Mathematical Programming (2019) 176:497–544
https://doi.org/10.1007/s10107-018-01357-w

FULL LENGTH PAPER

Series B

Distributed nonconvex constrained optimization over time-varying digraphs

Gesualdo Scutari[1] · Ying Sun[1]

Received: 24 July 2017 / Accepted: 16 December 2018 / Published online: 16 February 2019

Abstract

This paper considers nonconvex distributed constrained optimization over networks, modeled as directed (possibly time-varying) graphs. We introduce the first algorithmic framework for the minimization of the sum of a smooth nonconvex (nonseparable) function—the agent's sum-utility—plus a difference-of-convex function (with non-smooth convex part). This general formulation arises in many applications, from statistical machine learning to engineering. The proposed distributed method combines successive convex approximation techniques with a judiciously designed perturbed push-sum consensus mechanism that aims to track locally the gradient of the (smooth part of the) sum-utility. Sublinear convergence rate is proved when a fixed step-size (possibly different among the agents) is employed whereas asymptotic convergence to stationary solutions is proved using a diminishing step-size. Numerical results show that our algorithms compare favorably with current schemes on both convex and nonconvex problems.

Mathematics Subject Classification 90C33 · 90C90 · 91A10 · 49M27 · 65K15 · 65K10

Part of this work has been presented at the 2016 Asilomar Conference on System, Signal, and Computers [42] and the 2017 IEEE ICASSP Conference [41].

This work was supported by the USA National Science Foundation, Grants CIF 1564044 and CIF 1719205; the Office of Naval Research, Grant N00014-16-1-2244; and the Army Research Office, Grant W911NF1810238.

✉ Gesualdo Scutari
gscutari@purdue.edu

Ying Sun
sun578@purdue.edu

1 School of Industrial Engineering, Purdue University, West Lafayette, IN, USA

1 Introduction

This paper focuses on the following (possibly) *nonconvex* multiagent composite optimization problem:

$$\min_{\mathbf{x}\in\mathcal{K}} U(\mathbf{x}) \triangleq \underbrace{\sum_{i=1}^{I} f_i(\mathbf{x})}_{F(\mathbf{x})} + \underbrace{G^+(\mathbf{x}) - G^-(\mathbf{x})}_{G(\mathbf{x})}, \tag{P}$$

where $f_i : \mathbb{R}^m \to \mathbb{R}$ is the cost function of agent i, assumed to be smooth (possibly) *nonconvex*; $G : \mathbb{R}^m \to \mathbb{R}$ is a DC function, whose concave part $-G^-$ is smooth; and $\mathcal{K}$ is a closed convex subset of $\mathbb{R}^m$. The function G is generally used to promote some extra structure on the solution, like sparsity. Note that, differently from most of the papers in the literature, we do not require the (sub)gradient of f_i, G^- or G^+ to be (uniformly) bounded on $\mathcal{K}$. Agents are connected through a communication network, modeled as a directed graph, possibly time-varying. Moreover, each agent i knows only its own function f_i (as well as G and $\mathcal{K}$). In this setting, the agents want to cooperatively solve Problem (P) leveraging local communications with their immediate neighbors.

Distributed *nonconvex* optimization in the form (P) has found a wide range of applications in several areas, including network information processing, telecommunications, multi-agent control, and machine learning. In particular, Problem (P) is a key enabler of many emerging *nonconvex* "big data" analytic tasks, including nonlinear least squares, dictionary learning, principal/canonical component analysis, low-rank approximation, and matrix completion [18], just to name a few. Moreover, the DC structure of G allows to accommodate in an unified fashion convex and nonconvex sparsity-inducing surrogates of the ℓ_0 cardinality function (cf. Sect. 2). Time-varying communications arise, for instance, in mobile wireless networks (e.g., ad-hoc networks), wherein nodes are mobile and/or communicate throughout fading channels. Moreover, since nodes generally transmit at different power and/or communication channels are not symmetric, directed links is the natural assumption.

In most of the above scenarios, data processing and optimization need to be performed in a distributed but collaborative manner by the agents within the network. For instance, this is the case in data-intensive (e.g., sensor-network) applications wherein the sheer volume and spatial/temporal disparity of scattered data render centralized processing and storage infeasible or inefficient.

While distributed methods for convex optimization have been widely studied in the literature, there are no such schemes for (P) (cf. Sect. 1.1). We propose the first family of distributed algorithms that converge to stationary solutions of (P) over time-varying (directed) graphs. Asymptotic convergence is proved, under the use of either constant uncoordinate step-sizes from the agents or diminishing ones. When a constant step-size is employed, the algorithms are showed to achieve sublinear convergence rate. Furthermore, the technical tools we introduce are of independent interest. Our analysis hinges on a descent technique valid for nonconvex, nonsmooth, constrained problems based on a novel Lyapunov-like function (see Sect. 1.2 for the list of contributions).

1.1 Related works

The design of distributed algorithms for (P) faces the following challenges: (i) U is *nonconvex* and *nonseparable*; (ii) G^+ is *nonsmooth*; (iii) there are *constraints*; (iv) the graph is *directed* and *time-varying*, with *no specific* structure; and (v) the (sub)gradient of U is not assumed to be bounded on $\mathcal{K}$. We are not aware of any distributed design addressing (even a subset of) challenges (i)–(v), as documented next. Since the focus of this work is on distributed algorithms working on general network architectures, we omit to discuss the vast literature of schemes implementable on *specific* topologies, such as hierarchical networks (e.g., master-slave or shared memory systems); see, e.g., [6,15,16,32,36,37,48] and references therein for an entry point of this literature.

Distributed *convex* optimization Although the focus of this paper is mainly on non-convex optimization, we begin overviewing the much abundant literature of distributed algorithms for *convex* problems. We show in fact that, even in this simpler setting, some of the challenges (ii)–(v) remain unaddressed.

Primal methods While substantially different, primal methods can be generically abstracted as a combination of a local (sub)gradient-like step and a subsequent consensus-like update (or multiple consensus updates); examples include [23,27,31, 38,39]. Algorithms for adaptation and learning tasks based on in-network diffusion techniques were proposed in [8,11,35]. Schemes in [8,11,23,31,38,39] are applicable only to *undirected* graphs; [8,23,31,38,39] require the consensus matrices to be *double-stochastic* whereas [11] uses two matrices that are *row/column-stochastic*, respectively; furthermore, [11] is applicable only to strictly convex agents' cost functions having a *common* minimizer. When the graph is *directed*, double-stochastic weight matrices compliant to the graph might not exist or are not easy to be constructed in a distributed way [20]. This requirement was removed in [27] where the authors combined the sub-gradient algorithm [31] with push-sum consensus [24]. Other schemes applicable to digraphs are [49,50]. However, [31,49,50] cannot handle constraints. In fact, up until this work (and the associated conference papers [41,42]) it was not clear how to leverage push-sum-like protocols to deal with constraints over *digraphs*. Finally, as far as challenge (v) is concerned, only recent proposals [30,33,38,39,49,50,52] removed the assumption that the (sub-)gradient of U has to be bounded; however [30,33,38,49,50,52] can handle only *smooth and unconstrained* problems while [33,38,39,50,52] are not implementable over digraphs.

Dual-based methods This class of algorithms is based on a different approach: slack variables are first introduced to decouple the sum-utility function while forcing consistency among these local copies by adding consensus equality constraints (compliant with the graph topology). Lagrangian dual variables are then introduced to deal with such coupling constraints. The resulting algorithms build on primal–dual updates, aiming at converging to a saddle point of the (augmented) Lagrangian function. Examples of such algorithms include ADMM-like methods [9,22,46] as well as inexact primal–dual instances [10,25,26]. All these algorithms can handle only *static and undirected* graphs. Their extensions to time-varying graphs or digraphs seem not possible, because it is not clear how to enforce consensus via equality constraints over time-varying or directed networks. Furthermore, all the above schemes but [9] require U to be *smooth* and (P) to be *unconstrained*.

