ARCHIVIO ISTITUZIONALE
DELLA RICERCA

Alma Mater Studiorum Università di Bologna
Archivio istituzionale della ricerca

Distributed Big-Data Optimization via Block-wise Gradient Tracking

This is the final peer-reviewed author's accepted manuscript (postprint) of the following publication:

Published Version:
Notarnicola, I., Sun, Y., Scutari, G., Notarstefano, G. (2021). Distributed Big-Data Optimization via Block-wise Gradient Tracking. IEEE TRANSACTIONS ON AUTOMATIC CONTROL, 66(5), 2045-2060 [10.1109/TAC.2020.3008713].

Availability:
This version is available at: https://hdl.handle.net/11585/790157 since: 2021-01-21

Published:

DOI: http://doi.org/10.1109/TAC.2020.3008713

Terms of use:

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/).
When citing, please refer to the published version.

(Article begins on next page)

13 October 2026

This is the post peer-review accepted manuscript of:

I. Notarnicola, Y. Sun, G. Scutari and G. Notarstefano, "Distributed Big-Data Optimization via Block-wise Gradient Tracking," in IEEE Transactions on Automatic Control, 2020.

The published version is available online at:

https://doi.org/ 10.1109/TAC.2020.3008713

© 2020 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

Distributed Big-Data Optimization via Block-wise Gradient Tracking

Ivano Notarnicola*, *Member, IEEE*, Ying Sun*, Gesualdo Scutari, *Senior Member, IEEE*, Giuseppe Notarstefano, *Member, IEEE*

***Abstract*—We study *distributed big-data nonconvex* optimization in multi-agent networks. We consider the (constrained) minimization of the sum of a smooth (possibly) nonconvex function, i.e., the agents' sum-utility, plus a convex (possibly) nonsmooth regularizer. Our interest is on big-data problems in which there is a large number of variables to optimize. If treated by means of standard distributed optimization algorithms, these large-scale problems may be intractable due to the prohibitive local computation and communication burden at each node. We propose a novel distributed solution method where, at each iteration, agents update in an uncoordinated fashion only one block of the entire decision vector. To deal with the nonconvexity of the cost function, the novel scheme hinges on Successive Convex Approximation (SCA) techniques combined with a novel *block-wise* perturbed push-sum consensus protocol, which is instrumental to perform local block-averaging operations and tracking of gradient averages. Asymptotic convergence to stationary solutions of the nonconvex problem is established. Finally, numerical results show the effectiveness of the proposed algorithm and highlight how the block dimension impacts on the communication overhead and practical convergence speed.**

I. Introduction

Many modern control, estimation and learning applications lead to large-scale optimization problems, i.e., problems with a huge number of variables to optimize. These problems are often referred to as *big-data*, and call for the design of tailored algorithms. In this paper we consider *distributed* (nonconvex) *big-data optimization*. That is, we aim at solving large-scale optimization problems over networks in a distributed way by addressing the following two challenges: (i) *optimizing over (or even computing the gradient with respect to) all the variables can be too costly, and* (ii) *broadcasting to neighbors the entire solution estimate would incur an unaffordable communication overhead.* The literature on parallel and distributed methods is abundant; however, we are not aware of any work that can deal with both challenges (i) and (ii) over networks, as detailed next.

*These authors equally contributed and are in alphabetic order.

Preliminary short versions of this paper have appeared as [1], [2].

The work of Notarnicola and Notarstefano has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 638992 - OPT4SMART). The work of Sun and Scutari has been supported by the USA National Science Foundation under Grants CIF 1564044, CIF 1719205, CMMI 1832688; and in part by the Army Research Office under Grant W911NF1810238.

Notarnicola and Notarstefano are with the Department of Electrical, Electronic and Information Engineering, University of Bologna, Bologna, Italy, {ivano.notarnicola,giuseppe.notarstefano}@unibo.it

Sun and Scutari are with the School of Industrial Engineering, Purdue University, West-Lafayette, IN, USA, {sun578,gscutari}@purdue.edu.

A. Related Works

We organize the relevant literature in two main groups: centralized and parallel algorithms for large-scale optimization; and distributed algorithms applicable to multi-agent networks (with no specific topology).

Parallel algorithms. Parallel Block-Coordinate-Descent (BCD) methods are well-established methods in optimization; more recently, they have been proven to be particularly effective in solving very large-scale (mainly convex) optimization problems arising, e.g., from data-intensive applications. Examples include [3] for convex, smooth functions, and [4], [5] for composite optimization; a detailed overview of BCD methods can be found in [6], [7]. Parallel solution methods based on Successive Convex Approximation (SCA) techniques have been proposed in [8] to deal with nonconvex problems; see [9] for a recent research tutorial on the subject. In [10] block coordinate-descent and stochastic-gradient methods have been combined to optimize big-data, sum-of-utilities (cost) functions. These algorithms, however, are not implementable in a (fully) distributed setting; they are instead designed to be run on ad-hoc computational architectures, e.g., shared-memory systems or star networks.

Distributed multi-agent algorithms. The literature on distributed methods for multi-agent optimization is vast. Here, we discuss only primal-based algorithms, as they are more closely related to the approach proposed in this paper. Distributed subgradient methods have been proposed in the early works [11], [12], to solve convex, problems over undirected graphs. The extension to nonconvex costs has been developed in [13]. The generalization to (time-varying) digraphs was studied in [14] and [15] for convex and nonconvex objectives, respectively; these schemes combine distributed (sub-)gradient with push-sum consensus [16] updates. A Nesterov acceleration of the mentioned approach applied to convex, smooth problems has been proposed in [17] with a convergence rate analysis. Local, private constraints are handled in [18] and [19], where distributed methods based on a random projection (asynchronous) subgradient and a proximal minimization are proposed respectively. All these methods need to use a diminishing step-size to converge to an exact, consensual solution, thus converging at a sub-linear rate. On the other hand, with a constant (sufficiently small) step-size, they can be faster, but they would converge only to a neighborhood of the solution set.

Primal-based distributed methods that converge to an exact consensual solution using fixed step-sizes are available in the literature; they can be roughly grouped as i) [20], [21]; ii) [22]–[24], iii) [9], [25]–[30]; and iv) [31]–[34]. While substan-

tially different, these schemes build on the idea of correcting the decentralized gradient- (or Newton-) related direction to cancel the steady state error in it. More specifically, in [20] and its proximal variant [21], the gradient direction is corrected using iterate and gradient information of the last two iterations. In [22]–[24], the novel idea of distributively estimating a Newton-Raphson direction by means of suitable average consensus ratios has been introduced. In [35] the same approach has been extended to deal with directed, asynchronous networks with lossy communications. The third and fourth class of works is based on the idea of gradient tracking: each agent updates its own local variables along a surrogate direction that tracks the gradient of the sum-utility (which is not locally available). This idea was proposed independently in [25], [26] for constrained nonsmooth nonconvex problems, and in [27], [30] for strongly convex, unconstrained, smooth, optimization. The works [9], [28], [29] extended the algorithms to (possibly) time-varying digraphs (still in the nonconvex setting of [25], [26]). A convergence rate analysis of the scheme [27] was later developed in [31], [32], [36], [37], with [31], [36] considering time-varying (directed) graphs. Another scheme, still based on the idea of gradient tracking, has been recently proposed in [34]. All the above methods are based on the optimization and communication at each iteration of the *entire* set of variables of every agents (or some related quantities of the same size).

First attempts to block-wise distributed optimization have been proposed in [38]–[40] for a structured, *partitioned* optimization set-up in which the cost function of each agent depends on its (block) variables and those of its neighbors. In [41] a distributed stochastic gradient method has been proposed whereby agents optimize at each iteration only a subset of their variables (still communicating the entire vector).

B. Major Contributions

We propose a distributed algorithm over networks for, possibly nonconvex, big-data optimization problems, that explicitly accounts for challenges (i) and (ii). To cope with these two challenges, we propose a distributed scheme in which, at every iteration, each agent optimizes over and communicates only *one block* of the local solution estimate (and of auxiliary vectors) rather than all the components. Blocks are selected in an uncoordinated fashion by means of an "essentially cyclic rule". Such a rule states that all blocks must be selected within a finite number of iterations, thus guaranteeing all of them to be persistently updated during the algorithmic evolution. Specifically, inspired to the two optimization algorithms NEXT (in-Network succEssive conveX approximaTion) [25], [26] and SONATA (distributed Successive cONvex Approximation algorithm over Time-varying digrAphs) [28], [29], *not suitable for big-data* problems, we propose a block-iterative two-step (optimization and averaging) procedure, named BLOCK-SONATA. Each agent solves a (small) local optimization problem, depending only on the selected block, with cost function being a strongly convex surrogate of the nonconvex sum-cost function, whose gradient is a local estimate of the total gradient of the (smooth part of the) sum-cost function. The (block-wise) optimization step is combined with a twofold *block-wise* perturbed averaging scheme on the local solution estimate and on the local estimate of the total gradient. This scheme guarantees both the asymptotic agreement of the local solution estimates and the tracking of total gradient. We remark that this novel block-wise perturbed averaging protocol extends a (static) block averaging protocol proposed for an abstract message passing model in [42], and is thus of independent interest for other distributed computation tasks. It can be used by agents of a network to reach consensus or track the average of local signals by exchanging with neighboring agents only one block of their local vector. For the proposed distributed optimization algorithm we prove that: local solution estimates are asymptotically consensual to their (weighted) average, and any limit point of the average sequence is a stationary solution of the optimization problem. The algorithm analysis has two key distinctive features. First, a proper convergence analysis of the block-wise perturbed averaging scheme is developed based on suitable block-induced time-varying digraphs. Second, errors due to inexact block-wise minimizations and to uncoordinated block updates are properly handled to show that a suitably designed merit function decreases along the algorithmic evolution.

The rest of the paper is organized as follows. In Section II we present the problem set-up. In Section III we introduce a block-wise perturbed consensus scheme that will act as a building block for our distributed big-data optimization algorithm presented in Section IV, along with its convergence properties. In Section V we provide numerical computations to test our algorithm. Finally, the convergence analysis is deferred to the appendix.

II. Distributed Big-Data Optimization Set-up

We consider a multi-agent system composed of N agents, aiming at cooperatively solving the following composite (possibly) nonconvex large-scale optimization problem

$$\begin{aligned} \min_{\mathbf{x}} \quad & U(\mathbf{x}) \triangleq \sum_{i=1}^{N} f_i(\mathbf{x}) + \sum_{\ell=1}^{B} r_\ell(\mathbf{x}_\ell) \\ \text{subj. to } & \mathbf{x}_\ell \in \mathcal{K}_\ell, \quad \ell \in \{1,\dots,B\}, \end{aligned} \tag{1}$$

where $\mathbf{x}$ is the vector of optimization variables, partitioned in B blocks as

$$\mathbf{x} \triangleq [\mathbf{x}_1^\top, \dots, \mathbf{x}_B^\top]^\top,$$

with $\mathbf{x}_\ell \in \mathbb{R}^d$, $\ell \in \{1,\dots,B\}$; $f_i : \mathbb{R}^{dB} \to \mathbb{R}$ is the cost function of agent i, assumed to be smooth but (possibly) nonconvex; $r_\ell : \mathbb{R}^d \to \mathbb{R}$, $\ell \in \{1,\dots,B\}$, is a convex (possibly) nonsmooth function; and $\mathcal{K}_\ell$, $\ell \in \{1,\dots,B\}$, is a closed convex set; we denote by $\mathcal{K} \triangleq \mathcal{K}_1 \times \cdots \times \mathcal{K}_B$ the feasible set of (1). The nonsmooth terms r_ℓ are usually used to promote some extra structure in the solution, such as (group) sparsity. We study (1) under the following assumption.

Assumption II.1 (On the Optimization Problem)**.**

(i) Each $\mathcal{K}_\ell \neq \emptyset$ is closed and convex;

(ii) Each $f_i : \mathbb{R}^{dB} \to \mathbb{R}$ is $\mathcal{C}^1$ on (an open set containing) $\mathcal{K}$;

(iii) Each ∇f_i is globally L_i-Lipschitz continuous and bounded on $\mathcal{K}$;

(iv) Each $r_\ell : \mathbb{R}^d \to \mathbb{R}$ is convex (possibly nonsmooth) on $\mathcal{K}_\ell$, with bounded subgradients on $\mathcal{K}_\ell$;

(v) U is coercive on $\mathcal{K}$, i.e., $\lim_{\mathbf{x}\in\mathcal{K},\|\mathbf{x}\|\to\infty} U(\mathbf{x}) = \infty$. □

The above assumptions are quite standard and satisfied by many practical problems; see, e.g., [8]. Here, we only remark that we do not assume any convexity of f_i. In the following, we also make the blanket assumption that each agent i knows only its own cost function f_i, the regularizers r_ℓ and the feasible set $\mathcal{K}$, but not the other agents' functions.

On the communication network: The communication among agents is modeled as a fixed, directed graph $\mathcal{G} = (\{1,\dots,N\},\mathcal{E})$, where $\mathcal{E} \subseteq \{1,\dots,N\}\times\{1,\dots,N\}$ is the set of edges. The edge $(i,j) \in \mathcal{E}$ models the fact that agent i can send a message to agent j. We denote by $\mathcal{N}_i$ the set of *in-neighbors* of node i in the fixed graph $\mathcal{G}$, i.e., $\mathcal{N}_i \triangleq \{j \in \{1,\dots,N\} \mid (j,i) \in \mathcal{E}\}$. We assume that $\mathcal{E}$ contains self-loops and, thus, $\mathcal{N}_i$ contains $\{i\}$ itself. We use the following assumption.

Assumption II.2. *The digraph $\mathcal{G}$ is strongly connected.* □

Algorithmic Desiderata: Our goal is to solve problem (1) in a distributed fashion, leveraging local communications among neighboring agents. As a major departure from current literature on distributed optimization, here we focus on *big-data* instances of (1) in which the vector of variables $\mathbf{x}$ is composed by a huge number of components (B is very large). In such problems, minimizing the sum-utility with respect to all the components of $\mathbf{x}$, or even computing the gradient or evaluating the value of a single function f_i, can require substantial computational efforts. Moreover, exchanging an estimate of the *entire* local decision variables $\mathbf{x}$ over the network (like current distributed schemes do) is not efficient or even feasible, due to the excessive communication overhead. We design next the first scheme able to deal with such challenges.

III. Block-wise Perturbed Push-Sum Consensus

In this section we design a building block of our distributed optimization algorithm, namely a *block-wise perturbed push-sum consensus algorithm*. We first devise the "unperturbed" instance of the scheme, suitable to solve the average consensus problem over digraphs via block-wise communication. Then, we introduce the general perturbed version of the scheme, which allows agents to solve more general tasks, such as tracking the average of given (time-varying) agents' signals. This scheme will then be employed in the block-wise averaging parts of our distributed optimization presented in Section IV.

A. Block-wise Push-sum Average Consensus

Consider a system of N agents aiming at reaching consensus on the average of given initial values. Let the communication network be modeled as a digraph $\mathcal{G}$ satisfying Assumption II.2. To solve this problem, one can leverage the popular push-sum (consensus) algorithm [16]. However, differently from this scheme in which agents need to exchange their *entire* local estimates at each iteration, here we consider a *block-wise* communication protocol. Specifically, at every iteration t each agent i updates its entire (average estimate) vector $\mathbf{z}^t_{(i,:)} \in \mathbb{R}^{dB}$, but it sends to out-neighbors *one block only*[1]. As customary in push-sum schemes, each agent i maintains also an additional variable in order to reach *average* consensus over directed graphs. In our scheme, an auxiliary variable is assigned to each block ℓ and we collect them in a single vector $\phi_{(i,:)} \in \mathbb{R}^B$.

We denote by $\ell^t_i \in \{1,\dots,B\}$ the index of the block that agent i chooses (according to a proper rule) to update and broadcasts to its out-neighbors at iteration t; and by $\mathbf{z}^t_{(i,\ell^t_i)} \in \mathbb{R}^d$ such a block. To update $\mathbf{z}^t_{(i,:)}$, agent i runs a push-sum consensus on each block ℓ of $\mathbf{z}^t_{(i,:)}$ *separately*, using only the information received from its in-neighbors that sent block ℓ at time t (if any). The same block-wise protocol is used for the auxiliary variable $\phi^t_{(i,:)}$.

Since no coordination is assumed among agents in selecting their blocks, different agents will likely select blocks with different indices, i.e., $\ell^t_i \neq \ell^t_j$, with $i \neq j$. This induces a *block-dependent* communication graph, one for each block index ℓ, which is, in general, a subgraph of $\mathcal{G}$. In this subgraph, agent j is an in-neighbor of agent i at time t if $j \in \mathcal{N}_i$ and $\ell^t_j = \ell$, i.e., agent j sent its block ℓ to i at time t. This suggests the definition of *block-dependent* neighbor sets. For each agent $i \in \{1,\dots,N\}$ and each block $\ell \in \{1,\dots,B\}$, define

$$\mathcal{N}^t_{i,\ell} \triangleq \{j \in \mathcal{N}_i \mid \ell^t_j = \ell\} \cup \{i\} \subseteq \mathcal{N}_i,$$

which includes, besides agent i, only the in-neighbors of agent i in $\mathcal{G}$ that sent (i.e., updated) block ℓ at time t. Consistently, we denote by $\mathcal{G}^t_\ell \triangleq (\{1,\dots,N\},\mathcal{E}^t_\ell)$ the *time-varying* subgraph of $\mathcal{G}$ associated to block ℓ at iteration t. Its edge set is $\mathcal{E}^t_\ell \triangleq \{(j,i) \in \mathcal{E} \mid j \in \mathcal{N}^t_{i,\ell}, i \in \{1,\dots,N\}\}$.

Note that each (time-varying) digraph $\mathcal{G}^t_\ell$ is induced by the block selection rules (independently) adopted by the agents, so that the connectivity properties of all digraphs are coupled; this interplay will be discussed shortly (cf. Assumption III.2 and Proposition III.3).

The following table "Block-wise Push-sum Average Consensus" formally introduces the algorithm from the perspective of agent i only. The algorithm consists of applying the push-sum consensus protocol in a *block-wise* fashion over the time-varying subgraphs $\mathcal{G}^t_\ell$ introduced above. As in the existing consensus protocols, $a^t_{ij\ell}$ in (2) are nonnegative weights to be properly chosen. We let $\mathbf{A}^t_\ell \triangleq [a^t_{ij\ell}]^N_{i,j=1}$ be the weight-matrix matching $\mathcal{G}^t_\ell$ (cf. Assumption III.1). Each agent $i \in \{1,\dots,N\}$ initializes its local variables as $\phi^0_{(i,\ell)} = 1$ and $\mathbf{z}^0_{(i,\ell)}$ an arbitrary value in $\mathbb{R}^d$ for all $\ell \in \{1,\dots,B\}$.

Convergence of the Block-wise Push-sum Average Consensus depends on the choice of the weight matrices as well as the block-selection rules employed by the agents (which affect the connectivity properties of each digraph sequence $\{\mathcal{G}^t_\ell\}_{t\geq 0}$, $\ell \in \{1,\dots,B\}$). Sufficient conditions on these parameters guaranteeing convergence are discussed next.

