, radially unbounded function $V : \mathbb{R}^n \times \mathbb{N} \to \mathbb{R}$, a nonnegative continuous function $\varphi : \mathbb{R}^n \to \mathbb{R}$ such that

$$V(\psi(\mathrm{x}, t), t+1) - V(\mathrm{x}, t) \leq -\varphi(\mathrm{x}), \tag{18}$$

for all $\mathrm{x} \in \mathbb{R}^n$ and $t \in \mathbb{N}$. Then, the trajectories $\{\mathrm{x}^t\}_{t\in\mathbb{N}}$ of system (17) satisfy

$$\lim_{t\to\infty} \inf_{\xi\in\mathcal{M}} \left\|\mathrm{x}^t - \xi\right\| = 0, \tag{19}$$

for all $\mathrm{x}^0 \in \mathbb{R}^n$, where $\mathcal{M} := \{\mathrm{x} \in \mathbb{R}^n \mid \varphi(\mathrm{x}) = 0\}$.

Proof: A direct application of Theorem A.1 with $n_1 = n$ and $n_2 = 1$, along with the identification $\mathrm{x}_1^t = \mathrm{x}^t$ and $\mathrm{x}_2^t = t$, and defining the functions as $\psi_1 = \psi$, $\psi_2(\mathrm{x}_1^t, \mathrm{x}_2^t) = \mathrm{x}_2^t + 1 = t + 1$, and $V = V$, leads to the sought proof. ■

B. Singular Perturbations and LaSalle-Yoshizawa Invariance Principle for time-varying discrete-time Systems

The next theorem extends [19, Thm. 1] and [20, Thm. II] providing a stability result for time-varying, singularly perturbed discrete-time systems in which the slow-dynamics does not have an exponentially stable equilibrium.

Theorem A.3: Consider the following time-varying dynamical system

$$\chi^{t+1} = \chi^t + \gamma\phi(\chi^t, \zeta^t) \tag{20a}$$
$$\zeta^{t+1} = g(\zeta^t, \chi^t, t), \tag{20b}$$

with states $\chi^t \in \mathbb{R}^n$, $\zeta^t \in \mathbb{R}^m$, vector fields $\phi : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$, $g : \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{N} \to \mathbb{R}^m$, and a tunable coefficient $\gamma > 0$. Assume that ϕ is Lipschitz continuous and g is Lipschitz continuous in (χ, ζ) uniformly in t, with parameters $L_\phi > 0$ and $L_g > 0$, respectively. Also, assume that there exists a Lipschitz continuous function, with parameter $L_h > 0$, $h : \mathbb{R}^n \to \mathbb{R}^m$ such that $g(h(\chi), \chi, t) = h(\chi)$ for all $\chi \in \mathbb{R}^n$ and $t \in \mathbb{N}$. Let us introduce the so-called reduced system

$$\chi_{\mathrm{rs}}^{t+1} = \chi_{\mathrm{rs}}^t + \gamma\phi(\chi_{\mathrm{rs}}^t, h(\chi_{\mathrm{rs}}^t)) \tag{21}$$

with state $\chi_{\mathrm{bl}}^t \in \mathbb{R}^n$, and the so-called boundary-layer system

$$\tilde{\zeta}_{\mathrm{bl}}^{t+1} = g(\tilde{\zeta}_{\mathrm{bl}}^t + h(\chi), \chi, t) - h(\chi) \tag{22}$$

with state $\tilde{\zeta}_{\mathrm{bl}}^t \in \mathbb{R}^m$ and $\chi \in \mathbb{R}^n$. Assume that there exists a continuous function $W : \mathbb{R}^m \times \mathbb{N} \to \mathbb{R}$ such that

$$b_1\left\|\tilde{\zeta}\right\|^2 \leq W(\tilde{\zeta}, t) \leq b_2\left\|\tilde{\zeta}\right\|^2 \tag{23a}$$
$$W(g(\tilde{\zeta} + h(\chi), \chi, t) - h(\chi), t+1) - W(\tilde{\zeta}, t) \leq -b_3\left\|\tilde{\zeta}\right\|^2 \tag{23b}$$
$$|W(\tilde{\zeta}, t) - W(\tilde{\zeta}', t)| \leq b_4\left\|\tilde{\zeta} - \tilde{\zeta}'\right\| \left(\left\|\tilde{\zeta}\right\| + \left\|\tilde{\zeta}'\right\|\right), \tag{23c}$$

for all $\tilde{\zeta}, \tilde{\zeta}' \in \mathbb{R}^m$, $\chi \in \mathbb{R}^n$, $t \in \mathbb{N}$, and some $b_1, b_2, b_3, b_4 > 0$. Further, assume there exist $\bar{\gamma}_1 > 0$ and a continuous, radially unbounded function $U : \mathbb{R}^n \to \mathbb{R}$, such that for all