In summary, even restricting to convex instances of (P), there exists no distributed algorithm in the literature that can deal with either constraints [issue (iii)] or nonsmooth U [issue (ii)] with unbounded (sub-)gradient [issue (v)] over (time-varying) digraphs. Also, it is not clear how to extend the convergence analysis developed in the above papers when U is no longer convex.

Distributed *nonconvex* optimization Distributed algorithms dealing with special instances of Problem (P) are scarce; they include primal methods [4,12,43,45] and dual-based schemes [21,54]. The key features of these algorithms are summarized in Table 1 and discussed next.

Primal methods The scheme in [4] combines the distributed stochastic projection algorithm, employing a diminishing step-size, with the random gossip protocol. It can handle *smooth* objective functions over *undirected static* graphs; no rate analysis of the scheme is known. In [43], the authors showed that the (randomly perturbed) push-sum gradient algorithm with diminishing (square summable) step-size, earlier proposed for convex objectives in [27], converges also when applied to nonconvex *smooth unconstrained* problems. Asymptotic convergence and a sublinear convergence rate were proved (the latter under the assumption that the set of stationary points of U is finite). The first, to our knowledge, provably convergent distributed scheme for (P), with $G^+ \neq 0$ and constraints $\mathcal{K}$, over time-varying graphs is NEXT [12]. The algorithm requires the consensus matrices to be doubly-stochastic. Asymptotic convergence was proved, when a diminishing step-size is employed; no rate analysis was provided. In [45], the authors studied *smooth* (possibly nonconvex) U over *undirected static* graphs and proposed a distributed instance of the Frank–Wolfe algorithm, coupled with the idea of gradient tracking, first introduced in NEXT (see discussion below for more details on the idea of gradient tracking). Under a diminishing step-size (and further technical assumptions on the set of stationary solutions), a sublinear convergence rate is proved. Finally, all the algorithms discussed above require that the (sub)gradient of U is *bounded* on $\mathcal{K}$ (or $\mathbb{R}^m$). This is a key assumption to prove convergence: in the analysis of descent, it permits to treat the optimization and consensus steps *separately*, with the consensus error being a *summable* perturbation.

Dual-based methods In [54] a distributed approximate dual subgradient algorithm, coupled with a consensus scheme (using double-stochastic weight matrices), is introduced to solve (P) over time-varying graphs. Assuming zero-duality gap, the algorithm is proved to asymptotically find a pair of primal–dual solutions of an auxiliary problem, which however might not be stationary for the original problem; also, consensus is not guaranteed. No rate analysis is provided. In [21], a proximal primal–dual algorithm is proposed to solve an *unconstrained*, *smooth* instance of (P) over *undirected static* graphs. The algorithm employs either a constant or increasing penalty parameter (which plays the role of the step-size); a global sublinear convergence rate is proved. The algorithm can also deal with nonsmooth convex regularizes and norm constraints when it is applied to some distributed matrix factorization problems.

Gradient-tracking The proposed algorithmic framework leverages the idea of gradient tracking: each agent updates its own local variables along a direction that is a proxy of the sum-gradient $\sum_{i=1}^{I} \nabla f_i$ at the current iteration, an information that is not locally available. The idea of tracking the gradient averages through the use of consensus coupled with distributed optimization was independently introduced in [12–14] (NEXT framework) for constrained, nonsmooth, nonconvex instances of (P) over time-varying graphs and in [52] for the case of strongly convex, unconstrained, smooth optimization over static undirected graphs. This tracking protocol was extended to arbitrary (time-varying) digraphs (without requiring doubly-stochastic weight matrices) in our conference work [42]. A convergence rate analysis of the scheme in [52] was later developed in [30,33], with [30] considering (time-varying) directed graphs. We refer the reader to Sect. 3 for a more detailed discussion on this topic.

1.2 Summary of contributions

We summarize our major contributions as follows; see also Table 1.

1. **Novel algorithmic framework** We propose the first provably convergent distributed algorithmic framework for the general class of Problem (P), addressing *all* challenges (i)–(iv). The proposed approach hinges on Successive Convex Approximation (SCA) techniques, coupled with a judiciously designed perturbed push-sum consensus mechanism that aims to track locally the gradient of F. Both communication and tracking protocols are implementable on arbitrary *time-varying* undirected or *directed* graphs, and in the latter case only column-stochasticity of the weight matrices is required. Also, *feasibility* of the iterates is preserved at each iteration. Either constant or diminishing step-size rules can be used in the same scheme, and convergence to stationary solutions of Problem (P) is established.
2. **Iteration complexity** We prove that the proposed scheme has sublinear convergence rate as long as the positive step-size is smaller than an explicit upper bound; different step-sizes among the agents can also be used. To the best of our knowledge, this is the first convergence/complexity result of distributed algorithms employing a constant step-size for nonconvex (constrained) optimization over (time-varying) digraphs.
3. **New Lyapunov-like function and descent technique** We improve upon existing convergence techniques and introduce new ones. Current analysis of distributed algorithms has trouble handling *nonconvex, nonsmooth, constrained* optimization.

Table 1 Distributed nonconvex optimization: current works and contribution of this paper

		Proj-DGM [4]	**NEXT** [12]	**Push-sum DGM** [43]	**Prox-PDA** [21]	**DeFW** [45]	**SONATA** This work
nonsmooth G^+			✓				✓
constraints		✓	✓			$\mathcal{K}$ compact	✓
unbounded gradient					✓		✓
network topology:	**time-varying**		✓	✓			✓
	digraph		restricted	✓			✓
step-size:	**constant**				✓		✓
	diminishing	✓	✓	✓	✓	✓	✓
complexity				✓	✓	✓	✓

Moreover, in the presence of unbounded (sub-)gradients of the objective function, descent on the objective function while treating optimization and consensus errors separately no longer works. A new convergence analysis is introduced to overcome this difficulty based on a novel "Lyapunov"-like function that properly combines suitably defined weighted average dynamics, consensus and tracking disagreements.

4. **Broader class of problems and convergence results** The proposed algorithmic framework and convergence results are applicable to a significantly larger class of (constrained) optimization problems and network topology than current distributed schemes, including several instances arising from machine learning, signal processing, and data analytic applications (cf. Sect. 2.1). Moreover, we contribute to the theory of distributed algorithms also for convex problems, being our schemes the first able to provably deal with either constraints [issue (iii)] or nonsmooth U [issue (ii)] with nonbounded (sub-)gradient [issue (v)] over (time-varying) digraphs. Finally, our algorithm contains as special cases several recently gradient-based algorithms whose convergence was proved under more restrictive assumptions on the optimization problem and network topology (cf. Sect. 5).

Finally, preliminary numerical results show that the proposed schemes compare favorably with state-of-the-art algorithms.

The rest of the paper is organized as follows. The problem setting is discussed in Sect. 2 along with some motivating applications. Some preliminary results, including a perturbed push-sum consensus scheme over time-varying digraphs, are introduced in Sect. 3. Section 4 describes the proposed algorithmic framework along with its convergence properties, whose proofs are given in Sect. 6. Finally, some numerical results are presented in Sect. 7.