On the choice of $\mathbf{A}^t_\ell$: We make the following assumption on each $\mathbf{A}^t_\ell$, which is standard for the push-sum algorithm.

[1] We use the semicolon in the subscript of a vector to denote the vector stacking column-wise all its blocks.

Block-wise Push-sum Average Consensus

Select $\ell_i^t \in \{1,\dots,B\}$
For each $j \in \mathcal{N}_i$ receive $\phi^t_{(j,\ell_j^t)}$ and $\mathbf{z}^t_{(j,\ell_j^t)}$
For each $\ell \in \{1,\dots,B\}$ compute

$$\begin{aligned}\phi^{t+1}_{(i,\ell)} &= \sum_{j\in\mathcal{N}^t_{i,\ell}} a^t_{ij\ell}\,\phi^t_{(j,\ell)}\\ \mathbf{z}^{t+1}_{(i,\ell)} &= \sum_{j\in\mathcal{N}^t_{i,\ell}} \frac{a^t_{ij\ell}\,\phi^t_{(j,\ell)}}{\phi^{t+1}_{(i,\ell)}}\,\mathbf{z}^t_{(j,\ell)}\end{aligned} \tag{2}$$

Assumption III.1. *Given the sequence of graphs $\{\mathcal{G}^t_\ell\}_{t\geq 0}$, $\ell \in \{1,\dots,B\}$ and $t \geq 0$, each matrix $\mathbf{A}^t_\ell$ satisfies the following:*
(a) $a^t_{ij\ell} \geq \kappa > 0$, if $(j,i) \in \mathcal{E}^t_\ell$; and $a^t_{ij\ell} = 0$, if $(j,i) \notin \mathcal{E}^t_\ell$;
(b) it is column stochastic, that is, $\mathbf{1}^\top \mathbf{A}^t_\ell = \mathbf{1}^\top$. □

A natural question is whether a matrix $\mathbf{A}^t_\ell$ satisfying Assumption III.1 can be build by the agents using only local information. Next, we propose a simple procedure to locally build a valid $\mathbf{A}^t_\ell$. As the underlying communication digraph is static and strongly connected (cf. Assumption II.2), we assume that a column stochastic matrix $\tilde{\mathbf{A}}$ matching $\mathcal{G}$ is available, i.e., $\tilde{a}_{ij} > 0$ if $(j,i) \in \mathcal{E}$ and $\tilde{a}_{ij} = 0$ otherwise; and $\mathbf{1}^\top \tilde{\mathbf{A}} = \mathbf{1}^\top$. To construct $\mathbf{A}^t_\ell$ in a distributed way, we start noticing that at iteration t, an agent j either sends a block ℓ to all its out-neighbors in $\mathcal{G}$, $\ell = \ell_j^t$, or to none, $\ell \neq \ell_j^t$. Thus, let us focus on the j-th column of $\mathbf{A}^t_\ell$, denoted by $\mathbf{A}^t_\ell(:,j)$. If agent j does not send block ℓ at iteration t, $\ell \neq \ell_j^t$, all the components of $\mathbf{A}^t_\ell(:,j)$ will be zero except for $a^t_{jj\ell}$. Thus, for the j-th column to be stochastic, it must be $\mathbf{1}^\top \mathbf{A}^t_\ell(:,j) = a^t_{jj\ell} = 1$ (i.e., $\mathbf{A}^t_\ell(:,j)$ is the j-th vector of the canonical basis). This is a legitimate choice since $j \in \mathcal{N}^t_{j,\ell}$ by definition. Vice versa, if j sends block ℓ, all its out-neighbors in $\mathcal{G}$ will receive it and, thus, column $\mathbf{A}^t_\ell(:,j)$ has the same nonzero entries as column $\tilde{\mathbf{A}}(:,j)$ of $\tilde{\mathbf{A}}$. Since $\tilde{\mathbf{A}}$ is column stochastic, one can set $\mathbf{A}^t_\ell(:,j) = \tilde{\mathbf{A}}(:\,j)$. Note that, since $\tilde{\mathbf{A}}$ is assumed to be locally known, each agent can construct its own weights satisfying the above rule without requiring any additional coordination with other agents. In summary, for each $i \in \{1,\dots,N\}$ and $\ell \in \{1,\dots,B\}$, weights $a^t_{ij\ell}$ can be chosen as

$$a^t_{ij\ell} \triangleq \begin{cases} \tilde{a}_{ij}, & \text{if } j \in \mathcal{N}_i \text{ and } \ell = \ell_j^t,\\ 1, & \text{if } j = i \text{ and } \ell \neq \ell_i^t,\\ 0, & \text{otherwise.}\end{cases} \tag{3}$$

On the choice of the block selection rule: To guarantee convergence of the Block-wise Push-sum Average Consensus over time-varying digraphs, it is well known that some long-term connectivity property is required on the digraph sequence [16]. Here, we use T-strong connectivity: for each $\ell \in \{1,\dots,B\}$, the time-varying digraphs $\{\mathcal{G}^t_\ell\}$ are T-strongly connected, i.e., the union digraph $\bigcup_{h=0}^{T-1} \mathcal{G}^{t+h}_\ell$ is strongly connected $\forall\, t \geq 0$.

The T-strong connectivity requirement imposes a condition on the way the blocks are selected. Note that $\mathcal{G}^t_\ell$ is a subgraph of $\mathcal{G}$ such that if agent i selects (sends) block ℓ at time t, then the edges in $\mathcal{E}$ leaving node i are also present in $\mathcal{E}^t_\ell$. Hence, since $\mathcal{G}$ is strongly connected (cf. Assumption II.2), the following general *essentially cyclic* rule is enough to guarantee that each $\{\mathcal{G}^t_\ell\}$ is T-strongly connected.

Assumption III.2 (Block Selection Rule)**.** *For each agent $i \in \{1,\dots,N\}$ there exists a (finite) constant $T_i > 0$ such that*

$$\bigcup_{h=0}^{T_i-1} \{\ell_i^{t+h}\} = \{1,\dots,B\}, \text{ for all } t \geq 0. \qquad \square$$

Note that the above rule does not impose any coordination among the agents; they select their own block independently. Therefore, at a given iteration t, different agents may update different blocks. Moreover, some blocks can be updated more often than others. On the other hand, such a rule guarantees that, within a finite time window of length $T \leq \max_{i\in\{1,\dots,N\}} T_i$, all the blocks have been updated at least once by all agents. This is sufficient to ensure that $\mathcal{G}^t_\ell$ is T-strongly connected, as formally stated next.

Proposition III.3. *Under Assumption II.2 and III.2, there exits a $0 < T \leq \max_{i\in\{1,\dots,N\}} T_i$, such that, for each $\ell \in \{1,\dots,B\}$, the union digraph $\bigcup_{h=0}^{T-1} \mathcal{G}^{t+h}_\ell$ is strongly connected, for all $t \geq 0$.*

Proof. For each block ℓ, the set of out-neighbors of each agent i in $\bigcup_{h=0}^{T-1} \mathcal{G}^{t+h}_\ell$ equals the set of the out-neighbors of agent i in $\mathcal{G}$ due to Assumption III.2. Hence, by Assumption II.2, $\bigcup_{h=0}^{T-1} \mathcal{G}^{t+h}_\ell$ is strongly connected. □

B. Block-wise Perturbed Push-sum

We can now generalize the Block-wise Push-sum Average Consensus introducing in the agents' local updates a local block-wise, time-varying (deterministic) perturbation, denoted by $\boldsymbol{\epsilon}^t_{(i,\ell)}$. Later in this subsection we will specify a perturbation to perform signal tracking, while in the following section proper perturbations will be devised to design the proposed distributed optimization algorithm. The block-wise perturbed push-sum can be obtained by replacing the update (2) with the following perturbed version

$$\begin{aligned}\phi^{t+1}_{(i,\ell)} &= \sum_{j\in\mathcal{N}^t_{i,\ell}} a^t_{ij\ell}\,\phi^t_{(j,\ell)},\\ \mathbf{z}^{t+1}_{(i,\ell)} &= \sum_{j\in\mathcal{N}^t_{i,\ell}} \frac{a^t_{ij\ell}\,\phi^t_{(j,\ell)}}{\phi^{t+1}_{(i,\ell)}}\left(\mathbf{z}^t_{(j,\ell)} + \boldsymbol{\epsilon}^t_{(j,\ell)}\right),\end{aligned} \tag{4}$$

for all $\ell \in \{1,\dots,B\}$, where $\boldsymbol{\epsilon}^t_{(j,\ell)} \in \mathbb{R}^d$ are suitable perturbations that agents inject in their update. This scheme is a building block of proposed block-wise distributed optimization algorithm that will be introduced in the next section. Convergence of the block-wise perturbed push-sum algorithm is stated in the following proposition.

Proposition III.4. *Consider the block-wise perturbed push-sum consensus* (4)*, with weight matrix $\mathbf{A}^t_\ell$ defined according to* (3)*. Then, under Assumptions II.2 and III.2, for all $i \in \{1,\dots,N\}$, it holds*

$$\left\|\mathbf{z}^t_{(i,:)} - \frac{1}{N}\sum_{j=1}^N \mathbf{z}^t_{(j,:)}\right\| \leq \sum_{\ell=1}^B \left\|\mathbf{z}^t_{(i,\ell)} - \frac{1}{N}\sum_{j=1}^N \mathbf{z}^t_{(j,\ell)}\right\|$$

$$\leq c_1(\rho)^t + c_2 \sum_{\ell=1}^{B} \sum_{h=1}^{t} (\rho)^{t-h} \sum_{j=1}^{N} \|\boldsymbol{\epsilon}_{(j,\ell)}^h\|_1,$$

for some $\rho \in (0,1)$ *and positive constants* c_1 *and* c_2. □

The proof of the proposition can be obtained by the proof of [14, Lemma 1], which we report in Appendix as Lemma A.3, in vector form, as a preliminary result needed for our analysis. As a corollary (with no proof), the previous result states that if the perturbations $\boldsymbol{\epsilon}_{(i,\ell)}^t$ are vanishing, i.e, $\lim_{t\to\infty} \|\boldsymbol{\epsilon}_{(i,\ell)}^t\| = 0$, for all $\ell \in \{1,\dots,B\}$ and $i \in \{1,\dots,N\}$, it holds $\lim_{t\to\infty} \left\| \mathbf{z}_{(i,:)}^t - \frac{1}{N}\sum_{j=1}^N \mathbf{z}_{(j,:)}^t \right\| = 0$, for all $i \in \{1,\dots,N\}$. Clearly, for $\boldsymbol{\epsilon}_{(i,\ell)}^t = \mathbf{0}$ for all $t \geq 0$, $\ell \in \{1,\dots,B\}$ and $i \in \{1,\dots,N\}$, the block-wise perturbed push-sum reduces to the Block-wise Push-sum Average Consensus.

Several tasks can be accomplished by suitably choosing the perturbation $\boldsymbol{\epsilon}_{(i,\ell)}^t$ in (4). As a case study, in the following we show how to choose the perturbation, in a block-wise fashion, in order to track the average of time-varying signals over graphs. The resulting block-wise tracking scheme will be part of the proposed distributed optimization algorithm.

Block-wise average signal tracking. Consider the problem of tracking the average of N time-varying signals over a graph $\mathcal{G}$, [43], [44]. Specifically, assume each agent i has access to a time-varying signal, say $\{\mathbf{u}_i^t\}_{t\in\mathbb{N}}$, with each $\mathbf{u}_i^t \in \mathbb{R}^{dB}$, and aims at tracking the average signal $\bar{\mathbf{u}}^t \triangleq (1/N)\cdot\sum_{i=1}^N \mathbf{u}_i^t$ by exchanging information over the network. Existing tracking schemes, e.g. ones used in distributed optimization algorithms [24]–[34], [36], [37], require the acquisition and communication at each iteration of the entire signal $\mathbf{u}_i^t$, which might be too costly in a big-data setting. To cope with the curse of dimensionality, we can leverage the block-wise perturbed push-sum consensus algorithm: to track distributedly $\bar{\mathbf{u}}^t$, one can show that it is sufficient to set $\boldsymbol{\epsilon}_{(i,\ell)}^t$ in (4) to

$$\boldsymbol{\epsilon}_{(i,\ell)}^t = \frac{1}{\phi_{(i,\ell)}^t}\Big(\mathbf{u}_{i,\ell}^{t+1} - \mathbf{u}_{i,\ell}^t\Big), \tag{5}$$

where $\mathbf{u}_{i,\ell}^t$ denotes the ℓ-th block of $\mathbf{u}_i^t$.

While the tracking scheme (4)–(5) unlocks block-communications over networks, it requires, at each iteration, to potentially perform (5) for all the blocks $\ell \in \{1,\dots,B\}$, i.e. the evaluation (acquisition) of the *entire* signal $\mathbf{u}_i^t$. When the cost of acquiring $\mathbf{u}_i^t$ is non-negligible, e.g., $\mathbf{u}_i^t$ can be the gradient of a function with respect to a large number of variables, it is advisable to modify the protocol so that, at each iteration, only *one* block of $\mathbf{u}_i^t$ is used. To this end, we propose to replace $\mathbf{u}_i^t$ with a surrogate local variable, denoted by $\widehat{\mathbf{u}}_i^t$, initialized as $\widehat{\mathbf{u}}_i^0 = \mathbf{u}_i^0$. At each iteration t, agent i acquires only a block of $\mathbf{u}_i^t$, say the ℓ_i^t-th block, and updates $\widehat{\mathbf{u}}_i^t$ as

$$\widehat{\mathbf{u}}_{i,\ell}^t = \begin{cases} \mathbf{u}_{i,\ell}^t, & \text{if } \ell = \ell_i^t, \\ \widehat{\mathbf{u}}_{i,\ell}^{t-1}, & \text{if } \ell \neq \ell_i^t, \end{cases}$$

where, as in (5), $\widehat{\mathbf{u}}_{i,\ell}^t$ denotes the ℓ-th block of $\widehat{\mathbf{u}}_i^t$. That is, the vector $\widehat{\mathbf{u}}_i^t$ collects agent i's most recent information on $\mathbf{u}_i^t$.

The modified block-tracking scheme then reads

$$\phi_{(i,\ell)}^{t+1} = \sum_{j\in\mathcal{N}_{i,\ell}^t} a_{ij\ell}^t\, \phi_{(j,\ell)}^t,$$

$$\mathbf{z}_{(i,\ell)}^{t+1} = \sum_{j\in\mathcal{N}_{i,\ell}^t} \frac{a_{ij\ell}^t}{\phi_{(i,\ell)}^{t+1}} \Big(\phi_{(j,\ell)}^t \mathbf{z}_{(j,\ell)}^t + (\widehat{\mathbf{u}}_{j,\ell}^{t+1} - \widehat{\mathbf{u}}_{j,\ell}^t)\Big).$$

IV. Block-SONATA Distributed Algorithm

In this section we introduce our distributed big-data optimization algorithm (cf. Section IV-A) along with its convergence properties (cf. Section IV-B). Some extensions of the basic scheme are discussed in Section IV-C.

A. Algorithm Description

The proposed distributed algorithm takes inspiration from two existing optimization algorithms, namely: NEXT (in-Network succEssive conveX approximaTion) [25], [26] and SONATA (distributed Successive cONvex Approximation algorithm over Time-varying digrAphs) [28], [29]. These algorithms combine successive convex approximation techniques with a distributed gradient tracking mechanism to solve convex and nonconvex optimization problems over time-varying (di)graphs. Specifically, they consist of a two-step procedure in which each agent: (i) solves a local strongly convex approximation of the target optimization problem, and (ii) runs a twofold averaging scheme to reach consensus among the local solution estimates and to "track" the average of the gradient of agents' cost functions (the smooth part). As all the other existing schemes, they are not designed to solve big-data optimization problems over networks: they require that, at every iteration, agents solve a huge-scale optimization problem and communicate their entire solution estimate to neighbors.

We propose a distributed algorithm, named Block-SONATA, based on a block-wise execution of steps (i) and (ii) above. It copes with big-data optimization problems by unlocking for the first time block-wise optimization and communications. While the intuitive idea behind this block extension might look simple, we will show that the convergence analysis of Block-SONATA is quite challenging. Indeed it calls for new techniques to deal with local inexact (block-wise) optimization and communications, the latter inducing block-dependent time-varying digraphs in the consensus updates.

Block-SONATA reads as follows. Each agent maintains a local solution estimate $\mathbf{x}_{(i,:)}^t \in \mathbb{R}^{dB}$ of problem (1), with the same block structure as the optimization variable $\mathbf{x}$, with $\mathbf{x}_{(i,\ell)}^t \in \mathbb{R}^d$ being its ℓ-th block-component. All these estimates are iteratively updated with the goal of being asymptotically consensual to a stationary solution of problem (1). Agents also update a local auxiliary variable $\mathbf{y}_{(i,:)}^t \in \mathbb{R}^{dB}$ that is meant to track $\frac{1}{N}\sum_{j=1}^N \nabla f_j(\mathbf{x}_{(j,:)}^t)$ (which is not known locally by the agents), i.e., to get, for any agent i, $\lim_{t\to\infty} \|\mathbf{y}_{(i,:)}^t - \frac{1}{N}\sum_{j=1}^N \nabla f_j(\mathbf{x}_{(j,:)}^t)\| = 0$. The update of the $\mathbf{x}$- and $\mathbf{y}$-variables is described next.

Block-wise local optimization step. At iteration t, every agent i selects a block $\ell_i^t \in \{1,\dots,B\}$ according to an essentially cyclic rule satisfying Assumption III.2. As for the

optimization step, agent i computes a descent direction with respect to the selected block (only) by solving a strongly convex approximation of problem (1) (based on its current solution and gradient estimates, respectively $\mathbf{x}^t_{(i,:)}$ and $\mathbf{y}^t_{(i,\ell^t_i)}$). Specifically, it solves

$$\widetilde{\mathbf{x}}^t_{(i,\ell^t_i)} = \underset{\mathbf{x}_{\ell^t_i}\in\mathcal{K}_{\ell^t_i}}{\operatorname{argmin}} \ \widehat{f}_{i,\ell^t_i}\big(\mathbf{x}_{\ell^t_i};\mathbf{x}^t_{(i,:)},\mathbf{y}^t_{(i,\ell^t_i)}\big) + r_{\ell^t_i}(\mathbf{x}_{\ell^t_i}),$$

with

$$\widehat{f}_{i,\ell^t_i}\big(\mathbf{x}_{\ell^t_i};\mathbf{x}^t_{(i,:)},\mathbf{y}^t_{(i,\ell^t_i)}\big) = \tilde{f}_{i,\ell^t_i}(\mathbf{x}_{\ell^t_i};\mathbf{x}^t_{(i,:)}) + (N\mathbf{y}^t_{(i,\ell^t_i)} - \nabla_{\ell^t_i} f_i(\mathbf{x}^t_{(i,:)}))^\top(\mathbf{x}_{\ell^t_i} - \mathbf{x}^t_{(i,\ell^t_i)}),$$

where $\tilde{f}_{i,\ell^t_i}(\mathbf{x}_{\ell^t_i};\mathbf{x}^t_{(i,:)})$ is a strongly convex approximation of f_i satisfying the following assumption.