$\gamma \in (0, \bar{\gamma}_1)$, it holds

$$U(\chi + \gamma\phi(\chi, h(\chi))) - U(\chi) \leq -\gamma c_1 \left\|\phi(\chi, h(\chi))\right\|^2 \quad (24a)$$

$$U(\chi + \chi') - U(\chi + \chi'') \leq c_2 \left\|\phi(\chi, h(\chi))\right\| \|\chi' - \chi''\| + c_3 \left(\|\chi'\|^2 + \|\chi''\|^2 \right), \quad (24b)$$

for all $\chi, \chi', \chi'' \in \mathbb{R}^n$, and some $c_1, c_2, c_3 > 0$. Then, there exists $\bar{\gamma} \in (0, \bar{\gamma}_1)$ such that, for all $\gamma \in (0, \bar{\gamma})$, the trajectories $\{\chi^t, \zeta^t\}_{t \in \mathbb{N}}$ of system (20) satisfies

$$\lim_{t \to \infty} \inf_{\xi \in \mathcal{M}} \left\| \begin{bmatrix} \chi^t \\ \zeta^t \end{bmatrix} - \begin{bmatrix} \xi \\ h(\chi^t) \end{bmatrix} \right\| = 0,$$

for all (χ^0, ζ^0), where $\mathcal{M} := \{\chi \in \mathbb{R}^n \mid \phi(\chi, h(\chi)) = 0\}$.■
A sketch of proof of Theorem A.3 is given in the next section. The whole proof will be given in a forthcoming document.

C. Sketch of proof of Theorem A.3

We start by defining $\tilde{\zeta}^t := \zeta^t - h(\chi^t)$, and rewrite (20) as

$$\chi^{t+1} = \chi^t + \gamma\phi(\chi^t, \tilde{\zeta}^t + h(\chi^t)) \quad (25a)$$

$$\tilde{\zeta}^{t+1} = g(\tilde{\zeta}^t + h(\chi^t), \chi^t, t) - h(\chi^{t+1}). \quad (25b)$$

Firstly, pick the function U satisfying (24). By using the conditions in (24), it is possible to bound its increment along trajectories of system (25a), namely $\Delta U(\chi^t) := U(\chi^{t+1}) - U(\chi^t)$, as

$$\begin{aligned} &\Delta U(\chi^t) \\ &\leq -\gamma c_1 \left\|\phi(\chi^t, h(\chi^t))\right\|^2 + \gamma c_2 L_\phi \left\|\phi(\chi^t, h(\chi^t))\right\| \left\|\tilde{\zeta}^t\right\| \\ &\quad + \gamma^2 c_3 2 \left\|\phi(\chi^t, h(\chi^t))\right\|^2 + \gamma^2 c_3 L_\phi^2 \left\|\tilde{\zeta}^t\right\|^2 \\ &\quad + \gamma^2 c_3 2 L_\phi \left\|\tilde{\zeta}^t\right\| \left\|\phi(\chi^t, h(\chi^t))\right\|. \end{aligned} \quad (26)$$

Secondly, pick W satisfying (23). By evaluating the increment $\Delta W(\tilde{\zeta}^t, t) := W(\tilde{\zeta}^{t+1}, t+1) - W(\tilde{\zeta}^t, t)$ along trajectories of system (25b), it is possible to show that

$$\begin{aligned} \Delta W(\tilde{\zeta}^t, t) \leq &-b_3 \left\|\tilde{\zeta}^t\right\|^2 + \gamma^2 b_4 L_3^2 L_\phi^2 \left\|\tilde{\zeta}^t\right\|^2 \\ &+ \gamma^2 b_4 L_3^2 \left\|\phi(\chi^t, h(\chi^t))\right\|^2 \\ &+ \gamma^2 b_4 2 L_3^2 L_\phi \left\|\tilde{\zeta}^t\right\| \left\|\phi(\chi^t, h(\chi^t))\right\| \\ &+ \gamma b_4 2 L_3 L_g L_\phi \left\|\tilde{\zeta}^t\right\|^2 \\ &+ \gamma b_4 2 L_3 L_g \left\|\phi(\chi^t, h(\chi^t))\right\| \left\|\tilde{\zeta}^t\right\|. \end{aligned} \quad (27)$$

Finally, let $V : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{N} \to \mathbb{R}$ be defined as

$$V(\chi^t, \tilde{\zeta}^t, t) := U(\chi^t) + W(\tilde{\zeta}^t, t).$$

By exploiting the bounds (26) and (27), it is possible to show that the increment of V along trajectories of system (25) is nonnegative for sufficiently small values of γ.