Notation The set of nonegative (resp. positive) natural number is denoted by $\mathbb{N}_+$ (resp. $\mathbb{N}_{++}$). A vector $\mathbf{x}$ is viewed as a column vector; matrices are denoted by bold letters. We work with the space $\mathbb{R}^m$, equipped with the standard Euclidean norm, which is denoted by $\| \bullet \|$; when the argument of $\| \bullet \|$ is a matrix, the default norm is the spectral norm. When some other (vector or matrix) norms are used, such as ℓ_1-norm, or infinity-norm, we will use the notation $\| \bullet \|_p$ with the corresponding value of p. The transpose of a vector $\mathbf{x}$ is denoted by $\mathbf{x}^\top$. The Kronecker product is denoted by $\otimes$. We use $\mathbf{1}$ to denote a vector with all entries equal to 1, and $\mathbf{I}$ to denote the identity matrix; With some abuse of notation, the dimensions of $\mathbf{1}$ and $\mathbf{I}$ will not be given explicitly but understood within the context. Given $I \in \mathbb{N}_{++}$, we define $[I] \triangleq \{1, \ldots, I\}$.

2 Problem setup and motivating examples

We study Problem (P) under the following assumptions.

Assumption A (*On Problem* (P)) Given Problem (P), suppose that

A.1 The set $\mathcal{K} \subseteq \mathbb{R}^m$ is (nonempty) closed and convex;
A.2 Each $f_i : \mathcal{O} \to \mathbb{R}$ is C^1, where $\mathcal{O} \supseteq \mathcal{K}$ is an open set, and ∇f_i is L_i-Lipschitz on $\mathcal{K}$;

A.3 $G^+ : \mathcal{K} \to \mathbb{R}$ is convex (possibly nonsmooth), and $G^- : \mathcal{O} \to \mathbb{R}$ is C^1 with ∇G^- being L_G-Lipschitz on $\mathcal{K}$;
A.4 U is lower bounded on $\mathcal{K}$.

We also made the blanket assumption that each agent i knows only its own function f_i and the regularizer G but not the functions of the other agents.

Assumptions A.1 A.2 and A.4 are quite standard and satisfied by several problems of practical interest. We remark that, as a major departure from most of the literature on distributed algorithms, we do not assume that the gradient of F (and G^-) is bounded on the feasible set $\mathcal{K}$. This, together with the nonconvexity of G as stated in A.3, opens the way to design for the first time distributed algorithms for a gamut of new applications, including several big-data problems in statistical learning; see Sect. 2.1 for details.

On the network topology Agents communicate through a (possibly) time-varying network. Specifically, time is slotted with n denoting the iteration index (time-slot); in each time-slot n, the communication network of agents is modeled as a (possibly) time-varying digraph $\mathcal{G}^n = ([I], \mathcal{E}^n)$, where $[I] = \{1, \ldots, I\}$ denotes the set of agents—the vertices of the graph—and the set of edges $\mathcal{E}^n$ represents the agents' communication links; we use $(i, j) \in \mathcal{E}^n$ to indicate that the link is directed from node i to node j. The in-neighborhood of agent i at time n is defined as $\mathcal{N}_i^{\text{in}}[n] = \{j \mid (j, i) \in \mathcal{E}^n\} \cup \{i\}$ (we included in the set node i itself, for notational simplicity); it represents the set of agents which node i can receive information from. The out-neighborhood of agent i is $\mathcal{N}_i^{\text{out}}[n] = \{j \mid (i, j) \in \mathcal{E}^n\} \cup \{i\}$—the set of agents receiving information from node i (including node i itself). The out-degree of agent i is defined as $d_i^n \triangleq \left|\mathcal{N}_i^{\text{out}}[n]\right|$. To let information propagate over the network, we assume that the graph sequence $\{\mathcal{G}^n\}_{n\in\mathbb{N}_+}$ possesses some "long-term" connectivity property, as formally stated next.

Assumption B (*On graph connectivity*) The graph sequence $\{\mathcal{G}^n\}_{n\in\mathbb{N}_+}$ is B-strongly connected, i.e., there exists a finite integer $B > 0$ such that the graph with edge set $\cup_{t=k}^{k+B-1}\mathcal{E}^t$ is strongly connected, for all $k \geq 0$.

We conclude this section discussing some instances of Problem (P) in the context of statistical learning.

2.1 Distributed sparse statistical learning

We consider two distributed nonconvex problems in statistical learning, namely: (i) a nonconvex sparse linear regression problem; and (ii) the sparse Principal Component Analysis (PCA) problem.

Nonconvex Sparse Linear Regression Consider the problem of retrieving a sparse signal $\mathbf{x} \in \mathbb{R}^m$ from the observations $\{\mathbf{b}_i\}_{i=1}^I$, where each $\mathbf{b}_i = \mathbf{A}_i\mathbf{x}$ is a linear measurement of the signal acquired by agent i. A mainstream approach in the literature is to solve the following optimization problem

$$\min_{\mathbf{x}} \quad \sum_{i=1}^{I} \|\mathbf{b}_i - \mathbf{A}_i\mathbf{x}\|^2 + \lambda \cdot G(\mathbf{x}), \tag{1}$$

Table 2 Examples of nonconvex surrogates of the ℓ_0 function having a DC structure [cf. (2)]

Penalty function	Expression
Exp [7]	$g_{\exp}(x) = 1 - e^{-\theta\|x\|}$
$\ell_p(0 < p < 1)$ [19]	$g_{\ell_p^+}(x) = (\|x\| + \epsilon)^{1/\theta}$,
$\ell_p(p < 0)$ [34]	$g_{\ell_p^-}(x) = 1 - (\theta\|x\| + 1)^p$
SCAD [17]	$g_{\text{scad}}(x) = \begin{cases} \frac{2\theta}{a+1}\|x\|, & 0 \leq \|x\| \leq \frac{1}{\theta} \\ \frac{-\theta^2\|x\|^2+2a\theta\|x\|-1}{a^2-1}, & \frac{1}{\theta} < \|x\| \leq \frac{a}{\theta} \\ 1, & \|x\| > \frac{a}{\theta} \end{cases}$
Log [47]	$g_{\log}(x) = \frac{\log(1+\theta\|x\|)}{\log(1+\theta)}$

where the quadratic term measures the model fitness whereas the regularizer G is used to promote sparsity in the solution, and $\lambda > 0$ is chosen to balance the trade-off between the model fitness and solution sparsity. Problem (1) is clearly an instance of (P). Note that each agent knows only its own function f_i (since $\mathbf{b}_i$ is own only by agent i). Also, ∇f_i is not bounded on $\mathbb{R}^m$.

To promote sparsity on the solution, the ideal choice for G would be the cardinality of $\mathbf{x}$ (a.k.a. ℓ_0 "norm" of $\mathbf{x}$). However, its combinatorial nature makes the resulting optimization problem numerically intractable as the variable dimension m becomes large. Several convex and, more recently, also nonconvex surrogates of the ℓ_0 function have been proposed in the literature. The structure of G, as stated in Assumption A.3, captures either choices. For instance, one can choose as regularizer in (1), the ℓ_2 or ℓ_1 norm of $\mathbf{x}$ (and thus $G^- = 0$), which leads to the ridge and LASSO regression problems, respectively. Moreover, a vast class of nonconvex surrogates can also be considered, including the SCAD [17], the "transformed" ℓ_1 [53], the logarithmic [47], and the exponential [7]; see Table 2. It is well documented that nonconvex regularizers outperform the ℓ_1 norm in enhancing solution sparsity. Quite interestingly, all the widely used nonconvex surrogates listed in Table 2 enjoy the following separable DC structure (see, e.g., [1,44] and references therein)

$$G(\mathbf{x}) = \sum_{i=1}^{m} g(x_i), \quad \text{with} \quad g(x_i) = \underbrace{\eta(\theta)\,|x_i|}_{\triangleq g^+(x_i)} - \underbrace{\big(\eta(\theta)\,|x_i| - g(x_i)\big)}_{\triangleq g^-(x_i)}, \tag{2}$$

where the expression of $g : \mathbb{R} \to \mathbb{R}$ is given in Table 2; and $\eta(\theta)$ is a fixed given function, defined in Table 3 for each of the surrogate g listed in Table 2. The parameter θ controls the tightness of the approximation of the ℓ_0 function: in fact, it holds that $\lim_{\theta\to+\infty} g(x_i) = 1$ if $x_i \neq 0$, otherwise $\lim_{\theta\to+\infty} g(x_i) = 0$. Note that g^- is convex and has Lipschitz continuous first derivative dg^-/dx [44], whose closed form is given in Table 3.