Assumption IV.1 (On the surrogate functions)**.** *Given problem* (1) *under Assumption II.1, each surrogate function $\tilde{f}_{i,\ell} : \mathcal{K}_\ell \times \mathcal{K} \to \mathbb{R}$ is chosen so that*

(i) $\tilde{f}_{i,\ell}(\bullet;\mathbf{x})$ is uniformly strongly convex with constant $\tau_i > 0$ on $\mathcal{K}_\ell$;
(ii) $\nabla\tilde{f}_{i,\ell}(\mathbf{x}_\ell;\mathbf{x}) = \nabla_\ell f_i(\mathbf{x})$, for all $\mathbf{x}\in\mathcal{K}$;
(iii) $\nabla\tilde{f}_{i,\ell}(\mathbf{x}_\ell;\bullet)$ is uniformly Lipschitz continuous on $\mathcal{K}$;

where $\nabla\tilde{f}_{i,\ell}$ denotes the partial gradient of $\tilde{f}_{i,\ell}$ with respect to its first argument. □

Several choices for $\tilde{f}_{i,\ell}$ are possible; we refer the interested reader to [1], [2], [8], [26], [29] for more details and examples. We point out that each strongly convex function $\widehat{f}_{i,\ell}\big(\mathbf{x}_\ell;\mathbf{x}^t_{(i,:)},\mathbf{y}^t_{(i,\ell)}\big)$ satisfies $\nabla\widehat{f}_{i,\ell}\big(\mathbf{x}^t_{(i,\ell)};\mathbf{x}^t_{(i,:)},\mathbf{y}^t_{(i,\ell)}\big) = N\mathbf{y}^t_{i,\ell}$, thus it asymptotically encodes first order information of $\sum_i f_i$, namely $\sum_{i=1}^N \nabla f_i(\mathbf{x}^t_{(i,:)})$. As a clarifying example, one can consider the simplest first order approximation of f_i given by its linearization about the current iterate $\mathbf{x}^t_{(i,:)}$,

$$\widehat{f}_{i,\ell}\big(\mathbf{x}_\ell;\mathbf{x}^t_{(i,:)},\mathbf{y}^t_{(i,\ell)}\big) = (N\mathbf{y}^t_{(i,\ell)})^\top(\mathbf{x}_\ell - \mathbf{x}^t_{(i,\ell)}) + \tau_i\|\mathbf{x}_\ell - \mathbf{x}^t_{(i,\ell)}\|^2.$$

Given $\widetilde{\mathbf{x}}^t_{(i,\ell^t_i)}$, agent i computes and broadcasts to its neighbors the feasible point $\mathbf{x}^t_{(i,\ell^t_i)} + \gamma^t\Delta\mathbf{x}^t_{(i,\ell^t_i)}$, with $\Delta\mathbf{x}^t_{(i,\ell^t_i)} = \widetilde{\mathbf{x}}^t_{(i,\ell^t_i)} - \mathbf{x}^t_{(i,\ell^t_i)}$ being a local feasible descent direction and γ^t a step-size.

Block-wise averaging and gradient tracking step. As for the consensus steps, agent i collects all the updated blocks from its neighbors and runs two instances of the block-wise perturbed push-sum consensus scheme, described in Section III (Cf. eq. (4)). The first one is meant to make the local solution estimates, $\mathbf{x}^t_{(i,:)}$, consensual toward their average; the second, involving a local gradient estimate $\mathbf{y}^t_{(i,:)}$, serves as a tracking scheme for the gradient signal $\sum_{i=1}^N \nabla f_i(\mathbf{x}^t_{(i,:)})$.

The BLOCK-SONATA distributed algorithm is summarized (from the perspective of node i) in the next table.

Remark IV.2. *We would like to stress that agents send* only one *block per iteration. Hence, each "for" loop over ℓ [cf.* (8)–(10)*] consists of at most $|\mathcal{N}_i|$ non-trivial consensus steps. Thus, each agent i receives exactly $|\mathcal{N}_i|(2d+1)$ updated quantities per iteration. Moreover, due to the presence of the weights $a^t_{ij\ell}$, each non-trivial consensus step requires to sum at most $|\mathcal{N}_i|$ terms over all the blocks.* □

BLOCK-SONATA

Initialization: $\mathbf{x}^0_{(i,:)} \in \mathcal{K}$ arbitrary and $\mathbf{y}^0_{(i,:)} = \nabla f_i(\mathbf{x}^0_{(i,:)})$

Local Optimization:

Select $\ell^t_i \in \{1,\dots,B\}$ and compute

$$\widetilde{\mathbf{x}}^t_{(i,\ell^t_i)} = \underset{\mathbf{x}_{\ell^t_i}\in\mathcal{K}_{\ell^t_i}}{\operatorname{argmin}} \ \widehat{f}_{i,\ell^t_i}\big(\mathbf{x}_{\ell^t_i};\mathbf{x}^t_{(i,:)},\mathbf{y}^t_{(i,\ell^t_i)}\big) + r_{\ell^t_i}(\mathbf{x}_{\ell^t_i}) \tag{6}$$

$$\Delta\mathbf{x}^t_{(i,\ell)} = \begin{cases} \widetilde{\mathbf{x}}^t_{(i,\ell^t_i)} - \mathbf{x}^t_{(i,\ell^t_i)}, & \text{if } \ell = \ell^t_i, \\ \mathbf{0}, & \text{otherwise.} \end{cases} \tag{7}$$

Averaging and Gradient Tracking:

For each $j\in\mathcal{N}_i$ receive $\phi^t_{(j,\ell^t_j)}$ and $\mathbf{x}^t_{(j,\ell^t_j)} + \gamma^t\Delta\mathbf{x}^t_{(j,\ell^t_j)}$.

For each $\ell\in\{1,\dots,B\}$ compute

$$\phi^{t+1}_{(i,\ell)} = \sum_{j\in\mathcal{N}^t_{i,\ell}} a^t_{ij\ell}\,\phi^t_{(j,\ell)} \tag{8}$$

$$\mathbf{x}^{t+1}_{(i,\ell)} = \sum_{j\in\mathcal{N}^t_{i,\ell}} \frac{a^t_{ij\ell}\,\phi^t_{(j,\ell)}}{\phi^{t+1}_{(i,\ell)}}\Big(\mathbf{x}^t_{(j,\ell)} + \gamma^t\Delta\mathbf{x}^t_{(j,\ell)}\Big) \tag{9}$$

For each $j\in\mathcal{N}_i$ receive $\big(\phi^t_{(j,\ell^t_j)}\mathbf{y}^t_{(j,\ell^t_j)} + \nabla_{\ell^t_j} f_j\big(\mathbf{x}^{t+1}_{(j,:)}\big) - \nabla_{\ell^t_j} f_j\big(\mathbf{x}^t_{(j,:)}\big)\big)$

For each $\ell\in\{1,\dots,B\}$ compute

$$\mathbf{y}^{t+1}_{(i,\ell)} = \sum_{j\in\mathcal{N}^t_{i,\ell}} \frac{a^t_{ij\ell}}{\phi^{t+1}_{(i,\ell)}}\Big(\phi^t_{(j,\ell)}\mathbf{y}^t_{(j,\ell)} + \nabla_\ell f_j\big(\mathbf{x}^{t+1}_{(j,:)}\big) - \nabla_\ell f_j\big(\mathbf{x}^t_{(j,:)}\big)\Big). \tag{10}$$

B. Algorithm Convergence

We now provide the main convergence result of BLOCK-SONATA. We first introduce the following assumption on the step-size sequence $\{\gamma^t\}_{t\geq 0}$ [cf. (9)].

Assumption IV.3 (On the step-size)**.** *The sequence $\{\gamma^t\}_{t\geq 0}$, with each $0 < \gamma^t \leq 1$, satisfies: (i) $\gamma^{t+1} \leq \gamma^t$, for all $t\geq 0$; and (ii) $\sum_{t=0}^{\infty}\gamma^t = \infty$ and $\sum_{t=0}^{\infty}(\gamma^t)^2 < \infty$.* □

The above conditions are standard and satisfied by most practical diminishing step-size rules. For example, the following rule, proposed in [8], satisfies Assumption IV.3 and has been found very effective in our experiments: $\gamma^{t+1} = \gamma^t(1-\mu\gamma^t)$, with $\gamma^0\in(0,1]$ and $\mu\in(0,1/\gamma^0)$.

We are now in the position to state the main convergence result, as given below.

Theorem IV.4. *Let $\{(\mathbf{x}^t_{(i,:)})_{i=1}^N\}_{t\geq 0}$ be the sequences generated by* BLOCK-SONATA *and consider their weighted average*

$$\bar{\boldsymbol{s}}^t = \frac{1}{N}\Big(\sum_{i=1}^N \phi^t_{(i,\ell)}\mathbf{x}^t_{(i,\ell)}\Big)_{\ell=1}^B.$$

Suppose that Assumptions II.1, II.2, III.1, III.2, IV.1 and IV.3 are satisfied; then the following statements hold true:

(i) consensus: $\|\mathbf{x}^t_{(i,:)} - \bar{\boldsymbol{s}}^t\| \to 0$ as $t\to\infty$, for all $i\in\{1,\dots,N\}$;

(ii) `convergence`: *$\{\bar{\mathbf{s}}^t\}_{t\geq 0}$ is bounded and every of its limit points is a stationary solution of problem* (1).

Proof. See the Appendix. □

Theorem IV.4 states two results. First, a consensus is asymptotically achieved among the local estimates $\mathbf{x}^t_{(i,:)}$ over all the blocks. Second, the weighted average estimate $\bar{\mathbf{s}}^t$ converges to the set $\mathcal{S}$ of stationary solutions of problem (1).

C. Alternative Formulations and Generalizations

In this subsection, we discuss some extensions and generalizations of the basic BLOCK-SONATA. First, we start describing a special instance for an unconstrained version of problem (1) with all $r_\ell = 0$. If one chooses the simplest surrogate in (6), namely the linearization of f_i about the current iterate, then BLOCK-SONATA reads

$$\phi^{t+1}_{(i,\ell)} = \sum_{j\in\mathcal{N}^t_{i,\ell}} a^t_{ij\ell}\,\phi^t_{(j,\ell)}$$

$$\mathbf{x}^{t+1}_{(i,\ell)} = \sum_{j\in\mathcal{N}^t_{i,\ell}} \frac{a^t_{ij\ell}\,\phi^t_{(j,\ell)}}{\phi^{t+1}_{(i,\ell)}}\Big(\mathbf{x}^t_{(j,\ell)} - \gamma^t \mathbf{y}^t_{(j,\ell)}\Big)$$

$$\mathbf{y}^{t+1}_{(i,\ell)} = \sum_{j\in\mathcal{N}^t_{i,\ell}} \frac{a^t_{ij\ell}}{\phi^{t+1}_{(i,\ell)}}\Big(\phi^t_{(j,\ell)}\,\mathbf{y}^t_{(j)} + \nabla_\ell f_j(\mathbf{x}^{t+1}_{(j,:)}) - \nabla_\ell f_j(\mathbf{x}^t_{(j,:)})\Big),$$

which is a block-wise implementation of distributed algorithms based on a gradient tracking scheme as, e.g., [26], [29]–[34].

Combine-Then-Adapt Averaging. The block-wise consensus and tracking updates as in (9) and (10) are performed in the so-called Adapt-Then-Combine (ATC) fashion. We remark that they can be also performed adopting the other scheme used in the literature, namely the so-called Combine-Then-Adapt (CTA) way [45]. The CTA form of the averaging and gradient tracking step of BLOCK-SONATA reads

$$\mathbf{x}^{t+1}_{(i,\ell)} = \sum_{j\in\mathcal{N}^t_{i,\ell}} \frac{a^t_{ij\ell}\phi^t_{(j,\ell)}}{\phi^{t+1}_{(i,\ell)}}\,\mathbf{x}^t_{(j,\ell)} + \gamma^t\phi^t_{(i,\ell)}\Delta\mathbf{x}^t_{(i,\ell)}$$

$$\mathbf{y}^{t+1}_{(i,\ell)} = \sum_{j\in\mathcal{N}^t_{i,\ell}} \frac{a^t_{ij\ell}\,\phi^t_{(j,\ell)}}{\phi^{t+1}_{(i,\ell)}}\mathbf{y}^t_{(j,\ell)} + \frac{\nabla_\ell f_i\big(\mathbf{x}^{t+1}_{(i,:)}\big) - \nabla_\ell f_i\big(\mathbf{x}^t_{(i,:)}\big)}{\phi^{t+1}_{(i,\ell)}}.$$

Following the same steps of the proof and with minor changes of the absolute constants, one can show that Theorem IV.4 also applies to the CTA form of BLOCK-SONATA, which thus converges under the same condition of its ATC counterpart.

Block-Wise Gradient Computation. In order to perform (10) (and also its CTA counterpart), agent i needs to compute the entire gradient $\nabla f_i(\mathbf{x}^{t+1}_{(i,:)})$ [recall from (3) that $a_{ii\ell} > 0$, for all ℓ]. This potential drawback can be overcome considering a slightly different version of BLOCK-SONATA in which $\nabla_\ell f_i$ is replaced by an auxiliary variable $\widehat{\mathbf{g}}_{(i,\ell)}$, which is iteratively updated as

$$\widehat{\mathbf{g}}^{t+1}_{(i,\ell)} = \begin{cases} \nabla_{\ell^t_i} f_i\big(\mathbf{x}^{t+1}_{(i,:)}\big), & \text{if } \ell = \ell^t_i, \\ \widehat{\mathbf{g}}^t_{(i,\ell)}, & \text{otherwise.} \end{cases}$$

Thus, step (10) must be replaced by

$$\mathbf{y}^{t+1}_{(i,\ell)} = \sum_{j\in\mathcal{N}^t_{i,\ell}} \frac{a^t_{ij\ell}\phi^t_{(j,\ell)}}{\phi^{t+1}_{(i,\ell)}}\mathbf{y}^t_{(j,\ell)} + \frac{\widehat{\mathbf{g}}^{t+1}_{(i,\ell)} - \widehat{\mathbf{g}}^t_{(i,\ell)}}{\phi^{t+1}_{(i,\ell)}}.$$

Remark IV.5. *The auxiliary mechanism of $\widehat{\mathbf{g}}$ imposes that, at each iteration t of* BLOCK-SONATA*, each agent i computes two components of the same gradient, $\nabla f_i(\mathbf{x}^t_{(i,:)})$, rather than one. This twofold computation can be avoided by using a slight modification of the scheme in which the block index is selected after the optimization step, see [2] for further details.* □

V. NUMERICAL STUDY: APPLICATION TO SPARSE REGRESSION

In this section we apply BLOCK-SONATA to the distributed sparse regression problem. Consider a network of N agents taking linear measurements of a sparse signal $\mathbf{x}_0 \in \mathbb{R}^m$, with measurement matrix $\mathbf{D}_i \in \mathbb{R}^{n_i\times m}$. The observation taken by agent i can be expressed as $\mathbf{b}_i = \mathbf{D}_i\mathbf{x}_0 + \mathbf{n}_i$, where $\mathbf{n}_i \in \mathbb{R}^{n_i}$ accounts for the measurement noise. The estimation of the underlying signal $\mathbf{x}_0$ is obtained solving the following problem

$$\min_{\mathbf{x}\in\mathcal{K}} \sum_{i=1}^N \underbrace{\|\mathbf{D}_i\mathbf{x} - \mathbf{b}_i\|_2^2}_{f_i(\mathbf{x})} + r(\mathbf{x}), \tag{11}$$

where $\mathbf{x} \in \mathbb{R}^m$; $\mathcal{K}$ is the box constraint set $\mathcal{K} \triangleq [k_L, k_U]^m$, with $k_L \leq k_U$; and $r : \mathbb{R}^m \to \mathbb{R}$ is a difference-of-convex (DC) sparsity-promoting regularizer, given by

$$r(\mathbf{x}) \triangleq \lambda \cdot \sum_{j=1}^m r_0(x_j), \quad r_0(x_j) \triangleq \frac{\log(1+\theta|x_j|)}{\log(1+\theta)},$$

where λ and θ are positive tuning parameters.

The first step to apply BLOCK-SONATA is to build a valid surrogate $\tilde{f}_{i,\ell}$ of f_i (cf. Assumption IV.1). To this end, we first rewrite r_0 as a difference-of-convex function. It is not difficult to check that

$$r_0(x) = \underbrace{\eta(\theta)\,|x|}_{r_0^+(x)} - \underbrace{\big(\eta(\theta)\,|x| - r_0(x)\big)}_{r_0^-(x)},$$

where $r_0^+ : \mathbb{R} \to \mathbb{R}$ is convex non-smooth with $\eta(\theta) \triangleq \theta/\log(1+\theta)$, and $r_0^- : \mathbb{R} \to \mathbb{R}$ is convex with Lipschitz continuous first order derivative given by

$$\frac{dr_0^-}{dx}(x) = \operatorname{sign}(x)\cdot\frac{\theta^2|x|}{\log(1+\theta)(1+\theta|x|)}.$$

Denoting the coordinates associated with the ℓ-th block as $\mathcal{I}_\ell \subset \{1,\dots,B\}$, let us define the matrix $\mathbf{D}_{i,\ell}$ [resp. $\mathbf{D}_{i,-\ell}$] constructed by picking the columns of $\mathbf{D}_i$ that belongs [resp. does not belong] to $\mathcal{I}_\ell$. Then, the following is a valid surrogate function for each agent i that satisfy Assumption IV.1. We consider $\tilde{f}_i$ obtained as the linearization of f_i and $-r_0^-$, about the current solution estimate, which leads to

$$\tilde{f}_{i,\ell}(\mathbf{x}_\ell;\mathbf{x}^t_{(i,:)}) = \big(2\mathbf{D}^\top_{i,\ell}(\mathbf{D}_i - \mathbf{b}_i)\big)^\top(\mathbf{x}_\ell - \mathbf{x}^t_{(i,\ell)}) + \frac{\tau_i}{2}\|\mathbf{x}_\ell - \mathbf{x}^t_{(i,\ell)}\|^2 - \sum_{k\in\mathcal{I}_\ell}\underbrace{\frac{dr_0^-((\mathbf{x}^t_{(i,\ell)})_k)}{dx}}_{v^t_{ik}}(\mathbf{x}_\ell - \mathbf{x}^t_{(i,\ell)})_k,$$

where x is a scalar variable and, e.g., $(\mathbf{x}^t_{(i,\ell)})_k$ denotes the k-th scalar component of $\mathbf{x}^t_{(i,\ell)}$. Note that the minimizer of $\tilde{f}_{i,\ell}$ can be computed in closed form, and is given by

$$\mathbf{x}^{t+1}_{(i,\ell)} = \mathcal{P}_{\mathcal{K}_\ell}\Big(\mathcal{S}_{\frac{\lambda\eta}{\tau_i}}\Big(\mathbf{x}^t_{(i,\ell)} - \frac{1}{\tau_i}(2\,\mathbf{D}^\top_{i,\ell}(\mathbf{D}_i - \mathbf{b}_i) - \mathbf{v}^t_{i,\ell})\Big)\Big)$$

where $\mathbf{v}^t_{i,\ell} \triangleq (v^t_{ik})_{k\in\mathcal{I}_\ell}$, $\mathcal{S}_\lambda(\mathbf{x}) \triangleq \text{sign}(\mathbf{x}) \cdot \max\{|\mathbf{x}| - \lambda, 0\}$ (operations are performed element-wise), and $\mathcal{P}_{\mathcal{K}_\ell}$ is the Euclidean projection onto $\mathcal{K}_\ell$.