By evaluating its increment along the trajectories of system (25) and exploiting the bounds (26) and (27), it is possible to show the nonnegativeness of the increment of V for sufficiently small values of γ. Finally, with this result at hand, the proof follows by applying Theorem A.2.

REFERENCES

[1] A. Agarwal and L. Bottou, "A lower bound for the optimization of finite sums," in *International conference on machine learning*, pp. 78–86, PMLR, 2015.

[2] D. P. Bertsekas *et al.*, "Incremental gradient, subgradient, and proximal methods for convex optimization: A survey," *Optimization for Machine Learning*, vol. 2010, no. 1-38, p. 3, 2011.

[3] H. Robbins and S. Monro, "A stochastic approximation method," *The annals of mathematical statistics*, pp. 400–407, 1951.

[4] A. Defazio, F. Bach, and S. Lacoste-Julien, "SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives," *Advances in neural information processing systems*, vol. 27, 2014.

[5] M. Schmidt, N. Le Roux, and F. Bach, "Minimizing finite sums with the stochastic average gradient," *Mathematical Programming*, vol. 162, pp. 83–112, 2017.

[6] S. Shalev-Shwartz and T. Zhang, "Stochastic dual coordinate ascent methods for regularized loss minimization," *Journal of Machine Learning Research*, vol. 14, no. 1, 2013.

[7] J. Mairal, "Optimization with first-order surrogate functions," in *International Conference on Machine Learning*, pp. 783–791, PMLR, 2013.

[8] R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction," *Advances in neural information processing systems*, vol. 26, 2013.

[9] J. Konečnỳ and P. Richtárik, "Semi-stochastic gradient descent methods," *Frontiers in Applied Mathematics and Statistics*, vol. 3, p. 9, 2017.

[10] L. M. Nguyen, J. Liu, K. Scheinberg, and M. Takáč, "SARAH: A novel method for machine learning problems using stochastic recursive gradient," in *International conference on machine learning*, pp. 2613–2621, PMLR, 2017.

[11] Y. Nesterov, "Efficiency of coordinate descent methods on huge-scale optimization problems," *SIAM Journal on Optimization*, vol. 22, no. 2, pp. 341–362, 2012.

[12] S. J. Wright, "Coordinate descent algorithms," *Mathematical programming*, vol. 151, no. 1, pp. 3–34, 2015.

[13] F. Facchinei, G. Scutari, and S. Sagratella, "Parallel selective algorithms for nonconvex big data optimization," *IEEE Transactions on Signal Processing*, vol. 63, no. 7, pp. 1874–1889, 2015.

[14] Y. Nesterov and B. T. Polyak, "Cubic regularization of newton method and its global performance," *Mathematical Programming*, vol. 108, no. 1, pp. 177–205, 2006.

[15] S. J. Reddi, S. Sra, B. Póczos, and A. Smola, "Fast incremental method for smooth nonconvex optimization," in *IEEE 55th conference on decision and control (CDC)*, pp. 1971–1977, 2016.

[16] S. J. Reddi, A. Hefny, S. Sra, B. Poczos, and A. Smola, "Stochastic variance reduction for nonconvex optimization," in *International conference on machine learning*, pp. 314–323, PMLR, 2016.

[17] X. Cai, C. Song, S. Wright, and J. Diakonikolas, "Cyclic block coordinate descent with variance reduction for composite nonconvex optimization," in *International Conference on Machine Learning*, pp. 3469–3494, PMLR, 2023.

[18] Y. Xu and W. Yin, "A globally convergent algorithm for nonconvex optimization based on block coordinate update," *Journal of Scientific Computing*, vol. 72, no. 2, pp. 700–734, 2017.

[19] G. Carnevale and G. Notarstefano, "Nonconvex distributed optimization via Lasalle and singular perturbations," *IEEE Control Systems Letters*, vol. 7, pp. 301–306, 2022.

[20] G. Carnevale, N. Bastianello, G. Notarstefano, and R. Carli, "Admm-tracking gradient for distributed optimization over asynchronous and unreliable networks," *arXiv preprint arXiv:2309.14142*, 2023.

[21] W. M. Haddad and V. Chellaboina, "Nonlinear dynamical systems and control," in *Nonlinear Dynamical Systems and Control*, Princeton university press, 2011.