It is not difficult to check that Problem (1), with any of the regularizers discussed above, is an instance of (P) and satisfies Assumption A. Also, note that the gradient of the smooth part is not bounded on $\mathbb{R}^m$.

Table 3 Explicit expression of $\eta(\theta)$ and dg^-/dx [cf. (2)]

g	$\eta(\theta)$	dg_θ^-/dx
$g_{\exp}$	θ	$\text{sign}(x)\cdot\theta\cdot(1-e^{-\theta\lvert x\rvert})$
$g_{\ell_p^+}$	$\frac{1}{\theta}\epsilon^{1/\theta-1}$	$\frac{1}{\theta}\,\text{sign}(x)\cdot[\epsilon^{\frac{1}{\theta}-1}-(\lvert x\rvert+\epsilon)^{\frac{1}{\theta}-1}]$
$g_{\ell_p^-}$	$-p\cdot\theta$	$-\text{sign}(x)\cdot p\cdot\theta\cdot[1-(1+\theta\lvert x\rvert)^{p-1}]$
g_{scad}	$\frac{2\theta}{a+1}$	$\begin{cases}0, & \lvert x\rvert\le\frac{1}{\theta}\\ \text{sign}(x)\cdot\frac{2\theta(\theta\lvert x\rvert-1)}{a^2-1}, & \frac{1}{\theta}<\lvert x\rvert\le\frac{a}{\theta}\\ \text{sign}(x)\cdot\frac{2\theta}{a+1}, & \text{otherwise}\end{cases}$
$g_{\log}$	$\frac{\theta}{\log(1+\theta)}$	$\text{sign}(x)\cdot\frac{\theta^2\lvert x\rvert}{\log(1+\theta)(1+\theta\lvert x\rvert)}$

Sparse PCA Consider finding the sparse principal component of a distributed data set given by the rows of a set of matrices $\mathbf{D}_i$'s (each $\mathbf{D}_i$ is own by agent i). The problem can be formulated as

$$\max_{\|\mathbf{x}\|_2\le 1}\quad \sum_{i=1}^{I}\|\mathbf{D}_i\mathbf{x}\|^2-\lambda\cdot G(\mathbf{x}), \tag{3}$$

where G can be any of the sparse-promoting regularizers discussed in the previous example. Clearly, Problem (3) is another (nonconvex) instance of Problem (P) (satisfying Assumption A).

3 Preliminaries: the perturbed condensed push-sum algorithm

The proposed algorithmic framework combines local optimization based on SCA with constrained consensus and tracking of gradient averages over digraphs.

The consensus problem over graphs has been widely studied in the literature; a renowned distributed scheme solving this problem over (possibly time-varying) digraphs is the so-called push-sum algorithm [24]. A perturbed version of the push-sum scheme has been introduced in [27] to solve *unconstrained* optimization problems over (time-varying) digraphs. However, it is not clear how to leverage the push-sum update and extend these optimization schemes to deal with constraints. In this section, we introduce a reformulation of the *perturbed* push-sum protocol [27]—termed *perturbed condensed* push-sum—that is more suitable for the integration with *constrained* optimization. This scheme will be then used to build the gradient tracking and constrained consensus mechanisms embedded in the proposed algorithmic framework (cf. Sect. 4).

Consider a network of I agents, as introduced in Sect. 2, communicating over a time-varying digraph (cf. Assumption B). Each agent i controls a vector of variables $\mathbf{x}_{(i)}\in\mathbb{R}^m$ as well as a scalar ϕ_i that are iteratively updated, based upon the information received from its immediate neighbors. Let $\mathbf{x}_{(i)}^n$ and ϕ_i^n denote the values of $\mathbf{x}_{(i)}$ and ϕ_i at iteration $n\in\mathbb{N}_+$. We let agents' updates be subject to a(n adversarial) perturbation; we denote by $\boldsymbol{\delta}_i^n\in\mathbb{R}^m$ the perturbation injected in the update of agent i at iteration n. Given $\mathbf{x}_{(i)}^n$ and ϕ_i^n, the perturbed condensed consensus algorithm reads:

$$\phi_i^{n+1} = \sum_{j=1}^{I} a_{ij}^n \phi_j^n, \tag{4a}$$

$$\mathbf{x}_{(i)}^{n+1} = \frac{1}{\phi_i^{n+1}} \sum_{j=1}^{I} a_{ij}^n \phi_j^n \mathbf{x}_{(j)}^n + \boldsymbol{\delta}_i^{n+1}, \tag{4b}$$

for all $n \in \mathbb{N}_+$ and $i \in [I]$, where $\mathbf{x}_{(i)}^0$ are arbitrarily chosen and ϕ_i^0 are positive scalars such that $\sum_{i=1}^I \phi_i^0 = I$; and $\mathbf{A}^n \triangleq (a_{ij}^n)_{i,j=1}^I$ is a (possibly) time-varying matrix of weights whose nonzero pattern is compliant with the topology of the graph $\mathcal{G}^n$, in the sense of the assumption below.

Assumption C (*On the weight matrix* $\mathbf{A}^n$) Each $\mathbf{A}^n \triangleq (a_{ij}^n)_{i,j=1}^I$ is compliant with $\mathcal{G}^n$, that is,

C1. $a_{ii}^n \geq \kappa > 0$, for all $i \in [I]$;
C2. $a_{ij}^n \geq \kappa > 0$, if $(j, i) \in \mathcal{E}^n$; and $a_{ij}^n = 0$ otherwise.

Under Assumption C, the protocol (4) is implementable in a distributed fashion: each agent i updates its own variables using only the information $\phi_j^n \mathbf{x}_{(j)}^n$ and ϕ_j^n received from its current in-neighbors (and its own). We study convergence of (4) under the following further (standard) assumption on $\mathbf{A}^n$.

Assumption D (*Column stochasticity*) Each matrix $\mathbf{A}^n$ is column stochastic, that is, $\mathbf{1}^\top \mathbf{A}^n = \mathbf{1}^\top$.

The role of the extra variables ϕ_i is to dynamically rebuild the row stochasticity of the equivalent weight matrix governing variables' updates, which is a key condition to lock consensus. This can be easily seen rewriting the dynamics (4b) in terms of the equivalent weights $\mathbf{W}^n \triangleq (w_{ij})_{i,j=1}^I$:

$$\mathbf{x}_{(i)}^{n+1} = \sum_{j=1}^{I} w_{ij}^n \, \mathbf{x}_{(j)}^n, \qquad w_{ij}^n \triangleq \frac{a_{ij}^n \phi_j^n}{\phi_i^{n+1}}. \tag{5}$$

It is not difficult to check that, under Assumption D, $\mathbf{W}^n$ is row-stochastic.