We test our algorithm, considering the following simulation set-up. The variable dimension m is set to be 400, $\mathcal{K}$ is set to be $[-10, 10]^{400}$, and the regularization parameters are set to $\lambda = 0.15$ and $\theta = 7$. The network is composed of $N = 30$ agents, communicating over an undirected graph $\mathcal{G}$, obtained using an Erdős-Rényi random model. We considered two extreme network topologies: a densely and a poorly connected one, which have algebraic connectivity equal to 25 and 5, respectively. The components of the ground-truth signal $\mathbf{x}_0$ are i.i.d., generated according to the Normal distribution $\mathcal{N}(0, 1)$. To impose sparsity on $\mathbf{x}_0$, we set the smallest 80% of the entries of $\mathbf{x}_0$ to zero. Each agent i has a measurement matrix $\mathbf{D}_i \in \mathbb{R}^{300\times 400}$ with i.i.d. $\mathcal{N}(0, 1)$ distributed entries (with ℓ_2-normalized rows), and the observation noise $\mathbf{n}_i \in \mathbb{R}^{300}$ has entries i.i.d. distributed according to $\mathcal{N}(0, 0.5)$.

We compare our algorithm with the (sub)gradient-projection algorithm proposed in [13]. Note that there is no formal proof of convergence for such an algorithm in the nonconvex setting; moreover it is designed for the non-block-wise case, i.e., $B = 1$. We used the following tuning for the algorithms. The diminishing step-size is chosen as $\gamma^t = \gamma^{t-1}(1 - \mu\gamma^{t-1})$, with $\gamma^0 = 0.3$ and $\mu = 10^{-3}$; the proximal parameter $\tau_i = 10$ for all i. To evaluate the algorithmic performance we used three merit functions. The first one measures the distance from stationarity of the average of the agents' iterates $\bar{\boldsymbol{s}}^t = \frac{1}{N}(\sum_{i=1}^N \phi^t_{(i,\ell)}\mathbf{x}^t_{(i,\ell)})^B_{\ell=1}$, and is given by

$$J^t \triangleq \Big\|\bar{\boldsymbol{s}}^t - \mathcal{P}_{\mathcal{K}}\Big(\mathcal{S}_{\lambda\eta}\Big(\bar{\boldsymbol{s}}^t - \Big(\sum_{i=1}^N \nabla f_i(\bar{\boldsymbol{s}}^t) - r(\bar{\boldsymbol{s}}^t)\Big)\Big)\Big)\Big\|_\infty.$$

Note that J^t is a valid merit function: it is continuous and it is zero if and only if the $\bar{\boldsymbol{s}}^t$ is a stationary solution of problem (11). The other two merit functions quantify the consensus disagreement at each iteration among the solution estimates and the trackers. They are defined as

$$D^t \triangleq \max_{i\in\{1,\dots,N\}} \|\mathbf{x}^t_{(i,:)} - \bar{\boldsymbol{s}}^t\|,$$
$$R^t \triangleq \max_{i\in\{1,\dots,N\}} \|\mathbf{y}^t_{(i,:)} - \bar{\boldsymbol{\sigma}}^t\|,$$

where the average $\bar{\boldsymbol{s}}^t$ is defined as before, while the average tracker is $\bar{\boldsymbol{\sigma}}^t = \frac{1}{N}(\sum_{i=1}^N \phi^t_{(i,\ell)}\mathbf{y}^t_{(i,\ell)})^B_{\ell=1}$.

The performance of BLOCK-SONATA for different choices of the block dimension B are reported in Figure 1. To fairly compare the algorithms run for different block sizes, we plot J^t, D^t and R^t versus the average agents' "message exchanges", defined as t/B, where t is the iteration counter used in the algorithm description. The figures show that stationarity, consensus and correct tracking have been achieved

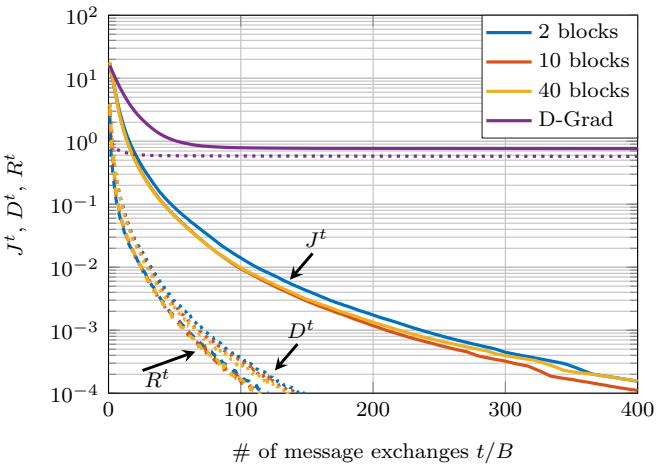


Figure 1. Optimality measurement J^t (solid), consensus error D^t (dotted) and tracking error R^t (dashed) versus the number of message exchange for several choices of the number of blocks B.

by BLOCK-SONATA within 200 message exchanges while the plain gradient scheme [13] is much slower.

Let t_{end} be the completion time up to a tolerance 10^{-3}, i.e., the iteration counter of the distributed algorithm such that $J^{t_{\text{end}}} < 10^{-3}$. Fig. 2 shows the normalized completion time t_{end}/B versus the number of blocks B. It highlights how the communication cost reduces by increasing the number of blocks. Notice that for $B = 1$ BLOCK-SONATA reduces to SONATA (cf. [28], [29], hence Fig. 2 shows also how the proposed scheme compares with the state of the art for non-big-data problems.

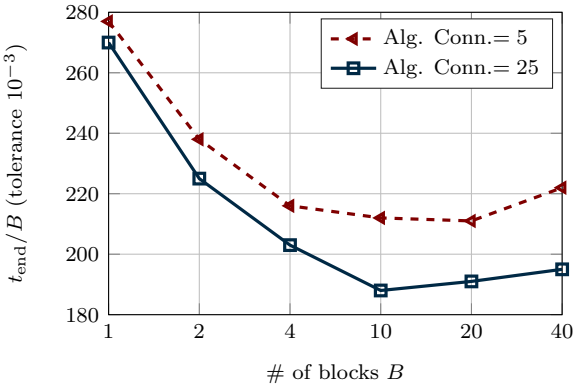


Figure 2. Completion time required to obtain $J^t < 10^{-3}$ versus the number of blocks B for two network topologies having different algebraic connectivity.

VI. CONCLUSIONS

In this paper we proposed a novel block-iterative distributed scheme for nonconvex, big-data optimization problems over networks. That is, we addressed large-scale optimization problems in which the dimension of the decision vector is huge via a distributed algorithm (over network) in which each agent optimizes over and communicates one block only of the entire

decision vector. Specifically, at each iteration, agents solve a local optimization problem involving only one block of the decision vector The optimization step is combined with a novel block-wise perturbed consensus protocol based on the communication to neighboring agents of one block only. We proved that agents achieve consensus to their (weighted) average, and that any limit point of the average sequence is a stationary solution of the optimization problem. Finally, we provided numerical results corroborating our theoretical findings and highlighting the impact of the block dimension on algorithm performance. Future research directions include the extensions of the design of distributed big-data optimization methods under more general network scenarios, including delays, noise or packet losses.

APPENDIX

To study convergence of BLOCK-SONATA, it is convenient to introduce some auxiliary variables, namely: $\mathbf{s}^t_{(i,:)} \triangleq (\mathbf{s}^t_{(i,\ell)})_{\ell=1}^B$ and $\boldsymbol{\sigma}^t_{(i,:)} \triangleq (\boldsymbol{\sigma}^t_{(i,\ell)})_{\ell=1}^B$, for all $i \in \{1,\dots,N\}$. Steps (8), (9), and (10) in BLOCK-SONATA can be then rewritten as: for all $\ell \in \{1,\dots,B\}$ and $i \in \{1,\dots,N\}$,

$$\phi^{t+1}_{(i,\ell)} = \sum_{j\in\mathcal{N}_i} a^t_{ij\ell}\, \phi^t_{(j,\ell)}, \tag{12}$$

$$\mathbf{s}^{t+1}_{(i,\ell)} = \sum_{j\in\mathcal{N}_i} a^t_{ij\ell}\Big(\mathbf{s}^t_{(j,\ell)} + \gamma^t \phi^t_{(j,\ell)}\Delta\mathbf{x}^t_{(j,\ell)}\Big), \tag{13}$$

$$\mathbf{x}^{t+1}_{(i,\ell)} = \frac{\mathbf{s}^{t+1}_{(i,\ell)}}{\phi^{t+1}_{(i,\ell)}}, \tag{14}$$

$$\boldsymbol{\sigma}^{t+1}_{(i,\ell)} = \sum_{j\in\mathcal{N}_i} a^t_{ij\ell}\Big(\boldsymbol{\sigma}^t_{(j,\ell)} + \nabla_\ell f_j\big(\mathbf{x}^{t+1}_{(j,:)}\big) - \nabla_\ell f_j\big(\mathbf{x}^t_{(j,:)}\big)\Big), \tag{15}$$

$$\mathbf{y}^{t+1}_{(i,\ell)} = \frac{\boldsymbol{\sigma}^{t+1}_{(i,\ell)}}{\phi^{t+1}_{(i,\ell)}}, \tag{16}$$

with each $\boldsymbol{\sigma}^0_{(i,:)} \triangleq \nabla f_i(\mathbf{x}^0_{(i,:)})$.

Averaging (13) and (15) over $i \in \{1,\dots,N\}$ and using the column stochasticity of each $\mathbf{A}^t_\ell$, yields the following dynamics for the block-averages: for each $\ell \in \{1,\dots,B\}$,

$$\bar{\boldsymbol{s}}^{t+1}_\ell = \bar{\boldsymbol{s}}^t_\ell + \gamma^t \frac{1}{N}\sum_{i=1}^N \phi^t_{(i,\ell)}\Delta\mathbf{x}^t_{(i,\ell)}, \tag{17}$$

$$\bar{\boldsymbol{\sigma}}^{t+1}_\ell = \bar{\boldsymbol{\sigma}}^t_\ell + \frac{1}{N}\sum_{i=1}^N \Big(\nabla_\ell f_i(\mathbf{x}^{t+1}_{(i,:)}) - \nabla_\ell f_i(\mathbf{x}^t_{(i,:)})\Big), \tag{18}$$

where $\bar{\boldsymbol{s}}^t_\ell \triangleq (1/N)\cdot\sum_{i=1}^N \mathbf{s}^t_{(i,\ell)}$ and $\bar{\boldsymbol{\sigma}}^t_\ell \triangleq (1/N)\cdot\sum_{i=1}^N \boldsymbol{\sigma}^t_{(i,\ell)}$. We also define $\bar{\boldsymbol{s}}^t \triangleq (\bar{\boldsymbol{s}}^t_\ell)_{\ell=1}^B$ and $\bar{\boldsymbol{\sigma}}^t \triangleq (\bar{\boldsymbol{\sigma}}^t_\ell)_{\ell=1}^B$. To prove Theorem IV.4, it is sufficient to show that: (i) all the local copies $\mathbf{x}^t_{(i,:)}$ converge to $\bar{\boldsymbol{s}}^t$; and (ii) every limit point of $\{\bar{\boldsymbol{s}}^t\}_{t\geq 0}$ is a stationary solution of problem (1).

Notice that given a linear dynamical system in the form (17), one can always write

$$\bar{\boldsymbol{s}}^{t+\theta_t}_\ell = \bar{\boldsymbol{s}}^t_\ell + \sum_{h=t}^{t+\theta_t-1} \mathbf{u}^h_\ell$$

for every integer $\theta_t \in [0,T]$, where we used the shorthand $\mathbf{u}^h_\ell = \gamma^h \frac{1}{N}\sum_{i=1}^N \phi^h_{(i,\ell)}\Delta\mathbf{x}^h_{(i,\ell)}$. Thus, if the input $\mathbf{u}^h_\ell$ is vanishing, i.e., $\lim_{h\to\infty}\|\mathbf{u}^h_\ell\| = 0$, there holds

$$\begin{aligned}\lim_{t\to\infty}\|\bar{\boldsymbol{s}}^{t+\theta_t}_\ell - \bar{\boldsymbol{s}}^t_\ell\| &\leq \lim_{t\to\infty}\sum_{h=t}^{t+\theta_t-1}\|\mathbf{u}^h_\ell\| \\ &\leq \lim_{t\to\infty}\sum_{h=t}^{t+T-1}\|\mathbf{u}^h_\ell\| = 0.\end{aligned} \tag{19}$$

Structure of the proof: The proof is organized as follows. In Section A, we introduce some preliminary results that will be used in the rest of the sections, namely: (i) a formal description of the perturbed push-sum algorithm along with its convergence properties; and (ii) a list of key properties of a best-response map $\widetilde{\mathbf{x}}^t$ and related quantities. Theorem IV.4(i) is proven in Section B, where convergence of the consensus updates (14) and tracking mechanism (16) is studied. More specifically, first we prove that $\lim_{t\to\infty}\|\mathbf{x}^t_{(i,:)} - \bar{\boldsymbol{s}}^t\| = 0$, for all $i \in \{1,\dots,N\}$ (cf. Proposition B.9), showing thus asymptotic consensus of the local estimates $\mathbf{x}^t_{(i,:)}$; and, second, $\lim_{t\to\infty}\|\mathbf{y}^t_{(i,:)} - \bar{\boldsymbol{\sigma}}^t\| = 0$, for all $i \in \{1,\dots,N\}$ (cf. Proposition B.10), which together with

$$\bar{\boldsymbol{\sigma}}^t = \frac{1}{N}\sum_{i=1}^N \nabla f_i(\mathbf{x}^t_{(i,:)}), \quad \forall t \geq 0, \tag{20}$$

proves that each $\mathbf{y}^t_{(i,:)}$ tracks asymptotically the average of the cost function gradients. In Section C, we study the descent properties of a suitably defined Lyapunov-like function along the trajectory $\{(\mathbf{x}^t_{(i,:)})_{i=1}^N, \bar{\boldsymbol{s}}^t\}_{t\geq 0}$. This result is instrumental to show (subsequence) convergence of $\{\bar{\boldsymbol{s}}^t\}_{t\geq 0}$ to stationary solutions of problem (1) [in the sense of Theorem IV.4(ii)], which is proven in Section D.

A. Technical preliminaries

1) Perturbed push-sum consensus: Consider a network of N agents communicating, at each time slot t, over the graph $\mathcal{G}^t \triangleq (\{1,\dots,N\}, \mathcal{E}^t)$. The vector form of the perturbed push-sum protocol introduced in [14] reads: for all $t \geq 0$,

$$\begin{aligned}\psi^{t+1}_i &= \sum_{j\in\mathcal{N}^t_i} a^t_{ij}\psi^t_j \\ \boldsymbol{\eta}^{t+1}_i &= \sum_{j\in\mathcal{N}^t_i} a^t_{ij}(\boldsymbol{\eta}^t_j + \boldsymbol{\epsilon}^t_j) \\ \mathbf{z}^{t+1}_i &= \frac{\boldsymbol{\eta}^{t+1}_i}{\psi^{t+1}_i},\end{aligned} \tag{21}$$

where $\psi_i \in \mathbb{R}$, $\boldsymbol{\eta}_i \in \mathbb{R}^n$, $\mathbf{z}_i \in \mathbb{R}^n$ are agent i's local variables, with $\psi^0_i = 1$, and $\{\boldsymbol{\epsilon}^t_i\}_{t\geq 0}$ is a given perturbation sequence (known by agent i only). The graph $\mathcal{G}^t$ and weight matrix $\mathbf{A}^t \triangleq (a_{ij})_{i,j=1}^N$ satisfy the following assumptions.

Assumption A.1. *The graph sequence $\{\mathcal{G}^t\}_{t\geq 0}$ is strongly connected, i.e., there exists an integer $T > 0$ such that the union digraph $\bigcup_{h=0}^{T-1}\mathcal{G}^{t+h} \triangleq (\{1,\dots,N\}, \cup_{h=0}^{T-1}\mathcal{E}^{t+h})$ is strongly connected for all $t \geq 0$.* □

Assumption A.2. *Each weight matrix $\mathbf{A}^t$ matches graph $\mathcal{G}^t$, that is, it satisfies*

(1) $a^t_{ij} = 0$, if $(j,i) \notin \mathcal{E}^t$; and $a^t_{ij} \geq \kappa > 0$, if $(j,i) \in \mathcal{E}^t$;

(2) $a_{ii}^t \geq \kappa > 0$*, for all* $i \in \{1,\dots,N\}$*;*
(3) $\mathbf{A}^t$ *is column stochastic, i.e.,* $\mathbf{1}^\top \mathbf{A}^t = \mathbf{1}^\top$. □

The convergence properties of the (scalar version of the) perturbed push-sum protocol have been studied in [14, Lemma 1], as summarized below [for the vector case (21)].

Lemma A.3. *Consider the perturbed push-sum protocol* (21) *under Assumptions A.1 and A.2. Then, the following hold:*

(1) For all $t \geq 0$ *and some* $\rho \in (0,1)$*,* $c_1, c_2 > 0$*,*

$$\left\| \mathbf{z}_i^{t+1} - \frac{1}{N}\sum_{j=1}^N (\boldsymbol{\eta}_j^t + \boldsymbol{\epsilon}_j^t) \right\| \leq c_1(\rho)^t + c_2 \sum_{h=1}^t (\rho)^{t-h} \sum_{i=1}^N \|\boldsymbol{\epsilon}_i^h\|_1; \tag{22}$$

(2) If the perturbations are vanishing, i.e., $\lim_{t\to\infty} \|\boldsymbol{\epsilon}_i^t\| = 0$*,* $i \in \{1,\dots,N\}$*, then*

$$\lim_{t\to\infty} \left\| \mathbf{z}_i^{t+1} - \frac{1}{N}\sum_{j=1}^N (\boldsymbol{\eta}_j^t + \boldsymbol{\epsilon}_j^t) \right\| = 0;$$

(3) The sequence $\{\psi_i^t\}_{t\geq 0}$ *satisfies*

$$\inf_{t\geq 0} \left(\min_{i\in\{1,\dots,N\}} \psi_i^t \right) \triangleq \delta > 0. \qquad \square$$

Note that, since $\mathbf{A}^t$ is column stochastic, we have $\bar{\boldsymbol{\eta}}^{t+1} \triangleq \frac{1}{N}\sum_{j=1}^N \boldsymbol{\eta}_j^{t+1} = \frac{1}{N}\sum_{j=1}^N (\boldsymbol{\eta}_j^t + \boldsymbol{\epsilon}_j^t)$. Therefore, the bound (22) can be written also as

$$\left\| \mathbf{z}_i^{t+1} - \bar{\boldsymbol{\eta}}^{t+1} \right\| \leq c_1(\rho)^t + c_2 \sum_{h=1}^t (\rho)^{t-h} \sum_{i=1}^N \|\boldsymbol{\epsilon}_i^h\|_1. \tag{23}$$

2) Properties of the best-response map $\widetilde{\mathbf{x}}^t$*:* In this subsection, we introduce some intermediate results dealing with key properties of a best-response map $\widetilde{\mathbf{x}}^t$ and related quantities. For notational simplicity, we state the results in a more abstract form, omitting time and agent index dependencies.

Consider the following optimization problem

$$\widetilde{\mathbf{x}} \triangleq \operatorname*{argmin}_{\mathbf{x}\in\mathcal{K}} \; h(\mathbf{x}) + r(\mathbf{x}) \tag{24}$$

where $\mathcal{K}$ is a closed convex set and h (resp. r) is a $\mathcal{C}^1$ (resp. convex, possibly nonsmooth) function on (an open set containing) $\mathcal{K}$. Given some $\mathbf{w} \in \mathcal{K}$, let us also introduce the function $\widehat{h}(\bullet; \mathbf{w}, \nabla h(\mathbf{w})) : \mathcal{K} \to \mathcal{K}$ (the explicit dependence of $\widehat{h}$ from $\mathbf{w}$ and $\nabla h(\mathbf{w})$ is immaterial for our discussion). We assume that $\widehat{h}(\bullet; \mathbf{w}, \nabla h(\mathbf{w}))$ satisfies the following conditions:

(1) $\widehat{h}(\bullet; \mathbf{w}, \nabla h(\mathbf{w}))$ is $\mathcal{C}^1$ (on an open set containing $\mathcal{K}$) and τ-strongly convex on $\mathcal{K}$;
(2) $\nabla \widehat{h}(\mathbf{w}; \mathbf{w}, \nabla h(\mathbf{w})) = \nabla h(\mathbf{w})$;
(3) $\nabla_{\mathbf{w}} \widehat{h}(\mathbf{x}; \mathbf{w}, \nabla h(\mathbf{w}))$ is uniformly Lipschitz continuous for all $\mathbf{x} \in \mathcal{K}$.

The function $\widehat{h}(\bullet; \mathbf{w}, \nabla h(\mathbf{w}))$ should be considered as a strongly convex approximation of h having the same gradient of h at $\mathbf{w}$. Given $\widehat{h}(\bullet; \mathbf{w}, \nabla h(\mathbf{w}))$, we can finally introduce the following optimization problem

$$\widehat{\mathbf{x}}(\mathbf{w}) = \operatorname*{argmin}_{\mathbf{x}\in\mathcal{K}} \; \widehat{h}(\mathbf{x}; \mathbf{w}, \nabla h(\mathbf{w})) + r(\mathbf{x}), \tag{25}$$

which can be considered as a convex approximation of (24).

The following results establish some key properties of the best-response maps $\widetilde{\mathbf{x}}$ and $\widehat{\mathbf{x}}$.

Lemma A.4. *Consider problem* (24) *under the further assumption that* h *is* τ*-strongly convex. Then, for all* $\mathbf{v} \in \mathcal{K}$*, the following hold:*

(i) $\|\widetilde{\mathbf{x}} - \mathbf{v}\| \leq \frac{1}{\tau}\|\nabla h(\mathbf{v})\| + \frac{1}{\tau}\|\widetilde{\nabla} r(\widetilde{\mathbf{x}})\|$*;*

(ii) $\nabla h(\mathbf{v})^\top (\widetilde{\mathbf{x}} - \mathbf{v}) \leq -\tau\|\widetilde{\mathbf{x}} - \mathbf{v}\|^2 - (r(\widetilde{\mathbf{x}}) - r(\mathbf{v}))$.

Proof. The proof follows readily from the first order optimality conditions of (24) and the convexity of r. □

Proposition A.5 ([8, Prop. 8])**.** *The best-response map* $\mathcal{K} \ni \mathbf{w} \longmapsto \widehat{\mathbf{x}}(\mathbf{w})$ *defined in* (25) *satisfies*

(1) $\widehat{\mathbf{x}}(\bullet)$ *is Lipschitz continuous on* $\mathcal{K}$*;*
(2) The set of the fixed-points of $\widehat{\mathbf{x}}(\bullet)$ *coincides with the set of stationary solutions of problem* (24)*; therefore* $\widehat{\mathbf{x}}(\bullet)$ *has a fixed point.* □

We can now customize the above results to our setting. Consider the best-response $\widetilde{\mathbf{x}}_{(i,\ell)}^t$ in (7); applying Lemma A.4(ii) we readily obtain the following.

Lemma A.6. *The best-response* $\widetilde{\mathbf{x}}_{(i,\ell)}^t$ *defined in* (7) *satisfies*

$$\left(\mathbf{y}_{(i,\ell)}^t\right)^\top \Delta\mathbf{x}_{(i,\ell)}^t \leq -\tau_i \|\Delta\mathbf{x}_{(i,\ell)}^t\|^2 - \left(r_\ell(\widetilde{\mathbf{x}}_{(i,\ell)}^t) - r_\ell(\mathbf{x}_{(i,\ell)}^t)\right), \tag{26}$$

for all $\ell \in \{1,\dots,B\}$. □

Finally, consider the best-response map $\mathcal{K} \ni \mathbf{w} \longmapsto \widehat{\mathbf{x}}_{(i,\ell)}(\mathbf{w})$, defined as

$$\widehat{\mathbf{x}}_{(i,\ell)}(\mathbf{w}) \triangleq \operatorname*{argmin}_{\mathbf{x}_\ell \in \mathcal{K}_\ell} \; \widehat{f}_{i,\ell}\Big(\mathbf{x}_\ell; \mathbf{w}, \frac{1}{N}\sum_{i=1}^N \nabla_\ell f_i(\mathbf{w})\Big) + r_\ell(\mathbf{x}_\ell). \tag{27}$$

Clearly (27) is an instance of (25). It follows readily from Proposition A.5 that $\widehat{\mathbf{x}}_{(i,\ell)}(\bullet)$ enjoys the following properties.

Lemma A.7. *The best-response* $\widehat{\mathbf{x}}_{(i,\ell)}(\bullet)$ *defined in* (27) *satisfies:*

(1) $\widehat{\mathbf{x}}_{(i,\ell)}(\bullet)$ *is* $\widehat{L}_{i,\ell}$*-Lipschitz continuous on* $\mathcal{K}$*;*
(2) The set of the fixed-points of $\widehat{\mathbf{x}}_{(i,:)}(\bullet) \triangleq \big(\widehat{\mathbf{x}}_{(i,\ell)}(\bullet)\big)_{\ell=1}^B$ *coincides with the set of stationary solutions of problem* (1). □

B. Convergence of Consensus and Tracking

In this subsection we prove that i) the local estimates $\mathbf{x}_{(i,:)}^t$ reach asymptotic consensus (cf. Proposition B.9); and ii) all $\mathbf{y}_{(i,:)}^t$ are asymptotically consensual while tracking the average of the gradients, namely $\frac{1}{N}\sum_{i=1}^N \nabla f_i(\mathbf{x}_{(:,i)}^t)$ (cf. Proposition B.10). Note that Proposition B.9 also proves statement (i) of Theorem IV.4.

1) Achieving consensus: We begin observing that, for each $\ell \in \{1,\dots,B\}$, the block-wise $\mathbf{x}$-update of BLOCK-SONATA [cf. (12)–(14)] is an instance of the perturbed push-sum algorithm (21), with $\boldsymbol{\epsilon}_i^t \triangleq \gamma^t \phi_{(i,\ell)}^t \Delta\mathbf{x}_{(i,\ell)}^t$ and $n = d$. By Lemma A.3(2), it follows that convergence of each $\mathbf{x}_{(i,\ell)}^t$ to the average $\bar{\mathbf{s}}_\ell^t$ can be readily proven showing that each $\Delta\mathbf{x}_{(i,\ell)}^t$ is *uniformly bounded*. In fact, this together with $\gamma^t \downarrow 0$ and $\phi_{(i,\ell)}^t \leq N$, for all $i \in \{1,\dots,N\}$ and $t \geq 0$, yields

$\lim_{t\to\infty}\|\boldsymbol{\epsilon}_i^t\|=0$ (cf. Proposition B.9). The following lemma proves that each $\Delta\mathbf{x}_{(i,\ell)}^t$ is uniformly bounded.

Lemma B.8. *Consider problem* (1) *under Assumption II.1, II.2, III.1, IV.1, IV.3. Let* $\{(\phi_{(i,:)}^t)_{i=1}^N\}_{t\geq 0}$, $\{(\mathbf{x}_{(i,:)}^t)_{i=1}^N\}_{t\geq 0}$ *and* $\{(\mathbf{y}_{(i,:)}^t)_{i=1}^N\}_{t\geq 0}$ *be generated by* BLOCK-SONATA. *Then, for all* $\ell\in\{1,\dots,B\}$ *and* $i\in\{1,\dots,N\}$, *the following holds:*

$$\sup_{t\geq 0}\left\|\mathbf{y}_{(i,\ell)}^t-\bar{\boldsymbol{\sigma}}_\ell^t\right\|<C_1, \tag{28}$$

and

$$\sup_{t\geq 0}\left\|\Delta\mathbf{x}_{(i,\ell)}^t\right\|<C_2, \tag{29}$$

where C_1 *and* C_2 *are some positive, finite scalars.*

Proof. We prove (28). Note that the gradient tracking in (12), (15) and (16) is an instance of the perturbed push-sum algorithm (21), with $\boldsymbol{\epsilon}_i^t\triangleq\nabla_\ell f_i\big(\mathbf{x}_{(i,:)}^{t+1}\big)-\nabla_\ell f_i\big(\mathbf{x}_{(i,:)}^t\big)$. By Lemma A.3 (cf. eq. (23)), we have

$$\begin{aligned}
&\left\|\mathbf{y}_{(i,\ell)}^t-\bar{\boldsymbol{\sigma}}_\ell^t\right\|\\
&\leq c_1(\rho)^{t-1}+\sum_{h=1}^{t-1}(\rho)^{t-1-h}\sum_{i=1}^N\left\|\nabla_\ell f_i(\mathbf{x}_{(i,:)}^{h+1})-\nabla_\ell f_i(\mathbf{x}_{(i,:)}^h)\right\|_1\\
&\leq c_1(\rho)^{t-1}+(2N\sqrt{d}\,M_F)\sum_{h=1}^{t-1}(\rho)^{t-1-h},
\end{aligned}$$

where $M_F\triangleq\sum_{i=1}^N M_{f_i}$ and M_{f_i} denotes the bound for ∇f_i [cf. Assumption II.1(iii)]. The above inequality proves (28). We prove now (29). Consider the case $\ell=\ell_i^t$ [for $\ell\neq\ell_i^t$, $\Delta\mathbf{x}_{(i,\ell)}^t=0$, trivially implying (29)]. Invoking Lemma A.4(i), with the following identifications: $\widetilde{\mathbf{x}}=\widetilde{\mathbf{x}}_{(i,\ell)}^t$, $h(\bullet)=\widehat{f}_{i,\ell_i^t}(\bullet;\mathbf{x}_{(i,:)}^t,\mathbf{y}_{(i,\ell_i^t)}^t)$, $r(\bullet)=r_{\ell_i^t}(\bullet)$, and $\mathcal{K}=\mathcal{K}_{\ell_i^t}$, yields

$$\left\|\Delta\mathbf{x}_{(i,\ell_i^t)}^t\right\|\leq\frac{N}{\tau_i}\left\|\mathbf{y}_{(i,\ell_i^t)}^t\right\|+\frac{B_r}{\tau_i},$$

where we used the fact that i) $\nabla\widehat{f}_{i,\ell_i^t}(\mathbf{x}_{(i,\ell_i^t)}^t;\mathbf{x}_{(i,:)}^t,\mathbf{y}_{(i,\ell_i^t)}^t)=N\mathbf{y}_{(i,\ell_i^t)}^t$ (cf. Assumption IV.1(ii)); and ii) $\|\widetilde{\nabla}r_{\ell_i^t}(\mathbf{x}_{(i,\ell_i^t)}^t)\|\leq B_r$ [cf. Assumption II.1(iv)]. By adding and subtracting $\bar{\boldsymbol{\sigma}}_{(i,\ell_i^t)}^t$ in $\|\mathbf{y}_{(i,\ell_i^t)}^t\|$ and using triangle inequality we can bound $\|\Delta\mathbf{x}_{(i,\ell_i^t)}^t\|$ as

$$\begin{aligned}
\|\Delta\mathbf{x}_{(i,\ell_i^t)}^t\|&\leq\frac{N}{\tau_i}\left\|\mathbf{y}_{(i,\ell_i^t)}^t-\bar{\boldsymbol{\sigma}}_{(i,\ell_i^t)}^t\right\|+\frac{N}{\tau_i}\left\|\bar{\boldsymbol{\sigma}}_{(i,\ell_i^t)}^t\right\|+\frac{B_r}{\tau_i}.\\
&\overset{(a)}{\leq}\frac{N}{\tau_i}\sum_{\ell=1}^B\left\|\mathbf{y}_{(i,\ell)}^t-\bar{\boldsymbol{\sigma}}_{(i,\ell)}^t\right\|\\
&\quad+\frac{1}{\tau_i}\Big\|\sum_{i=1}^N\nabla_{\ell_i^t}f_i(\mathbf{x}_{(i,:)}^t)\Big\|+\frac{B_r}{\tau_i}\\
&\overset{(b)}{\leq}\frac{N}{\tau_i}\sum_{\ell=1}^B\left\|\mathbf{y}_{(i,\ell)}^t-\bar{\boldsymbol{\sigma}}_{(i,\ell)}^t\right\|+\frac{N}{\tau_i}\cdot M_F+\frac{B_r}{\tau_i}\\
&\overset{(c)}{\leq}\frac{N}{\tau_i}\cdot B\cdot C_1+\frac{N}{\tau_i}\cdot M_F+\frac{B_r}{\tau_i}\triangleq C_2<\infty,
\end{aligned}$$

where in (a) we used (20); (b) follows from the boundedness of ∇f_i [cf. Assumption II.1(iii)]; and (c) comes from (28). □

We are now ready to characterize the dynamics of the consensus error, as given below.

Proposition B.9. *Consider problem* (1) *under Assumptions II.1, II.2, III.1, IV.1, IV.3. Let* $\{(\mathbf{x}_{(i,:)}^t)_{i=1}^N\}_{t\geq 0}$ *and* $\{(\mathbf{s}_{(i,:)}^t)_{i=1}^N\}_{t\geq 0}$ *be generated by* BLOCK-SONATA. *Then, the decision variables* $\mathbf{x}_{(i,:)}^t$ *are asymptotically consensual to* $\bar{\boldsymbol{s}}^t$*:*

$$\lim_{t\to\infty}\|\mathbf{x}_{(i,:)}^t-\bar{\boldsymbol{s}}^t\|=0, \tag{30}$$

for all $i\in\{1,\dots,N\}$. *Furthermore, the following hold:*

$$\sum_{t=0}^\infty\gamma^t\|\mathbf{x}_{(i,:)}^t-\bar{\boldsymbol{s}}^t\|<\infty, \tag{31}$$

$$\sum_{t=0}^\infty\|\mathbf{x}_{(i,:)}^t-\bar{\boldsymbol{s}}^t\|^2<\infty. \tag{32}$$

Proof. It is sufficient to prove (30)–(32) for each block ℓ.

Notice that the evolution of $\mathbf{x}_{(:,\ell)}^t$ [(12)–(14)] follows the dynamics of the perturbed push-sum algorithm (21), under the following identification: $n=d$, $\psi_i^t\triangleq\phi_{(i,\ell)}^t$, $\boldsymbol{\eta}_i^t\triangleq\mathbf{s}_{(i,\ell)}^t$, $\mathbf{z}_i^t\triangleq\mathbf{x}_{(i,\ell)}^t$, and $\boldsymbol{\epsilon}_{i,\ell}^t\triangleq\gamma^t\phi_{(i,\ell)}^t\Delta\mathbf{x}_{(i,\ell)}^t$. By Lemma B.8 [cf. (29)] and $\gamma^t\downarrow 0$, we infer $\lim_{t\to\infty}\boldsymbol{\epsilon}_{i,\ell}^t=\gamma^t\phi_{(i,\ell)}^t\Delta\mathbf{x}_{(i,\ell)}^t=0$. Invoking Lemma A.3(2), we conclude $\lim_{t\to\infty}\|\mathbf{x}_{(i,\ell)}^t-\bar{\boldsymbol{s}}_\ell^t\|=0$, which proves (30). We prove now (31). Using again the aforementioned connection with the perturbed push-sum algorithm (21), we can invoke Lemma A.3(1) [cf. (23)] and write

$$\begin{aligned}
&\sum_{t=0}^\infty\gamma^{t+1}\|\mathbf{x}_\ell^{t+1}-\bar{\boldsymbol{s}}_\ell^{t+1}\|\\
&\leq\sum_{t=0}^\infty\gamma^{t+1}\Big(c_1\,(\rho)^t+c_2\sum_{h=1}^t(\rho)^{t-h}\gamma^h\|\Delta\mathbf{x}_{(:,\ell)}^t\|_1\Big)\\
&\overset{(a)}{\leq}\sum_{t=0}^\infty\gamma^{t+1}\Big(c_1\,(\rho)^t+c_3\sum_{h=1}^t(\rho)^{t-h}\gamma^h\Big)\overset{(b)}{<}\infty,
\end{aligned}\tag{33}$$

for some finite, positive scalars c_1,c_2, and c_3, where (a) follows from the boundedness of $\|\Delta\mathbf{x}_{(:,:)}^t\|_1$ [cf. Lemma B.8]; and (b) is due to [12, Lemma 7].

Finally, to prove (32), we use the same bound of $\|\mathbf{x}_\ell^{t+1}-\bar{\boldsymbol{s}}_\ell^{t+1}\|$ as in (33), and write

$$\begin{aligned}
&\sum_{t=0}^\infty\|\mathbf{x}_\ell^{t+1}-\bar{\boldsymbol{s}}_\ell^{t+1}\|^2\\
&\leq\sum_{t=0}^\infty\Big(c_1^2(\rho)^{2t}+c_3^2\sum_{h=0}^t\sum_{s=0}^t\gamma^h\gamma^s(\rho)^{t-h}(\rho)^{t-s}\\
&\qquad+2c_1c_3\sum_{h=0}^t\gamma^h(\rho)^{t-h}(\rho)^t\Big)\overset{(a)}{<}\infty,
\end{aligned}$$

where (a) follows from [26, Lemma 7]. □

2) Asymptotic tracking: We conclude this section studying the dynamics of the gradient tracking scheme.