To state the main convergence result in compact form, we introduce the following notation. Let

$$\mathbf{x}^n \triangleq \left[\mathbf{x}_{(1)}^{n\,\top}, \ldots, \mathbf{x}_{(I)}^{\top}\right]^\top, \tag{6a}$$

$$\boldsymbol{\phi}^n \triangleq \left[\phi_1^n, \ldots, \phi_I^n\right]^\top, \tag{6b}$$

$$\boldsymbol{\delta}^n \triangleq \left[\boldsymbol{\delta}_1^{n\,\top}, \ldots, \boldsymbol{\delta}_I^{n\,\top}\right]^\top. \tag{6c}$$

Noting that, in the absence of perturbation (i.e., $\boldsymbol{\delta}^n = \mathbf{0}$), the weighed sum $\sum_{i=1}^I \phi_i^n \mathbf{x}_{(i)}^n$ is an invariant of (4), that is, $\sum_{i=1}^I \phi_i^{n+1} \mathbf{x}_{(i)}^{n+1} = \cdots = \sum_{i=1}^I \phi_i^0 \mathbf{x}_{(i)}^0$. We define the consensus disagreement at iteration n as the deviation of each $\mathbf{x}_{(i)}^n$ from the weighted average $(1/I) \sum_{i=1}^I \phi_i^n \mathbf{x}_{(i)}^n$:

$$\mathbf{e}_x^n \triangleq \mathbf{x}^n - \mathbf{1} \otimes \frac{1}{I} \sum_{i=1}^{I} \phi_i^n \mathbf{x}_{(i)}^n. \tag{7}$$

The dynamics of the error $\mathbf{e}_x^n$ are studied in the following proposition (whose proof is postponed to Sect. 3.2).

Proposition 1 *Let $\{\mathcal{G}^n\}_{n\in\mathbb{N}_+}$ be a sequence of digraphs satisfying Assumption B, and let $\{(\boldsymbol{\phi}^n, \mathbf{x}^n)\}_{n\in\mathbb{N}_+}$ be the sequence generated by the perturbed condensed push-sum protocol (4), for a given perturbation sequence $\{\boldsymbol{\delta}^n\}_{n\in\mathbb{N}_+}$ and weight matrices $\{\mathbf{A}^n\}_{n\in\mathbb{N}_+}$ satisfying Assumptions C–D. Then, there hold:*

(i) *[Bounded $\{\boldsymbol{\phi}^n\}_{n\in\mathbb{N}_+}$]:*

$$\begin{aligned} \inf_{n\in\mathbb{N}_+} \min_{i\in[I]} \phi_i^n &\geq \phi_{lb}, \qquad \phi_{lb} \triangleq \kappa^{2(I-1)B}, \\ \sup_{n\in\mathbb{N}_+} \max_{i\in[I]} \phi_i^n &\leq \phi_{ub}, \qquad \phi_{ub} \triangleq I - \kappa^{2(I-1)B}, \end{aligned} \tag{8}$$

with $B \geq 1$ and $\kappa \in (0, 1)$ defined in Assumptions B and C, respectively;

(ii) *[Error decay]: For all $n, k \in \mathbb{N}_+$, $n \geq k$,*

$$\|\mathbf{e}_x^n\| \leq \lambda^k \,\|\mathbf{e}_x^{n-k}\| + \lambda^t \cdot \sum_{t=0}^{k-1} \|\boldsymbol{\delta}^{n-t}\|, \tag{9}$$

where

$$\lambda^t \triangleq \min\left\{\sqrt{2}\, I,\, 2\, c_0\, I\, (\rho)^{\left\lfloor \frac{t}{(I-1)B} \right\rfloor}\right\},$$

and

$$c_0 \triangleq 2\left(1 + \tilde{\kappa}^{-(I-1)B}\right), \quad \rho \triangleq 1 - \tilde{\kappa}^{(I-1)B}, \quad \tilde{\kappa} \triangleq \kappa^{2(I-1)B+1}/I. \tag{10}$$

Furtheremore, there exists a finite $\bar{B} \in \mathbb{N}_+$ such that $\rho_{\bar{B}} \triangleq 2c_0 I (\rho)^{\left\lfloor \frac{\bar{B}}{(I-1)B} \right\rfloor} < 1$.

Remark 1 The perturbed consensus algorithm (4) was mainly designed for digraphs. However, when the graph is undirected, one can choose the weight matrix $\mathbf{A}^n$ to be double stochastic and get rid of the auxiliary variables $\boldsymbol{\phi}^n$, just setting in (4) $\boldsymbol{\phi}^0 = \mathbf{1}$. As a consequence, $\boldsymbol{\phi}^n \equiv \mathbf{1}$ and $\mathbf{W}^n \equiv \mathbf{A}^n$, for all $n \in \mathbb{N}_+$. In this case, using [29, Lemma 9], the expression of λ^t in Proposition 1 can be tightened by letting $\lambda^t \triangleq \min\{1, (\rho)^{\lfloor t/B \rfloor}\}$, with $\rho \triangleq \sqrt{1 - \kappa/\left(2I^2\right)}$.

3.1 Discussion

Proposition 1 provides a unified set of convergence conditions of the perturbed condensed push-sum scheme that are applicable to any given perturbation sequence

$\{\boldsymbol{\delta}^n\}_{n\in\mathbb{N}_+}$. We discuss next two special cases, namely: the plain average consensus problem and the distributed tracking of time-varying signals.

1. (Weighted) average consensus Setting in (4) $\boldsymbol{\delta}^n = \mathbf{0}$, for all $n \in \mathbb{N}_+$, (4) reduces to the plain (condensed) push-sum scheme. whose geometric convergence to the (weighted) average of the initial values, $(1/I)\sum_{i=1}^{I}\phi_i^0\,\mathbf{x}_{(i)}^0$, follows readily from Proposition 1. More specifically, using $\sum_{i=1}^{I}\phi_i^{n+1}\mathbf{x}_{(i)}^{n+1} = \cdots = \sum_{i=1}^{I}\phi_i^0\mathbf{x}_{(i)}^0$, (9) yields

$$\left\|\mathbf{x}^{n+1} - \mathbf{1}\otimes\frac{1}{I}\sum_{i=1}^{I}\phi_i^0\mathbf{x}_{(i)}^0\right\| \le 2\,c_0\,I\,(\rho)^{\left\lfloor\frac{n+1}{(I-1)B}\right\rfloor}\left\|\mathbf{e}_x^0\right\|,\quad n\in\mathbb{N}_+. \tag{11}$$

Note that, since the weight matrix $\mathbf{W}^n$ in (5) is row stochastic, if the initial values $\mathbf{x}_{(i)}^0$ all belong to a common set $\mathcal{K}$, then $\mathbf{x}_{(i)}^n \in \mathcal{K}$, for all $n \in \mathbb{N}_{++}$; that is feasibility of the iterates is preserved.