Proposition B.10. *Consider problem* (1) *under Assumptions II.1, II.2, III.1, IV.1, IV.3. Let* $\{(\mathbf{y}_{(i,:)}^t)_{i=1}^N\}_{t\geq 0}$ *be the sequence generated by* BLOCK-SONATA. *Then,* $\mathbf{y}_{(i,:)}^t$ *tracks the average of the gradients* $\sum_{j=1}^N\nabla_\ell f_j(\mathbf{x}_{(j,:)}^t)$ *asymptotically:*

$$\lim_{t\to\infty}\left\|\mathbf{y}_{(i,:)}^t-\frac{1}{N}\sum_{j=1}^N\nabla f_j(\mathbf{x}_{(j,:)}^t)\right\|=0, \tag{34}$$

for all $i\in\{1,\dots,N\}$. *Furthermore, the following holds:*

$$\sum_{t=0}^\infty\gamma^t\left\|\mathbf{y}_{(i,:)}^t-\frac{1}{N}\sum_{j=1}^N\nabla f_j(\mathbf{x}_{(j,:)}^t)\right\|<\infty. \tag{35}$$

Proof. It is sufficient to prove (34) and (35) for each block ℓ. Notice that the gradient tracking scheme given by (12), (15) and (16) is an instance of the perturbed push-sum consensus (21), with the identifications: $n = d$, $\psi_i^t \triangleq \phi_{(i,\ell)}^t$, $\boldsymbol{\eta}_i^t \triangleq \boldsymbol{\sigma}_{(i,\ell)}^t$, $\mathbf{z}_i^t \triangleq \mathbf{y}_{(i,\ell)}^t$, and $\boldsymbol{\epsilon}_{i,\ell}^t \triangleq \nabla_\ell f_i(\mathbf{x}_{(i,:)}^{t+1}) - \nabla_\ell f_i(\mathbf{x}_{(i,:)}^t)$. Therefore, (34) follows readily from Lemma A.3(2) and (20), once we have shown $\lim_{t\to\infty} \|\boldsymbol{\epsilon}_{i,\ell}^t\| = 0$, as proven next.

Since each $\nabla_\ell f_i$ is Lipschitz continuous [cf. Assumption IV.1(iii)], it suffices to prove $\lim_{t\to\infty} \|\mathbf{x}_{(i,:)}^{t+1} - \mathbf{x}_{(i,:)}^t\| = 0$. We have:

$$\begin{aligned}\left\|\mathbf{x}_{(i,:)}^{t+1} - \mathbf{x}_{(i,:)}^t\right\| &\overset{(a)}{\leq} \left\|\mathbf{x}_{(i,:)}^{t+1} - \bar{\boldsymbol{s}}^{t+1}\right\| + \left\|\mathbf{x}_{(i,:)}^t - \bar{\boldsymbol{s}}^t\right\| \\ &\quad + \frac{1}{N}\sum_{\ell=1}^B \sum_{i=1}^N \left\|\gamma^t \phi_{(i,\ell)}^t \Delta \mathbf{x}_{(i,\ell)}^t\right\| \\ &\overset{(b)}{\leq} \left\|\mathbf{x}_{(i,:)}^{t+1} - \bar{\boldsymbol{s}}^{t+1}\right\| + \left\|\mathbf{x}_{(i,:)}^t - \bar{\boldsymbol{s}}^t\right\| \\ &\quad + \gamma^t \sum_{\ell=1}^B \sum_{i=1}^N \left\|\Delta \mathbf{x}_{(i,\ell)}^t\right\|,\end{aligned} \tag{36}$$

where in (a) we used (17) while (b) follows from $\phi_{(i,\ell)}^t \leq N$. The desired result, $\lim_{t\to\infty} \|\mathbf{x}_{(i,:)}^{t+1} - \mathbf{x}_{(i,:)}^t\| = 0$, follows readily from (36), Proposition B.9 [cf. eq. (30)], Lemma B.8(2), and $\gamma^t \downarrow 0$ [cf. Assumption IV.3].

We prove now (35). Invoking Lemma A.3(2), we can write

$$\begin{aligned}&\sum_{t=0}^\infty \gamma^{t+1} \left\|\mathbf{y}_{(i,\ell)}^{t+1} - \frac{1}{N}\sum_{j=1}^N \nabla_\ell f_j(\mathbf{x}_{(j,:)}^{t+1})\right\| \\ &= \sum_{t=0}^\infty \gamma^{t+1} \left\|\mathbf{y}_{(i,\ell)}^{t+1} - \bar{\boldsymbol{\sigma}}_\ell^{t+1}\right\| \\ &\leq \sum_{t=0}^\infty \gamma^{t+1} \Big(c_1(\rho)^t \\ &\qquad + c_2 \sum_{h=1}^t (\rho)^{t-h} \sum_{i=1}^N \left\|\nabla_\ell f_i(\mathbf{x}_{(i,:)}^{h+1}) - \nabla_\ell f_i(\mathbf{x}_{(i,:)}^h)\right\| \Big) \\ &\leq \sum_{t=0}^\infty \gamma^{t+1} \Big(c_1(\rho)^t + c_4 \sum_{h=1}^t (\rho)^{t-h} \sum_{i=1}^N \left\|\mathbf{x}_{(i,:)}^{h+1} - \mathbf{x}_{(i,:)}^h\right\| \Big) \\ &\overset{(36)}{\leq} c_1 \sum_{t=0}^\infty \gamma^{t+1} (\rho)^t + c_4 \sum_{t=0}^\infty \gamma^{t+1} \sum_{h=1}^t (\rho)^{t-h} \sum_{i=1}^N \left\|\mathbf{x}_{(i,:)}^{h+1} - \bar{\boldsymbol{s}}^{h+1}\right\| \\ &+ c_4 \sum_{t=0}^\infty \gamma^{t+1} \sum_{h=1}^t (\rho)^{t-h} \sum_{i=1}^N \left\|\mathbf{x}_{(i,:)}^h - \bar{\boldsymbol{s}}^h\right\| \\ &+ c_5 \sum_{t=0}^\infty \gamma^{t+1} \sum_{h=1}^t (\rho)^{t-h} \gamma^h \overset{(a)}{<} \infty,\end{aligned}$$

for some positive, finite scalars c_4 and c_5, where (a) follows from [26, Lemma 7]. □

C. Lyapunov Function and its Descent Property

We begin introducing the following lemma that is instrumental for the rest of the proof.

Lemma C.11. *Consider problem* (1) *under Assumptions II.1, II.2, III.1, IV.1, IV.3; and let* $\{\phi_{(i,:)}^t\}_{t\geq 0}$ *and* $\{\mathbf{x}_{(i,:)}^t\}_{t\geq 0}$ *be the sequences generated by* BLOCK-SONATA. *Then, for all* $\ell \in \{1, \ldots, B\}$, *it holds*

$$\begin{aligned}&\sum_{i=1}^N \phi_{(i,\ell)}^{t+1}\, r_\ell(\mathbf{x}_{(i,\ell)}^{t+1}) - \sum_{i=1}^N \phi_{(i,\ell)}^t\, r_\ell(\mathbf{x}_{(i,\ell)}^t) \\ &\leq \gamma^t \frac{1}{N} \sum_{i=1}^N \phi_{(i,\ell)}^t \Big(r_\ell(\widetilde{\mathbf{x}}_{(i,\ell)}^t) - r_\ell(\mathbf{x}_{(i,\ell)}^t) \Big).\end{aligned}$$

Proof. The proof follows readily from the convexity of r_ℓ and the column stochasticity of $\mathbf{A}_\ell^t$. □

We are now ready to introduce our Lyapunov-like function: given $\bar{\boldsymbol{s}}_\ell^t$, $(\mathbf{x}_{(i,\ell)}^t)_{i=1}^N$, and $(\phi_{(i,\ell)}^t)_{i=1}^N$, define (we omit the dependence on the algorithm variables for notational simplicity)

$$V^t \triangleq \sum_{i=1}^N f_i(\bar{\boldsymbol{s}}^{t+1}) + \sum_{\ell=1}^B \sum_{i=1}^N \phi_{(i,\ell)}^t\, r_\ell(\mathbf{x}_{(i,\ell)}^t).$$

The descent properties of the above function along the trajectory of the algorithm are studied in the following proposition.

Proposition C.12. *Consider problem* (1)*, under Assumptions II.1, II.2, III.1, IV.1, IV.3; and let* $\{(\phi_{(i,:)}^t)_{i=1}^N\}_{t\geq 0}$, $\{\bar{\boldsymbol{s}}^t\}_{t\geq 0}$, *and* $\{(\mathbf{x}_{(i,:)}^t)_{i=1}^N\}_{t\geq 0}$ *be the sequences generated by* BLOCK-SONATA. *Then* $\{V^t\}_{t\geq 0}$ *satisfies:*

$$V^{t+1} \leq V^t - c_7 \sum_{\ell=1}^B \sum_{i=1}^N \gamma^t \|\Delta \mathbf{x}_{(i,\ell)}^t\|^2 + P^t, \tag{37}$$

with $\sum_{t=0}^\infty P^t < \infty$, *where* P^t *is defined as*

$$P^t \triangleq c_8\, \gamma^t \sum_{\ell=1}^B \sum_{i=1}^N \left\|\frac{1}{N}\sum_{j=1}^N \nabla_\ell f_j(\bar{\boldsymbol{s}}^t) - \mathbf{y}_{(i,\ell)}^t\right\| + c_6\, (\gamma^t)^2,$$

and c_6, c_7, *and* c_8 *are some positive, finite scalars.*

Proof. Applying the descent lemma to (17), with $L = \sum_{i=1}^N L_i$, yields

$$\begin{aligned}&\sum_{i=1}^N f_i(\bar{\boldsymbol{s}}^{t+1}) \\ &\leq \sum_{i=1}^N f_i(\bar{\boldsymbol{s}}^t) + \Big(\sum_{j=1}^N \nabla f_j(\bar{\boldsymbol{s}}^t)\Big)^\top \Big(\bar{\boldsymbol{s}}^{t+1} - \bar{\boldsymbol{s}}^t\Big) + \frac{L}{2}\|\bar{\boldsymbol{s}}^{t+1} - \bar{\boldsymbol{s}}^t\|^2 \\ &\leq \sum_{i=1}^N f_i(\bar{\boldsymbol{s}}^t) + \sum_{\ell=1}^B \Big(\sum_{j=1}^N \nabla_\ell f_j(\bar{\boldsymbol{s}}^t)\Big)^\top \Big(\frac{\gamma^t}{N}\sum_{i=1}^N \phi_{(i,\ell)}^t \Delta \mathbf{x}_{(i,\ell)}^t\Big) \\ &\quad + \frac{L}{2}\sum_{\ell=1}^B \Big\|\frac{1}{N}\gamma^t \sum_{i=1}^N \phi_{(i,\ell)}^t\, \Delta \mathbf{x}_{(i,\ell)}^t\Big\|^2 \\ &\overset{(a)}{\leq} \sum_{i=1}^N f_i(\bar{\boldsymbol{s}}^t) + \gamma^t \sum_{\ell=1}^B \sum_{i=1}^N \phi_{(i,\ell)}^t \big(\mathbf{y}_{(i,\ell)}^t\big)^\top \Delta \mathbf{x}_{(i,\ell)}^t \\ &\quad + \gamma^t \sum_{\ell=1}^B \sum_{i=1}^N \phi_{(i,\ell)}^t \Big(\frac{1}{N}\sum_{j=1}^N \nabla_\ell f_j(\bar{\boldsymbol{s}}^t) - \mathbf{y}_{(i,\ell)}^t\Big)^\top \Delta \mathbf{x}_{(i,\ell)}^t \\ &\quad + (\gamma^t)^2 \frac{L}{2} \sum_{\ell=1}^B \sum_{i=1}^N \frac{\phi_{(i,\ell)}^t}{N} \|\Delta \mathbf{x}_{(i,\ell)}^t\|^2, \\ &\overset{(b)}{\leq} \sum_{i=1}^N f_i(\bar{\boldsymbol{s}}^t) - \gamma^t h \sum_{\ell=1}^B \sum_{i=1}^N \phi_{(i,\ell)}^t \|\Delta \mathbf{x}_{(i,\ell)}^t\|^2 \\ &\quad - \gamma^t \sum_{\ell=1}^B \sum_{i=1}^N \phi_{(i,\ell)}^t \big(r_\ell(\widetilde{\mathbf{x}}_{(i,\ell)}^t) - r_\ell(\mathbf{x}_{(i,\ell)}^t)\big) \\ &\quad + \gamma^t \sum_{\ell=1}^B \sum_{i=1}^N \phi_{(i,\ell)}^t \Big\|\frac{1}{N}\sum_{j=1}^N \nabla_\ell f_j(\bar{\boldsymbol{s}}^t) - \mathbf{y}_{(i,\ell)}^t\Big\| \left\|\Delta \mathbf{x}_{(i,\ell)}^t\right\|\end{aligned}$$

$$+ c_6(\gamma^t)^2,$$

where in (a) we added and subtracted $\gamma^t \sum_{i=1}^N \sum_{\ell=1}^B \phi^t_{(i,\ell)}(\mathbf{y}^t_{(i,\ell)})^\top \Delta \mathbf{x}^t_{(i,\ell)}$; and in (b) we used Lemma A.6 [cf. (26)], Lemma B.8 [cf. (29)], we defined $\tau = \min_i \tau_i$, and c_6 is some positive, finite scalar.

Combining now the above chain of inequalities with Lemma C.11 and using Lemma A.3(3), we can write

$$V^{t+1} \leq V^t - c_7\,\gamma^t \sum_{\ell=1}^B \sum_{i=1}^N \|\Delta \mathbf{x}^t_{(i,\ell)}\|^2 + \underbrace{c_8 \gamma^t \sum_{\ell=1}^B \sum_{i=1}^N \Big\| \frac{1}{N} \sum_{j=1}^N \nabla_\ell f_j(\bar{\boldsymbol{s}}^t) - \mathbf{y}^t_{(i,\ell)} \Big\|}_{P^t} + c_6\,(\gamma^t)^2,$$

where c_7, and c_8 are some positive, finite scalars.

To conclude the proof, we show next that P^t is summable. Since $\sum_{t=0}^\infty (\gamma^t)^2 < \infty$ (cf. Assumption IV.3), it is sufficient to prove that the first term of P^t is summable, as shown below:

$$\begin{aligned}
&\lim_{k\to\infty} \sum_{t=0}^k \gamma^t \sum_{\ell=1}^B \sum_{i=1}^N \Big\| \frac{1}{N} \sum_{j=1}^N \nabla_\ell f_j(\bar{\boldsymbol{s}}^t) - \mathbf{y}^t_{(i,\ell)} \Big\| \\
&\overset{(a)}{\leq} \lim_{k\to\infty} \sum_{t=0}^k \gamma^t \sum_{\ell=1}^B \sum_{i=1}^N \Big\| \frac{1}{N} \sum_{j=1}^N \nabla_\ell f_j(\mathbf{x}^t_{(j,:)}) - \mathbf{y}^t_{(i,\ell)} \Big\| \\
&\quad + c_9 \lim_{k\to\infty} \sum_{t=0}^k \gamma^t \sum_{\ell=1}^B \sum_{i=1}^N \big\| \mathbf{x}^t_{(i,:)} - \bar{\boldsymbol{s}}^t \big\| \overset{(b)}{<} \infty,
\end{aligned}$$

where in (a) we used the Lipschitz continuity of ∇f_i; (b) follows from Prop. B.9 and B.10 with c_9 positive scalar. □

D. Asymptotic Convergence of $\{\bar{\mathbf{s}}^t\}_{t\geq 0}$

Since U is coercive and $\sum_{t=0}^\infty P^t < \infty$, (37) implies that i) $\{V^t\}_{\geq 0}$ is convergent; and ii) and $\{\bar{\mathbf{s}}^t\}_{t\geq 0}$ is bounded. Therefore, it must be

$$\sum_{t=0}^\infty \sum_{i=1}^N \sum_{\ell=1}^B \gamma^t \|\Delta \mathbf{x}^t_{(i,\ell)}\|^2 < \infty. \tag{38}$$

Recall that agents select their blocks to update according to an essential cyclic rule [cf. Assumption III.2]. This means that in any time window $[t, t+T-1]$, with $T > 0$ defined in Proposition III.3, any agent i selects all of its blocks at least once. Denote by $t + s^t_i(\ell)$ the last time agent i selects block ℓ in the time window $[t, t+T-1]$; notice that such a $s^t_i(\ell)$ is always well-defined and $s^t_i(\ell) \in [0, T-1]$. Finally, let

$$\boldsymbol{\Delta}^t \triangleq \sum_{i=1}^N \sum_{\ell=1}^B \|\Delta \mathbf{x}^{t+s^t_i(\ell)}_{(i,\ell)}\|. \tag{39}$$

The above quantity will play a key role to prove (subsequence) convergence of $\{\bar{\mathbf{s}}^t\}_{t\geq 0}$. We organize the rest of the proof in the following steps:

- **Step 1:** We prove $\lim_{t\to\infty} \boldsymbol{\Delta}^t = 0$, by showing that, first, $\liminf_{t\to\infty} \boldsymbol{\Delta}^t = 0$ [Step 1(a)], and, second, $\limsup_{t\to\infty} \boldsymbol{\Delta}^t = 0$ [Step 1(b)];
- **Step 2:** Using results in Step 1, we prove that every limit point of $\{\bar{\mathbf{s}}^t\}_{t\geq 0}$ is a stationary solution of problem (1).

Step 1(a) – $\liminf_{t\to\infty} \boldsymbol{\Delta}^t = 0$. For all $t \geq T-1$, we have

$$\begin{aligned}
&T \cdot \sum_{h=0}^t \sum_{\ell=1}^B \sum_{i=1}^N \gamma^h \|\Delta \mathbf{x}^h_{(i,\ell)}\|^2 \\
&\quad \geq \sum_{h=0}^{t-T+1} \sum_{s=0}^{T-1} \sum_{\ell=1}^B \sum_{i=1}^N \gamma^{h+s} \|\Delta \mathbf{x}^{h+s}_{(i,\ell)}\|^2 \\
&\quad \overset{(a)}{\geq} \sum_{h=0}^{t-T+1} \gamma^{h+T-1} \sum_{s=0}^{T-1} \sum_{\ell=1}^B \sum_{i=1}^N \|\Delta \mathbf{x}^{h+s}_{(i,\ell)}\|^2,
\end{aligned} \tag{40}$$

where (a) follows from Assumption IV.3(i). Using (38) and $\sum_{t=0}^\infty \gamma^t = \infty$, we deduce

$$\liminf_{t\to\infty} \sum_{s=0}^{T-1} \sum_{\ell=1}^B \sum_{i=1}^N \|\Delta \mathbf{x}^{t+s}_{(i,\ell)}\| = 0,$$

which leads to

$$0 = \liminf_{t\to\infty} \sum_{\ell=1}^B \sum_{i=1}^N \sum_{s=0}^{T-1} \|\Delta \mathbf{x}^{t+s}_{(i,\ell)}\| \geq \liminf_{t\to\infty} \boldsymbol{\Delta}^t.$$

Step 1(b) – $\limsup_{t\to\infty} \boldsymbol{\Delta}^t = 0$. We begin stating the following lemma, which proves that the best-response maps $\widetilde{\mathbf{x}}_{(i,\ell^t_i)}$ [cf. (6)] and $\widehat{\mathbf{x}}_{(i,\ell^t_i)}$ [cf. (27)], are asymptotically consistent along the trajectory of the algorithm.

Lemma D.13. *In the setting of* BLOCK-SONATA*, the best-response maps* $\widehat{\mathbf{x}}_{(i,\ell^t_i)}$ *and* $\widetilde{\mathbf{x}}^t_{(i,\ell^t_i)}$ *satisfy*

$$\lim_{t\to\infty} \Big\| \widehat{\mathbf{x}}_{(i,\ell^t_i)}\big(\mathbf{x}^t_{(i,:)}\big) - \widetilde{\mathbf{x}}^t_{(i,\ell^t_i)} \Big\| = 0, \quad \forall\, i \in \{1,\dots,N\}. \tag{41}$$

Proof. We use the shorthand $\widehat{\mathbf{x}}^t_{(i,\ell^t_i)}$ for $\widehat{\mathbf{x}}_{(i,\ell^t_i)}\big(\mathbf{x}^t_{(i,:)}\big)$. Invoking the optimality conditions of $\widehat{\mathbf{x}}_{(i,\ell^t_i)}(\mathbf{x}^t_{(i,:)})$ and $\widetilde{\mathbf{x}}^t_{(i,\ell^t_i)}$ yields

$$\begin{aligned}
&\big(\widetilde{\mathbf{x}}^t_{(i,\ell^t_i)} - \widehat{\mathbf{x}}^t_{(i,\ell^t_i)}\big)^\top \times \\
&\Big(\nabla_{\ell^t_i} \widehat{f}_{i,\ell^t_i}\big(\widehat{\mathbf{x}}^t_{(i,\ell^t_i)}; \mathbf{x}^t_{(i,:)}, \tfrac{1}{N} \sum_{j=1}^N \nabla_{\ell^t_i} f_j(\mathbf{x}^t_{(i,:)})\big) \\
&\quad + \widetilde{\nabla} r_{\ell^t_i}\big(\widehat{\mathbf{x}}^t_{(i,\ell^t_i)}\big)\Big) \geq 0,
\end{aligned} \tag{42}$$

and

$$\begin{aligned}
&\big(\widehat{\mathbf{x}}^t_{(i,\ell^t_i)} - \widetilde{\mathbf{x}}^t_{(i,\ell^t_i)}\big)^\top \times \\
&\Big(\nabla_{\ell^t_i} \widehat{f}_{i,\ell^t_i}(\widetilde{\mathbf{x}}^t_{(i,\ell^t_i)}; \mathbf{x}^t_{(i,:)}, \mathbf{y}^t_{(i,\ell^t_i)}) + \widetilde{\nabla} r_{\ell^t_i}\big(\widetilde{\mathbf{x}}^t_{(i,\ell^t_i)}\big)\Big) \geq 0.
\end{aligned} \tag{43}$$

Adding the two inequalities (42) and (43) and using the strong convexity of $\tilde{f}_i(\bullet; \mathbf{x}^t_{(i,:)})$ as well as the convexity of $r_{\ell^t_i}$, yields

$$\begin{aligned}
\Big\| \widetilde{\mathbf{x}}^t_{(i,\ell^t_i)} - \widehat{\mathbf{x}}^t_{(i,\ell^t_i)} \Big\| &\leq \frac{1}{\tau_i} \Big\| \frac{1}{N} \sum_{j=1}^N \nabla_{\ell^t_i} f_j(\mathbf{x}^t_{(i,:)}) - \mathbf{y}^t_{(i,\ell^t_i)} \Big\| \\
&\leq \frac{1}{\tau_i} \Big\| \frac{1}{N} \sum_{j=1}^N \nabla_{\ell^t_i} f_j(\mathbf{x}^t_{(i,:)}) - \frac{1}{N} \sum_{j=1}^N \nabla_{\ell^t_i} f_j(\mathbf{x}^t_{(j,:)}) \Big\| \\
&\quad + \frac{1}{\tau_i} \Big\| \frac{1}{N} \sum_{j=1}^N \nabla_{\ell^t_i} f_j(\mathbf{x}^t_{(j,:)}) - \mathbf{y}^t_{(i,\ell^t_i)} \Big\| \\
&\leq \frac{1}{\tau_i N} \sum_{j=1}^N L_j \Big\| \mathbf{x}^t_{(i,:)} - \mathbf{x}^t_{(j,:)} \Big\| \\
&\quad + \frac{1}{\tau_i} \Big\| \frac{1}{N} \sum_{j=1}^N \nabla_{\ell^t_i} f_j(\mathbf{x}^t_{(j,:)}) - \mathbf{y}^t_{(i,\ell^t_i)} \Big\| \\
&\leq \frac{1}{\tau_i N} \sum_{j=1}^N L_j \Big(\Big\| \mathbf{x}^t_{(i,:)} - \bar{\boldsymbol{s}}^t \| + \| \bar{\boldsymbol{s}}^t - \mathbf{x}^t_{(j,:)} \Big\| \Big)
\end{aligned}$$

$$+\frac{1}{\tau_i}\Big\|\frac{1}{N}\sum_{j=1}^{N}\nabla_{\ell_i^t} f_j(\mathbf{x}_{(j,:)}^t)-\mathbf{y}_{(i,\ell_i^t)}^t\Big\|.$$

Finally, by noticing that

$$\begin{aligned}&\operatorname*{limsup}_{t\to\infty}\Big\|\frac{1}{N}\sum_{i=1}^{N}\nabla_{\ell_i^t} f_i(\mathbf{x}_{(i,:)}^t)-\mathbf{y}_{(i,\ell_i^t)}^t\Big\|\\&\le\operatorname*{limsup}_{t\to\infty}\sum_{\ell=1}^{B}\Big\|\frac{1}{N}\sum_{i=1}^{N}\nabla_{\ell} f_i(\mathbf{x}_{(i,:)}^t)-\mathbf{y}_{(i,\ell)}^t\Big\|,\end{aligned}$$

and invoking Propositions B.9 and B.10, we obtain the desired result $\operatorname{limsup}_{t\to\infty}\|\widehat{\mathbf{x}}_{(i,\ell_i^t)}-\widetilde{\mathbf{x}}_{(i,\ell_i^t)}^t\|=0$. □

Now we prove by contradiction that $\operatorname{limsup}_{t\to\infty}\boldsymbol{\Delta}^t=0$. Suppose $\operatorname{limsup}_{t\to\infty}\boldsymbol{\Delta}^t>0$. Since $\operatorname{liminf}_{t\to\infty}\boldsymbol{\Delta}^t=0$, there exists a $\delta>0$ such that $\boldsymbol{\Delta}^t<\delta$ for infinitely many t and also $\boldsymbol{\Delta}^t>2\delta$ for infinitely many t. Therefore, one can always find an infinite set of indices, say $\mathcal{T}$, having the following property: for any $t\in\mathcal{T}$, there exists an integer $\theta_t>t$ such that

$$\begin{aligned}&\boldsymbol{\Delta}^t\le\delta,\ \ \boldsymbol{\Delta}^{\theta_t}\ge 2\delta,\\&\delta<\boldsymbol{\Delta}^h<2\delta,\qquad\quad t<h<\theta_t.\end{aligned}\tag{44}$$

Therefore, for all $t\in\mathcal{T}$, we have

$$\begin{aligned}\delta&\le\boldsymbol{\Delta}^{\theta_t}-\boldsymbol{\Delta}^t\\&=\sum_{i=1}^{N}\sum_{\ell=1}^{B}\Big(\Big\|\widetilde{\mathbf{x}}_{(i,\ell)}^{\theta_t+s_i^{\theta_t}(\ell)}-\mathbf{x}_{(i,\ell)}^{\theta_t+s_i^{\theta_t}(\ell)}\Big\|-\Big\|\widetilde{\mathbf{x}}_{(i,\ell)}^{t+s_i^{t}(\ell)}-\mathbf{x}_{(i,\ell)}^{t+s_i^{t}(\ell)}\Big\|\Big)\\&\le\sum_{i=1}^{N}\sum_{\ell=1}^{B}\Big(\Big\|\widetilde{\mathbf{x}}_{(i,\ell)}^{\theta_t+s_i^{\theta_t}(\ell)}-\widetilde{\mathbf{x}}_{(i,\ell)}^{t+s_i^{t}(\ell)}\Big\|+\Big\|\mathbf{x}_{(i,\ell)}^{\theta_t+s_i^{\theta_t}(\ell)}-\mathbf{x}_{(i,\ell)}^{t+s_i^{t}(\ell)}\Big\|\Big)\\&\le\sum_{i=1}^{N}\sum_{\ell=1}^{B}\Big(\Big\|\widetilde{\mathbf{x}}_{(i,\ell)}^{\theta_t+s_i^{\theta_t}(\ell)}-\widehat{\mathbf{x}}_{(i,\ell)}(\mathbf{x}_{(i,:)}^{\theta_t+s_i^{\theta_t}(\ell)})\Big\|\\&\qquad+\Big\|\widehat{\mathbf{x}}_{(i,\ell)}(\mathbf{x}_{(i,:)}^{\theta_t+s_i^{\theta_t}(\ell)})-\widehat{\mathbf{x}}_{(i,\ell)}(\mathbf{x}_{(i,:)}^{t+s_i^{t}(\ell)})\Big\|\\&\qquad+\Big\|\widehat{\mathbf{x}}_{(i,\ell)}(\mathbf{x}_{(i,:)}^{t+s_i^{t}(\ell)})-\widetilde{\mathbf{x}}_{(i,\ell)}^{t+s_i^{t}(\ell)}\Big\|\\&\qquad+\Big\|\mathbf{x}_{(i,\ell)}^{\theta_t+s_i^{\theta_t}(\ell)}-\mathbf{x}_{(i,\ell)}^{t+s_i^{t}(\ell)}\Big\|\Big)\\&\le(1+\widehat{L})\sum_{i=1}^{N}\sum_{\ell=1}^{B}\Big\|\mathbf{x}_{(i,:)}^{\theta_t+s_i^{\theta_t}(\ell)}-\mathbf{x}_{(i,:)}^{t+s_i^{t}(\ell)}\Big\|+e_1^t,\end{aligned}\tag{45}$$

where in the last inequality, we used the Lipschitz continuity of $\widehat{\mathbf{x}}_{(i,\ell)}(\bullet)$ [cf. Lemma. A.7], with $\widehat{L}\triangleq\max_i\max_\ell\widehat{L}_{i,\ell}$, and

$$\begin{aligned}e_1^t\triangleq\sum_{i=1}^{N}\sum_{\ell=1}^{B}\Big(&\Big\|\widehat{\mathbf{x}}_{(i,\ell)}(\mathbf{x}_{(i,:)}^{\theta_t+s_i^{\theta_t}(\ell)})-\widetilde{\mathbf{x}}_{(i,\ell)}^{\theta_t+s_i^{\theta_t}(\ell)}\Big\|\\&+\Big\|\widehat{\mathbf{x}}_{(i,\ell)}(\mathbf{x}_{(i,:)}^{t+s_i^{t}(\ell)})-\widetilde{\mathbf{x}}_{(i,\ell)}^{t+s_i^{t}(\ell)}\Big\|\Big).\end{aligned}\tag{46}$$

Adding and subtracting $\bar{\boldsymbol{s}}^{\theta_t+s_i^{\theta_t}(\ell)}$ and $\bar{\boldsymbol{s}}^{t+s_i^{t}(\ell)}$ in the first term of the last inequality in (45), and introducing

$$\begin{aligned}e_2^t\triangleq(1+\widehat{L})\sum_{i=1}^{N}\sum_{\ell=1}^{B}\Big(&\Big\|\mathbf{x}_{(i,:)}^{\theta_t+s_i^{\theta_t}(\ell)}-\bar{\boldsymbol{s}}^{\theta_t+s_i^{\theta_t}(\ell)}\Big\|\\&+\Big\|\bar{\boldsymbol{s}}^{t+s_i^{t}(\ell)}-\mathbf{x}_{(i,:)}^{t+s_i^{t}(\ell)}\Big\|\Big),\end{aligned}\tag{47}$$

we can write:

$$\delta\le(1+\widehat{L})\sum_{i=1}^{N}\sum_{\ell=1}^{B}\Big\|\bar{\boldsymbol{s}}^{\theta_t+s_i^{\theta_t}(\ell)}-\bar{\boldsymbol{s}}^{t+s_i^{t}(\ell)}\Big\|+e_1^t+e_2^t.\tag{48}$$

Since $\theta_t+s_i^{\theta_t}(\ell)$ is the *last* time at which block ℓ has been updated by agent i in $[\theta_t,\theta_t+T-1]$ and $\theta_t>t$, it must hold: $\theta_t+s_i^{\theta_t}(\ell)\ge t+s_i^t(\ell)$, for all $t\in\mathcal{T}$. We assume, without loss of generality, that $\theta_t+s_i^{\theta_t}(\ell)>t+s_i^t(\ell)$, for all $i\in\{1,\dots,N\}$ and $\ell\in\{1,\dots,B\}$. Hence, all the intervals $[t+s_i^t(\ell),\theta_t+s_i^{\theta_t}(\ell)]$ are nonempty. Using (17) to bound $\|\bar{\boldsymbol{s}}^{\theta_t+s_i^{\theta_t}(\ell)}-\bar{\boldsymbol{s}}^{t+s_i^t(\ell)}\|$ in (48), we can write

$$\begin{aligned}\delta&\le c_{10}\sum_{i=1}^{N}\sum_{\ell=1}^{B}\sum_{\ell'=1}^{B}\sum_{h=t+s_i^t(\ell)}^{\theta_t+s_i^{\theta_t}(\ell)-1}\sum_{j=1}^{N}\gamma^h\|\Delta\mathbf{x}_{(j,\ell')}^h\|+e_1^t+e_2^t\\&\le c_{11}\sum_{h=t}^{\theta_t+T-1}\sum_{j=1}^{N}\sum_{\ell=1}^{B}\gamma^h\|\Delta\mathbf{x}_{(j,\ell)}^h\|+e_1^t+e_2^t\\&=c_{11}\sum_{h=t}^{\theta_t+T-1}\sum_{i=1}^{N}\gamma^h\|\Delta\mathbf{x}_{(i,\ell_i^h)}^h\|+e_1^t+e_2^t,\end{aligned}$$

for some positive, finite scalars c_{10} and c_{11}, where the last equality follows from the fact that, at time t, agent i optimizes only block ℓ_i^t, implying $\|\Delta\mathbf{x}_{(i,\ell)}^t\|=0$, for all $\ell\neq\ell_i^t$.

Note that, for each $i\in\{1,\dots,N\}$, $t+T-1$ is the last time agent i selects block ℓ_i^{t+T-1} in the interval $[t,t+T-1]$. Therefore, for all $i\in\{1,\dots,N\}$,

$$\Big\|\Delta\mathbf{x}_{(i,\ell_i^{t+T-1})}^{t+T-1}\Big\|=\Big\|\Delta\mathbf{x}_{(i,\ell_i^{t+T-1})}^{t+s_i^t(\ell_i^{t+T-1})}\Big\|\overset{(39)}{\le}\boldsymbol{\Delta}^t.$$

Hence, we can write

$$\begin{aligned}\delta&\le c_{12}\Big(\sum_{h=t}^{t+T-1}\gamma^h\sum_{i=1}^{N}\|\Delta\mathbf{x}_{(i,\ell_i^h)}^h\|+\sum_{h=t+T}^{\theta_t+T-1}\gamma^h\boldsymbol{\Delta}^{h-T+1}\Big)\\&\quad+e_1^t+e_2^t\\&=c_{12}\sum_{h=t+1}^{\theta_t}\gamma^{h+T-1}\boldsymbol{\Delta}^h+e_1^t+e_2^t+e_3^t,\end{aligned}$$

for some positive, finite scalar c_{12}, where we set

$$e_3^t\triangleq c_{12}\sum_{h=t}^{t+T-1}\gamma^h\sum_{i=1}^{N}\|\Delta\mathbf{x}_{(i,\ell_i^h)}^h\|.\tag{49}$$

Since $\boldsymbol{\Delta}^h>\delta$, for $h\in[t+1,\theta_t]$ [cf. (44)], we have

$$\begin{aligned}\delta&\le c_{12}\sum_{h=t+1}^{\theta_t}\gamma^{h+T-1}\sum_{i=1}^{N}\sum_{\ell=1}^{B}\|\Delta\mathbf{x}_{(i,\ell)}^{h+s_i^h(\ell)}\|+e_1^t+e_2^t+e_3^t\\&\le\frac{c_{11}}{\delta}\sum_{h=t+1}^{\theta_t}\gamma^{h+T-1}\left(\sum_{i=1}^{N}\sum_{\ell=1}^{B}\Big\|\Delta\mathbf{x}_{(i,\ell)}^{h+s_i^h(\ell)}\Big\|\right)^2+e_1^t+e_2^t+e_3^t\\&\le c_{13}\sum_{h=t+1}^{\theta_t}\gamma^{h+T-1}\sum_{i=1}^{N}\sum_{\ell=1}^{B}\|\Delta\mathbf{x}_{(i,\ell)}^{h+s_i^h(\ell)}\|^2+e_1^t+e_2^t+e_3^t\\&\le c_{13}\sum_{h=t+1}^{\theta_t}\gamma^{h+T-1}\sum_{s=0}^{T-1}\sum_{i=1}^{N}\sum_{\ell=1}^{B}\|\Delta\mathbf{x}_{(i,\ell)}^{h+s}\|^2+e_1^t+e_2^t+e_3^t,\end{aligned}\tag{50}$$

for some positive, finite scalar c_{13}. Using now Proposition B.9, Lemma D.13, and (38), we infer that e_1^t, e_2^t, and e_3^t [defined in (46), (47), and (49), respectively] are asymptotically vanishing, that is, $e_1^t,e_2^t,e_3^t\underset{t\to\infty}{\longrightarrow}0$. Furthermore, since [due to (38) and (40)]

$$\sum_{h=0}^{\infty}\gamma^{h+T-1}\sum_{s=0}^{T-1}\sum_{\ell=1}^{B}\sum_{i=1}^{N}\|\Delta\mathbf{x}_{(i,\ell)}^{h+s}\|^2<\infty,$$

it must be

$$\lim_{t\to\infty}\sum_{h=t+1}^{\theta_t}\gamma^{h+T-1}\sum_{s=0}^{T-1}\sum_{i=1}^{N}\sum_{\ell=1}^{B}\|\Delta\mathbf{x}_{(i,\ell)}^{h+s}\|^2=0.$$

Therefore there must exist a sufficient large $\bar{t}\in\mathcal{T}$ such that

$$c_{13}\sum_{h=t+1}^{\theta_t}\gamma^{h+T-1}\sum_{s=0}^{T-1}\sum_{i=1}^{N}\sum_{\ell=1}^{B}\|\Delta\mathbf{x}_{(i,\ell)}^{h+s}\|^2+e_1^t+e_2^t+e_3^t\le\frac{\delta}{2}$$

for all $t\ge\bar{t}$, which contradicts (50). Thus, it must be $\limsup_{t\to\infty}\boldsymbol{\Delta}^t=0$, and hence

$$\lim_{t\to\infty}\sum_{i=1}^{N}\sum_{\ell=1}^{B}\|\Delta\mathbf{x}_{(i,\ell)}^{t+s_i^t(\ell)}\|=0. \tag{51}$$