2. Tracking of time-varying signals' averages Consider the problem of tracking distributively the average of time-varying signals. At each iteration $n \in \mathbb{N}_+$, each agent i evaluates (or generates) a signal sample $\mathbf{u}_i^n \in \mathbb{R}^m$ from the (time-varying) sequence $\{\mathbf{u}_i^n\}_{n\in\mathbb{N}_+}$. The goal is to design a distributed algorithm obeying the communication structure of the graphs $\mathcal{G}^n$ that tracks the average of the signals $\{\mathbf{u}_i^n\}_{n\in\mathbb{N}_+}$, that is,

$$\lim_{n\to\infty}\left\|\mathbf{x}^n - \mathbf{1}\otimes\bar{\mathbf{u}}^n\right\| = 0,\quad \bar{\mathbf{u}}^n \triangleq \frac{1}{I}\sum_{i=1}^{I}\mathbf{u}_i^n. \tag{12}$$

The perturbed condensed push-sum algorithm (4) can be readily used to accomplish this task by setting

$$\boldsymbol{\delta}_i^{n+1} = \frac{1}{\phi_i^{n+1}}\left(\mathbf{u}_i^{n+1} - \mathbf{u}_i^n\right),\qquad i\in[I],\ n\in\mathbb{N}_+, \tag{13}$$

and $\mathbf{x}_i^0 = \mathbf{u}_i^0$, $i \in [I]$. Convergence of this scheme is stated next.

Corollary 2 *Let $\{\mathbf{u}_i^n\}_{n\in\mathbb{N}_+}$ be a given sequence such that $\lim_{n\to\infty}\|\mathbf{u}_i^{n+1} - \mathbf{u}_i^n\| = 0$, for all $i \in [I]$. Consider the perturbed condensed push-sum protocol (4), under the assumptions of Proposition 1; and set $\boldsymbol{\delta}_i^{n+1}$ as in (13), $\mathbf{x}_i^0 = \mathbf{u}_i^0$, and $\phi_i^0 = 1$ for all $i \in [I]$. Then, (12) holds.*

Proof The proof follows readily from Proposition 1 and the following two facts:

(i) $(1/I)\sum_{i=1}^{I}\phi_i^{n+1}\mathbf{x}_{(i)}^{n+1} = \bar{\mathbf{u}}^{n+1}$; and (ii) [28, Lemma 7]

$$\lim_{n\to\infty}\|\boldsymbol{\delta}^n\| = 0 \ \Rightarrow\ \lim_{n\to\infty}\sum_{t=0}^{n-1}(\rho)^{\left\lfloor\frac{t}{(I-1)B}\right\rfloor}\|\boldsymbol{\delta}^{n-t}\| = 0.$$

□

3.2 Proof of Proposition 1

To prove Proposition 1, it is convenient to rewrite the perturbed consensus protocol (4) in a vector–matrix form. To do so, let us introduce the following quantities: given the weight matrix $\mathbf{A}^n$ compliant with $\mathcal{G}^n$ (cf. Assumption C) and $\mathbf{W}^n$ defined in (5), let

$$\mathbf{D}_{\boldsymbol{\phi}^n} \triangleq \mathrm{Diag}\left(\boldsymbol{\phi}^n\right), \tag{14a}$$

$$\widehat{\mathbf{D}}_{\boldsymbol{\phi}^n} \triangleq \mathbf{D}_{\boldsymbol{\phi}^n} \otimes \mathbf{I}, \tag{14b}$$

$$\widehat{\mathbf{A}}^n \triangleq \mathbf{A}^n \otimes \mathbf{I}, \tag{14c}$$

$$\widehat{\mathbf{W}}^n \triangleq \mathbf{W}^n \otimes \mathbf{I}, \tag{14d}$$

where Diag(•) denotes a diagonal matrix whose diagonal entries are the elements of the vector argument, and $\mathbf{I}$ is the $m \times m$ identity matrix. Under the column stochasticity of $\mathbf{A}^n$, it is not difficult to check that the following holds:

$$\mathbf{W}^n = \left(\mathbf{D}_{\boldsymbol{\phi}^{n+1}}\right)^{-1} \mathbf{A}^n \, \mathbf{D}_{\boldsymbol{\phi}^n} \quad \text{and} \quad \widehat{\mathbf{W}}^n = \left(\widehat{\mathbf{D}}_{\boldsymbol{\phi}^{n+1}}\right)^{-1} \widehat{\mathbf{A}}^n \, \widehat{\mathbf{D}}_{\boldsymbol{\phi}^n}. \tag{15}$$

Using the above notation and (6), the perturbed push-sum protocol (4) can be rewritten in matrix–vector form as

$$\boldsymbol{\phi}^{n+1} = \mathbf{A}^n \boldsymbol{\phi}^n \quad \text{and} \quad \mathbf{x}^{n+1} = \widehat{\mathbf{W}}^n \mathbf{x}^n + \boldsymbol{\delta}^{n+1}. \tag{16}$$

To study convergence of (16), it is convenient to introduce the following matrix products: given $n, k \in \mathbb{N}_+$, with $n \geq k$,

$$\begin{aligned} \mathbf{A}^{n:k} &\triangleq \begin{cases} \mathbf{A}^n \mathbf{A}^{n-1} \cdots \mathbf{A}^k, & \text{if } n > k, \\ \mathbf{A}^n, & \text{if } n = k, \end{cases} \\ \mathbf{W}^{n:t} &\triangleq \begin{cases} \mathbf{W}^n \mathbf{W}^{n-1} \cdots \mathbf{W}^k, & \text{if } n > k, \\ \mathbf{W}^n, & \text{if } n = k, \end{cases} \end{aligned} \tag{17}$$

and

$$\widehat{\mathbf{A}}^{n:k} \triangleq \mathbf{A}^{n:k} \otimes \mathbf{I}, \quad \widehat{\mathbf{W}}^{n:k} \triangleq \mathbf{W}^{n:k} \otimes \mathbf{I}. \tag{18}$$

Define the weight-averaging matrix

$$\mathbf{J}_{\boldsymbol{\phi}^n} \triangleq \frac{1}{I}\left(\mathbf{1}\,(\boldsymbol{\phi}^n)^\top\right) \otimes \mathbf{I}, \tag{19}$$

so that $\mathbf{J}_{\boldsymbol{\phi}^n}\, \mathbf{x}^n = \mathbf{1} \otimes \frac{1}{I}\sum_{i=1}^{I} \phi_i^n \mathbf{x}_{(i)}^n$. Also, it is not difficult to check the following chain of equalities hold among $\mathbf{J}_{\boldsymbol{\phi}}^n$, $\widehat{\mathbf{W}}^{n:t}$, and $\widehat{\mathbf{A}}^{n:t}$: for $n, k \in \mathbb{N}_+$, with $n \geq k$,

$$\mathbf{J}_{\boldsymbol{\phi}^{n+1}}\,\widehat{\mathbf{W}}^{n:k} \stackrel{(a)}{=} \mathbf{J}_{\mathbf{1}}\widehat{\mathbf{D}}_{\boldsymbol{\phi}^k} = \mathbf{J}_{\boldsymbol{\phi}^k} \stackrel{(b)}{=} \widehat{\mathbf{W}}^{n:k}\,\mathbf{J}_{\boldsymbol{\phi}^k}, \tag{20}$$

where in (a) we used the definition of $\widehat{\mathbf{W}}^n$ [cf. (15)], $\mathbf{J}_{\boldsymbol{\phi}^{n+1}}$ [cd. (20)], and the column stochasticity of $\widehat{\mathbf{A}}^n$; and (b) is due to the row stochasticity of $\widehat{\mathbf{W}}^{n:k}$.

The consensus error $\mathbf{e}_x^n$ in (7) can be rewritten as $\mathbf{e}_x^n = (\mathbf{I} - \mathbf{J}_{\boldsymbol{\phi}^n})\mathbf{x}^n$.