Step 2 – Every limit point of $\{\bar{\mathbf{s}}^t\}_{t\ge0}$ is stationary for (1). Let $\bar{\mathbf{s}}^\infty$ be a limit point of $\{\bar{\mathbf{s}}^t\}_{t\ge0}$; note that such a point exists, because $\{\bar{\mathbf{s}}^t\}_{t\ge0}$ is bounded (cf. Section D). By Lemma A.7, $\bar{\mathbf{s}}^\infty$ is a stationary solution of problem (1), if

$$\lim_{t\to\infty}\left\|\widehat{\mathbf{x}}_{(i,:)}\big(\bar{\boldsymbol{s}}^t\big)-\bar{\boldsymbol{s}}^t\right\|=0, \tag{52}$$

for all $i\in\{1,\dots,N\}$. To prove (52), we first bound $\|\widehat{\mathbf{x}}_{(i,:)}(\bar{\boldsymbol{s}}^t)-\bar{\boldsymbol{s}}^t\|$ as follows

$$\begin{aligned}
\left\|\widehat{\mathbf{x}}_{(i,:)}\big(\bar{\boldsymbol{s}}^t\big)-\bar{\boldsymbol{s}}^t\right\| &\le \sum_{\ell=1}^{B}\Big(\left\|\widehat{\mathbf{x}}_{(i,\ell)}\big(\bar{\boldsymbol{s}}^t\big)-\widehat{\mathbf{x}}_{(i,\ell)}\big(\bar{\boldsymbol{s}}^{t+s_i^t(\ell)}\big)\right\|\\
&\quad+\left\|\widehat{\mathbf{x}}_{(i,\ell)}\big(\bar{\boldsymbol{s}}^{t+s_i^t(\ell)}\big)-\bar{\boldsymbol{s}}_\ell^{t+s_i^t(\ell)}\right\|\\
&\quad+\left\|\bar{\boldsymbol{s}}_\ell^{t+s_i^t(\ell)}-\bar{\boldsymbol{s}}_\ell^{t}\right\|\Big)\\
&\overset{(a)}{\le}\sum_{\ell=1}^{B}\Big(\left\|\widehat{\mathbf{x}}_{(i,\ell)}\big(\bar{\boldsymbol{s}}^{t+s_i^t(\ell)}\big)-\bar{\boldsymbol{s}}_\ell^{t+s_i^t(\ell)}\right\|\\
&\quad+(1+\hat{L})\left\|\bar{\boldsymbol{s}}^{t+s_i^t(\ell)}-\bar{\boldsymbol{s}}^{t}\right\|\Big)\\
&\le\sum_{\ell=1}^{B}\Big(\left\|\widehat{\mathbf{x}}_{(i,\ell)}\big(\bar{\boldsymbol{s}}^{t+s_i^t(\ell)}\big)-\widehat{\mathbf{x}}_{(i,\ell)}\big(\mathbf{x}_{(i,:)}^{t+s_i^t(\ell)}\big)\right\|\\
&\quad+\left\|\widehat{\mathbf{x}}_{(i,\ell)}\big(\mathbf{x}_{(i,:)}^{t+s_i^t(\ell)}\big)-\widetilde{\mathbf{x}}_{(i,\ell)}^{t+s_i^t(\ell)}\right\|\\
&\quad+\left\|\Delta\mathbf{x}_{(i,\ell)}^{t+s_i^t(\ell)}\right\|\\
&\quad+\left\|\mathbf{x}_{(i,\ell)}^{t+s_i^t(\ell)}-\bar{\boldsymbol{s}}_\ell^{t+s_i^t(\ell)}\right\|\\
&\quad+(1+\hat{L})\left\|\bar{\boldsymbol{s}}^{t+s_i^t(\ell)}-\bar{\boldsymbol{s}}^{t}\right\|\Big)\\
&\overset{(b)}{\le}\sum_{\ell=1}^{B}\Big(\left\|\widehat{\mathbf{x}}_{(i,\ell)}\big(\mathbf{x}_{(i,:)}^{t+s_i^t(\ell)}\big)-\widetilde{\mathbf{x}}_{(i,\ell)}^{t+s_i^t(\ell)}\right\|\\
&\quad+\left\|\Delta\mathbf{x}_{(i,\ell)}^{t+s_i^t(\ell)}\right\|\\
&\quad+(1+\widehat{L})\left\|\mathbf{x}_{(i,:)}^{t+s_i^t(\ell)}-\bar{\boldsymbol{s}}^{t+s_i^t(\ell)}\right\|\\
&\quad+(1+\hat{L})\left\|\bar{\boldsymbol{s}}^{t+s_i^t(\ell)}-\bar{\boldsymbol{s}}^{t}\right\|\Big)
\end{aligned}$$

where in (a) and (b) we used the Lipschitz continuity of $\widehat{\mathbf{x}}_{(i,\ell)}(\bullet)$. We show next that the four terms on the RHS of the above inequality are all asymptotically vanishing, which proves (52). Invoking Proposition B.9 [cf. (30)], we have

$$\lim_{t\to\infty}\left\|\mathbf{x}_{(i,\ell)}^{t+s_i^t(\ell)}-\bar{\boldsymbol{s}}_\ell^{t+s_i^t(\ell)}\right\|=0,$$

for all $\ell\in\{1,\dots,B\}$ and $i\in\{1,\dots,N\}$. By definition of $t+s_i^t(\ell)$, there exists some $\overline{\mathcal{T}}\subseteq\mathbb{N}_+$, with $|\overline{\mathcal{T}}|=\infty$, such that

$$\begin{aligned}
&\lim_{t\to\infty}\left\|\widehat{\mathbf{x}}_{(i,\ell)}\big(\mathbf{x}_{(i,:)}^{t+s_i^t(\ell)}\big)-\widetilde{\mathbf{x}}_{(i,\ell)}^{t+s_i^t(\ell)}\right\|\\
&\quad=\lim_{\overline{\mathcal{T}}\ni t\to\infty}\left\|\widehat{\mathbf{x}}_{(i,\ell_i^t)}\big(\mathbf{x}_{(i,:)}^{t}\big)-\widetilde{\mathbf{x}}_{(i,\ell_i^t)}^{t}\right\|\overset{(41)}{=}0,
\end{aligned}$$

for all $\ell\in\{1,\dots,B\}$ and $i\in\{1,\dots,N\}$. Using (51), we have $\lim_{t\to\infty}\|\Delta\mathbf{x}_{(i,\ell)}^{t+s_i^t(\ell)}\|=0$, which, together to (19), yields

$$\lim_{t\to\infty}\left\|\bar{\boldsymbol{s}}^{t+s_i^t(\ell)}-\bar{\boldsymbol{s}}^{t}\right\|=0,$$

for all $\ell\in\{1,\dots,B\}$ and $i\in\{1,\dots,N\}$, completing the proof.

REFERENCES

[1] I. Notarnicola, Y. Sun, G. Scutari, and G. Notarstefano, "Distributed big-data optimization via block-iterative convexification and averaging," in *IEEE Conf. on Decision and Control (CDC)*, 2017, pp. 2281–2288.

[2] ——, "Distributed big-data optimization via block communications," in *IEEE Intern. Conf. on Comput. Advances in Multi-Sensor Adaptive Process. (CAMSAP)*, 2017, pp. 557–561.

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

[4] P. Richtárik and M. Takáč, "Parallel coordinate descent methods for big data optimization," *Mathematical Programming*, pp. 1–52, 2012.

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

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

[7] H.-J. M. Shi, S. Tu, Y. Xu, and W. Yin, "A primer on coordinate descent algorithms," *arXiv preprint arXiv:1610.00040*, 2016.

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

[9] G. Scutari and Y. Sun, "Parallel and distributed successive convex approximation methods for big-data optimization," in *Multi-Agent Optimization*, F. Facchinei and J.-S. Pang, Eds. Springer, C.I.M.E. Foundation Subseries (Lecture Notes in Mathematics), 2018, pp. 1–158.

[10] A. Mokhtari, A. Koppel, and A. Ribeiro, "Doubly random parallel stochastic methods for large scale learning," in *IEEE American Control Conf. (ACC)*, 2016, pp. 4847–4852.

[11] A. Nedić and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Trans. on Autom. Control*, vol. 54, no. 1, pp. 48–61, 2009.

[12] A. Nedić, A. Ozdaglar, and P. A. Parrilo, "Constrained consensus and optimization in multi-agent networks," *IEEE Trans. on Autom. Control*, vol. 55, no. 4, pp. 922–938, 2010.

[13] P. Bianchi and J. Jakubowicz, "Convergence of a multi-agent projected stochastic gradient algorithm for non-convex optimization," *IEEE Trans. on Autom. Control*, vol. 58, no. 2, pp. 391–405, 2013.

[14] A. Nedić and A. Olshevsky, "Distributed optimization over time-varying directed graphs," *IEEE Trans. on Autom. Control*, vol. 60, no. 3, pp. 601–615, 2015.

[15] T. Tatarenko and B. Touri, "Non-convex distributed optimization," *IEEE Trans. on Autom. Control*, vol. 62, no. 8, pp. 3744–3757, 2017.

[16] F. Bénézit, V. Blondel, P. Thiran, J. Tsitsiklis, and M. Vetterli, "Weighted gossip: Distributed averaging using non-doubly stochastic matrices," in *IEEE Intern. Symposium on Information Theory (ISIT)*, 2010, pp. 1753–1757.

[17] D. Jakovetić, J. Xavier, and J. M. Moura, "Fast distributed gradient methods," *IEEE Trans. on Autom. Control*, vol. 59, no. 5, pp. 1131–1146, 2014.

[18] S. Lee and A. Nedić, "Asynchronous gossip-based random projection algorithms over networks," *IEEE Trans. on Autom. Control*, vol. 61, no. 4, pp. 953–968, 2016.

[19] K. Margellos, A. Falsone, S. Garatti, and M. Prandini, "Distributed constrained optimization and consensus in uncertain networks via proximal minimization," *IEEE Trans. on Autom. Control*, vol. 63, no. 5, pp. 1372–1387, 2018.

[20] W. Shi, Q. Ling, G. Wu, and W. Yin, "Extra: An exact first-order algorithm for decentralized consensus optimization," *SIAM J. on Optimization*, vol. 25, no. 2, pp. 944–966, 2015.

[21] ——, "A proximal gradient algorithm for decentralized composite optimization," *IEEE Trans. on Signal Process.*, vol. 63, no. 22, pp. 6013–6023, 2015.

[22] F. Zanella, D. Varagnolo, A. Cenedese, G. Pillonetto, and L. Schenato, "Newton-Raphson consensus for distributed convex optimization," in *IEEE Conf. on Decision and Control and European Control Conf. (CDC-ECC)*, 2011, pp. 5917–5922.

[23] ——, "Asynchronous Newton-Raphson consensus for distributed convex optimization," in *3rd IFAC Workshop on Distributed Estimation and Control in Networked Systems*, 2012.

[24] D. Varagnolo, F. Zanella, A. Cenedese, G. Pillonetto, and L. Schenato, "Newton-Raphson consensus for distributed convex optimization," *IEEE Trans. on Autom. Control*, vol. 61, no. 4, pp. 994–1009, 2016.

[25] P. Di Lorenzo and G. Scutari, "Distributed nonconvex optimization over networks," in *IEEE Intern. Conf. on Comput. Advances in Multi-Sensor Adaptive Process. (CAMSAP)*, 2015, pp. 229–232.

[26] ——, "NEXT: In-network nonconvex optimization," *IEEE Trans. on Signal and Information Process. over Networks*, vol. 2, no. 2, pp. 120–136, 2016.

[27] J. Xu, S. Zhu, Y. C. Soh, and L. Xie, "Augmented distributed gradient methods for multi-agent optimization under uncoordinated constant stepsizes," in *IEEE Conf. on Decision and Control (CDC)*, 2015, pp. 2055–2060.

[28] Y. Sun, G. Scutari, and D. Palomar, "Distributed nonconvex multiagent optimization over time-varying networks," in *IEEE Asilomar Conf. on Signals, Systems, and Computers*, 2016, pp. 788–794.

[29] Y. Sun and G. Scutari, "Distributed nonconvex optimization for sparse representation," in *IEEE Intern. Conf. on Speech and Signal Process. (ICASSP)*, 2017, pp. 4044–4048.

[30] J. Xu, S. Zhu, Y. C. Soh, and L. Xie, "Convergence of asynchronous distributed gradient methods over stochastic networks," *IEEE Trans. on Autom. Control*, vol. 63, no. 2, pp. 434–448, 2018.

[31] A. Nedić, A. Olshevsky, and W. Shi, "Achieving geometric convergence for distributed optimization over time-varying graphs," *SIAM J. on Optimization*, vol. 27, no. 4, pp. 2597–2633, 2017.

[32] G. Qu and N. Li, "Harnessing smoothness to accelerate distributed optimization," *IEEE Trans. on Control of Network Systems*, vol. PP, no. 99, pp. 1–14, 2017.

[33] C. Xi, R. Xin, and U. A. Khan, "ADD-OPT: Accelerated distributed directed optimization," *IEEE Trans. on Autom. Control*, vol. 63, no. 5, pp. 1329–1339, 2018.

[34] R. Xin and U. A. Khan, "A linear algorithm for optimization over directed graphs with geometric convergence," *IEEE Control Systems Letters*, vol. 2, no. 3, pp. 325–330, 2018.

[35] N. Bof, R. Carli, G. Notarstefano, L. Schenato, and D. Varagnolo, "Multiagent Newton-Raphson optimization over lossy networks," *IEEE Trans. on Autom. Control*, vol. 64, no. 7, pp. 2983–2990, 2019.

[36] A. Nedić, A. Olshevsky, and W. Shi, "A geometrically convergent method for distributed optimization over time-varying graphs," in *IEEE Conf. on Decision and Control (CDC)*, 2016, pp. 1023–1029.

[37] G. Qu and N. Li, "Harnessing smoothness to accelerate distributed optimization," in *IEEE Conf. on Decision and Control (CDC)*, 2016, pp. 159–166.

[38] R. Carli and G. Notarstefano, "Distributed partition-based optimization via dual decomposition," in *IEEE Conf. on Decision and Control (CDC)*, 2013, pp. 2979–2984.

[39] I. Notarnicola and G. Notarstefano, "A randomized primal distributed algorithm for partitioned and big-data non-convex optimization," in *IEEE Conf. on Decision and Control (CDC)*, 2016, pp. 153–158.

[40] I. Notarnicola, R. Carli, and G. Notarstefano, "Distributed partitioned big-data optimization via asynchronous dual decomposition," *IEEE Trans. on Control of Network Sys.*, vol. 5, no. 4, pp. 1910–1919, 2018.

[41] C. Wang, Y. Zhang, B. Ying, and A. H. Sayed, "Coordinate-descent diffusion learning by networked agents," *IEEE Trans. on Signal Process.*, vol. 66, no. 2, pp. 352–367, 2016.

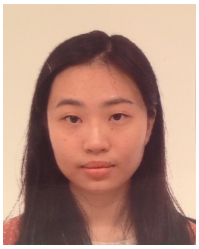
[42] J. Tsitsiklis, D. Bertsekas, and M. Athans, "Distributed asynchronous deterministic and stochastic gradient optimization algorithms," *IEEE Trans. on Autom. Control*, vol. 31, no. 9, pp. 803–812, 1986.

[43] M. Zhu and S. Martínez, "Discrete-time dynamic average consensus," *Automatica*, vol. 46, no. 2, pp. 322–329, 2010.

[44] S. S. Kia, B. Van Scoy, J. Cortés, R. A. Freeman, K. M. Lynch, and S. Martínez, "Tutorial on dynamic average consensus: The problem, its applications, and the algorithms," *IEEE Control Systems Magazine*, vol. 39, no. 3, pp. 40–72, 2019.

[45] A. H. Sayed, "Adaptation, learning, and optimization over networks," *Foundations and Trends in Machine Learning*, vol. 7, no. 4-5, pp. 311–801, 2014.

Ivano Notarnicola (S'16-M'19) is a Postdoctoral Researcher at University of Bologna, Bologna, Italy. He received the Laurea degree "summa cum laude" in Computer Engineering from the Università del Salento, Lecce, in 2014 and the Ph.D. degree "summa cum laude" in Engineering of Complex Systems from the same university in 2018. He was a visiting student at the University of Stuttgart in 2014 and a visiting scholar at Purdue University in 2017. His research interests include distributed methods for optimization, control and machine learning.



Ying Sun received the B.E. degree in electronic information from the Huazhong University of Science and Technology, Wuhan, China, in 2011, and the Ph.D. degree from the Department of Electronic and Computer Engineering at the Hong Kong University of Science and Technology, Hong Kong, in 2016. She is currently a postdoctoral researcher with the School of Industrial Engineering, Purdue University. Her research interests include optimization algorithms, statistical signal processing, and machine learning.

Gesualdo Scutari (S'05-M'06-SM'11) received the Electrical Engineering and Ph.D. degrees (both with honors) from the University of Rome "La Sapienza," Rome, Italy, in 2001 and 2005, respectively. He is the Thomas and Jane Schmidt Rising Star Associate Professor with the School of Industrial Engineering, Purdue University, West Lafayette, IN, USA; and he is the scientific director for the area of Big-Data Analytics at the Cyber Center and the thrust leader for Optimization at the CRISP Center at Purdue University. He had previously held several research appointments, namely, at the University of California at Berkeley, Berkeley, CA, USA; Hong Kong University of Science and Technology, Hong Kong; and University of Illinois at Urbana-Champaign, Urbana, IL, USA. His research interests include continuous and distributed optimization, equilibrium programming, and their applications to signal processing and machine learning. He is a Senior Area Editor of the IEEE Transactions On Signal Processing and an Associate Editor of the SIAM J. on Optimization; He served on the IEEE Signal Processing Society Technical Committee on Signal Processing for Communications (SPCOM). Among others, he was the recipient of the 2006 Best Student Paper Award at the IEEE ICASSP 2006, the 2013 NSF CAREER Award, and the 2015 IEEE Signal Processing Society Young Author Best Paper Award.

Giuseppe Notarstefano (M'11) is a Professor in the Department of Electrical, Electronic, and Information Engineering G. Marconi at Alma Mater Studiorum Università di Bologna. He was Associate Professor (June '16 - June '18) and previously Assistant Professor, Ricercatore, (Feb. '07 - May '16) at the Università del Salento, Lecce, Italy. He received the Laurea degree "summa cum laude" in Electronics Engineering from the Università di Pisa in 2003 and the Ph.D. degree in Automation and Operation Research from the Università di Padova in 2007. He has been visiting scholar at the University of Stuttgart, University of California Santa Barbara and University of Colorado Boulder. His research interests include distributed optimization, cooperative control in complex networks, applied nonlinear optimal control, and trajectory optimization and maneuvering of aerial and car vehicles. He serves as an Associate Editor for IEEE Transactions on Automatic Control, IEEE Transactions on Control Systems Technology and IEEE Control Systems Letters. He has been part of the Conference Editorial Board of IEEE Control Systems Society and EUCA. He is recipient of an ERC Starting Grant 2014.