To study the evolution of $\mathbf{e}_x^n$, we apply the x-update (16) recursively and obtain

$$\mathbf{x}^n = \widehat{\mathbf{W}}^{n-1:n-k}\,\mathbf{x}^{n-k} + \sum_{t=1}^{k-1} \widehat{\mathbf{W}}^{n-1:n-t}\,\boldsymbol{\delta}^{n-t} + \boldsymbol{\delta}^n. \tag{21}$$

Using (20) and (21), the weighted average $\mathbf{J}_{\boldsymbol{\phi}^n}\mathbf{x}^n$ can be written as

$$\mathbf{J}_{\boldsymbol{\phi}^n}\mathbf{x}^n = \mathbf{J}_{\boldsymbol{\phi}^{n-k}}\mathbf{x}^{n-k} + \sum_{t=1}^{k-1} \mathbf{J}_{\boldsymbol{\phi}^{n-t}}\boldsymbol{\delta}^{n-t} + \mathbf{J}_{\boldsymbol{\phi}^n}\boldsymbol{\delta}^n. \tag{22}$$

Subtracting (22) from (21) and using $\big(\widehat{\mathbf{W}}^{n-1:n-k} - \mathbf{J}_{\boldsymbol{\phi}^{n-k}}\big)\mathbf{J}_{\boldsymbol{\phi}^{n-k}} = \mathbf{0}$ [cf. (20)], we can bound the consensus error $\mathbf{e}_x^{n+1}$ as

$$\begin{aligned}\|\mathbf{e}_x^n\| \leq{}& \left\|\widehat{\mathbf{W}}^{n-1:n-k} - \mathbf{J}_{\boldsymbol{\phi}^{n-k}}\right\| \|\mathbf{e}_x^{n-k}\| + \sum_{t=1}^{k-1} \left\|\widehat{\mathbf{W}}^{n-1:n-t} - \mathbf{J}_{\boldsymbol{\phi}^{n-t}}\right\| \|\boldsymbol{\delta}^{n-t}\| \\ &+ \left\|\mathbf{I} - \mathbf{J}_{\boldsymbol{\phi}^n}\right\| \|\boldsymbol{\delta}^n\|.\end{aligned} \tag{23}$$

Convergence of the perturbed consensus protocol reduces to studying the dynamics of the matrix product $\|\widehat{\mathbf{W}}^{n:k} - \mathbf{J}_{\boldsymbol{\phi}^k}\|$, as done in the lemma below.

Lemma 3 *Let $\{\mathcal{G}^n\}_{n\in\mathbb{N}_+}$ be a sequence of digraphs satisfying Assumption B; let $\{\mathbf{A}^n\}_{n\in\mathbb{N}_+}$ be a sequence of weight matrices satisfying Assumptions C–D; and let $\{\mathbf{W}^n\}_{n\in\mathbb{N}_+}$ be the sequence of row stochastic matrices related to $\{\mathbf{A}^n\}_{n\in\mathbb{N}_+}$ by (15). There holds:*

$$\left\|\widehat{\mathbf{W}}^{n:k} - \mathbf{J}_{\boldsymbol{\phi}^k}\right\| \leq \min\left\{\sqrt{2}\,I,\, 2\,c_0 I(\rho)^{\left\lfloor \frac{n-k+1}{(I-1)B}\right\rfloor}\right\}, \quad n,\, k \in \mathbb{N}_+, \quad n \geq k, \tag{24}$$

where c_0 and ρ are defined in Proposition 1.

Proof See Appendix A. □

The error decay law (9) comes readily from (23), Lemma 3, and the following fact: $\left\|\mathbf{I} - \mathbf{J}_{\boldsymbol{\phi}^n}\right\| \leq \sqrt{2\,I} \leq \lambda^0 \triangleq \min\{2c_0 I, \sqrt{2}I\}$, which is proved below. Let $\mathbf{z} \in \mathbb{R}^{I\cdot m}$ be an arbitrary vector; let us partition $\mathbf{z}$ as $\mathbf{z} = [\mathbf{z}_1^\top, \ldots, \mathbf{z}_I^\top]^\top$, with each $\mathbf{z}_i \in \mathbb{R}^m$. Then,

$$\left\|(\mathbf{I}-\mathbf{J}_{\boldsymbol{\phi}^n})\,\mathbf{z}\right\| \leq \|\mathbf{z}-\mathbf{J}_{\mathbf{1}}\,\mathbf{z}\| + \left\|\mathbf{J}_{\mathbf{1}}\mathbf{z}-\mathbf{J}_{\boldsymbol{\phi}^n}\mathbf{z}\right\| \overset{(a)}{\leq} \|\mathbf{z}\| + \frac{\sqrt{I}}{I}\left\|\sum_{i=1}^{I}\mathbf{z}_i - \sum_{i=1}^{I}\phi_i^n\mathbf{z}_i\right\|$$
$$\leq \|\mathbf{z}\| + \frac{\sqrt{I}}{I}\sqrt{I^2-I}\;\|\mathbf{z}\| \leq \sqrt{2\,I}\;\|\mathbf{z}\|\,, \tag{25}$$

where in (a) we used $\|\mathbf{I}-\mathbf{J}_{\mathbf{1}}\| = 1$. □

4 Algorithmic design

We are ready to introduce the proposed distributed algorithm for Problem (P). To shed light on the core idea of the novel framework, we begin introducing an informal and constructive description of the algorithm (cf. Sect. 4.1), followed by its formal statement along with its convergence properties (cf. Sect. 4.2).

4.1 SONATA at-a-glance

Each agent i maintains and updates iteratively a *local copy* $\mathbf{x}_{(i)}$ of the global variable $\mathbf{x}$, along with an auxiliary variable $\mathbf{y}_{(i)} \in \mathbb{R}^m$; let $\mathbf{x}_{(i)}^n$ and $\mathbf{y}_{(i)}^n$ denote the values of $\mathbf{x}_{(i)}$ and $\mathbf{y}_{(i)}$ at iteration n, respectively. Roughly speaking, the update of these variables is designed so that all the $\mathbf{x}_{(i)}^n$ will be asymptotically consensual, converging to a stationary solution of (P); and each $\mathbf{y}_{(i)}$ tracks locally the average of the gradients $(1/I)\cdot\sum_{i=1}^{I}\nabla f_i$, an information that is not available at the agent's side. More specifically, the following two steps are performed iteratively and in parallel across the agents.

Step 1: Local SCA The nonconvexity of f_i together with the lack of knowledge of $\sum_{j\neq i} f_j$ in F, prevent agent i to solve Problem (P) directly. To cope with these issues, we leverage SCA techniques: at each iteration n, given the current iterate $\mathbf{x}_{(i)}^n$ and $\mathbf{y}_{(i)}^n$, agent i solves instead a convexification of (P), having the following form:

$$\widetilde{\mathbf{x}}_{(i)}^n \triangleq \underset{\mathbf{x}_{(i)}\in\mathcal{K}}{\operatorname{argmin}}\; \widetilde{F}_i\left(\mathbf{x}_{(i)};\mathbf{x}_{(i)}^n,\mathbf{y}_{(i)}^n\right) + G^+\left(\mathbf{x}_{(i)}\right), \tag{26}$$

and updates its $\mathbf{x}_{(i)}$ according to

$$\mathbf{x}_{(i)}^{n+1/2} = \mathbf{x}_{(i)}^n + \alpha^n\left(\widetilde{\mathbf{x}}_{(i)}^n - \mathbf{x}_{(i)}^n\right), \tag{27}$$

where $\alpha^n \in (0,1)$ is a step-size (to be properly chosen). In (26), $\widetilde{F}_i(\bullet;\mathbf{x}_{(i)}^n,\mathbf{y}_{(i)}^n)$ is chosen as:

$$\widetilde{F}_i\left(\mathbf{x}_{(i)};\mathbf{x}_{(i)}^n,\mathbf{y}_{(i)}^n\right) \triangleq \widetilde{f}_i\left(\mathbf{x}_{(i)};\mathbf{x}_{(i)}^n\right) - \nabla G^-\left(\mathbf{x}_{(i)}^n\right)^\top\left(\mathbf{x}_{(i)}-\mathbf{x}_{(i)}^n\right)$$
$$+\left(I\cdot\mathbf{y}_{(i)}^n - \nabla f_i\left(\mathbf{x}_{(i)}^n\right)\right)^\top\left(\mathbf{x}_{(i)}-\mathbf{x}_{(i)}^n\right), \tag{28}$$

where $\widetilde{f}_i(\bullet;\mathbf{x}^n_{(i)})$ is a strongly convex approximation of f_i at the current iterate $\mathbf{x}^n_{(i)}$ (see Assumption E below); the second term is the linearization of the smooth nonconvex function $-G^-$; and $\mathbf{y}^n_{(i)}$, as anticipated, aims at tracking the gradient average $(1/I)\sum_{j=1}^I \nabla f_j(\mathbf{x}^n_{(i)})$, that is, $\lim_{n\to\infty}\|\mathbf{y}^n_{(i)} - (1/I)\sum_{j=1}^I \nabla f_j(\mathbf{x}^n_{(i)})\| = 0$. This sheds light on the role of the last term in (28): under the claimed tracking properties of $\mathbf{y}^n_{(i)}$, there would hold:

$$\lim_{n\to\infty}\left\|\left(I\cdot\mathbf{y}^n_{(i)} - \nabla f_i\left(\mathbf{x}^n_{(i)}\right)\right) - \sum_{j\neq i}\nabla f_j(\mathbf{x}^n_{(i)})\right\| = 0. \tag{29}$$

Therefore, the last term in (28) can be seen as a proxy of the gradient sum $\sum_{j\neq i}\nabla f_j(\mathbf{x}^n_{(i)})$, which is not available at agent i's site. Building on the perturbed condensed push-sum protocol introduced in Sect. 3 we will show in Step 2 below how to update $\mathbf{y}^n_{(i)}$ so that (29) holds, using only *local* information.

The surrogate function $\widetilde{f}_i$ satisfies the following assumption.

Assumption E (*On surrogate function* $\widetilde{f}_i$) Let $\widetilde{f}_i : \mathcal{K}\times\mathcal{K}\to\mathbb{R}$ be a C^1 function with respect to its first argument, and such that

E1. $\nabla\widetilde{f}_i\,(\mathbf{x};\mathbf{x}) = \nabla f_i\,(\mathbf{x})$, for all $\mathbf{x}\in\mathcal{K}$;
E2. $\widetilde{f}_i\,(\bullet;\mathbf{y})$ is uniformly strongly convex on $\mathcal{K}$, with constant τ_i;
E3. $\nabla\widetilde{f}_i\,(\mathbf{x};\bullet)$ is uniformly Lipschitz continuous on $\mathcal{K}$, with constant $\tilde{L}_i$;

where $\nabla\widetilde{f}_i(\mathbf{x};\mathbf{y})$ denotes the partial gradient of $\widetilde{f}_i$ with respect to the first argument, evaluated at $(\mathbf{x},\mathbf{y})$.

Conditions E1–E3 are quite natural: $\widetilde{f}_i$ should be regarded as a (simple) convex, local, approximation of f_i at $\mathbf{x}$ that preserves the first order properties of f_i. A gamut of choices for $\widetilde{f}_i$ satisfying Assumption E are available; some representative examples are discussed in Sect. 4.4.

Step 2: Information mixing and gradient tracking To complete the description of the algorithm, we need to introduce a mechanism to ensure that

(i) the local estimates $\mathbf{x}^n_{(i)}$'s asymptotically converge to a common value; and

(ii) each $\mathbf{y}^n_{(i)}$ tracks the gradient sum $\sum_{j\neq i}\nabla f_j(\mathbf{x}^n_{(i)})$. To this end, we leverage the perturbed condensed push-sum protocol introduced in Sect. 3. Specifically, given $\mathbf{x}^{n+1/2}_{(j)}$'s, each $\mathbf{x}_{(i)}$ is updated according to [cf. (4)]

$$\phi_i^{n+1} = \sum_{j=1}^I a^n_{ij}\phi^n_j,\quad \mathbf{x}^{n+1}_{(i)} = \frac{1}{\phi_i^{n+1}}\sum_{j=1}^I a^n_{ij}\,\phi^n_j\,\mathbf{x}^{n+1/2}_{(j)}, \tag{30}$$

where the a^n_{ij} are chosen to satisfy Assumption C. Note that, the updates in (30) can be performed in a distributed way: each agent j only needs to select the set of weights $\{a^n_{ij}\}_{i=1}^I$ and send $a^n_{ij}\phi^n_j$ and $a^n_{ij}\phi^n_j\mathbf{x}^{n+1/2}_{(j)}$ to its out-neighbors while summing up the information received from its in-neighbors.

To update the $\mathbf{y}_{(i)}^n$'s we leverage again the perturbed condensed push-sum scheme (4), with with $\boldsymbol{\epsilon}_i^{n+1} = (1/\phi_i^{n+1})\left(\nabla f_i(\mathbf{x}_{(i)}^{n+1}) - \nabla f_i(\mathbf{x}_{(i)}^n)\right)$ [cf. (13)]. The resulting gradient tracking mechanism reads

$$\mathbf{y}_{(i)}^{n+1} = \frac{1}{\phi_i^{n+1}} \sum_{j=1}^{I} a_{ij}^n \phi_j^n \, \mathbf{y}_{(j)}^n + \frac{1}{\phi_i^{n+1}} \left(\nabla f_i\left(\mathbf{x}_{(i)}^{n+1}\right) - \nabla f_i\left(\mathbf{x}_{(i)}^n\right)\right), \tag{31}$$

with $\mathbf{y}_{(i)}^0 = \nabla f_i(\mathbf{x}_{(i)}^0)$.

Note that the update of $\mathbf{y}_{(i)}^n$ can be performed locally by agent i, with the same signaling of that of (30).

4.2 The SONATA algorithm

We can now formally introduce the proposed algorithm, SONATA, just combining steps (26), (27), (30), and (31)—see Algorithm 1.

Algorithm 1 SONATA

Data: $\mathbf{x}_{(i)}^0 \in \mathcal{K}$, for all i; $\boldsymbol{\phi}^0 = \mathbf{1}$; $\mathbf{y}^0 = \mathbf{g}^0$. Set $n = 0$.

`[S.1]` If $\mathbf{x}^n$ satisfies termination criterion: STOP;

`[S.2]` `[Distributed Local Optimization]` Each agent i

Compute locally $\widetilde{\mathbf{x}}_{(i)}^n$ solving problem (26);

Update its local variable $\mathbf{x}_{(i)}^{n+1/2} \triangleq \mathbf{x}_{(i)}^n + \alpha^n (\widetilde{\mathbf{x}}_{(i)}^n - \mathbf{x}_{(i)}^n)$;

`[S.3]` `[Information Mixing]` Each agent i compute

(a) `Consensus`

$$\phi_i^{n+1} = \sum_{j=1}^{I} a_{ij}^n \phi_j^n \tag{32}$$

$$\mathbf{x}_{(i)}^{n+1} = \frac{1}{\phi_i^{n+1}} \sum_{j=1}^{I} a_{ij}^n \, \phi_j^n \, \mathbf{x}_{(j)}^{n+1/2}; \tag{33}$$

(b) `Gradient tracking`

$$\mathbf{y}_{(i)}^{n+1} = \frac{1}{\phi_i^{n+1}} \sum_{j=1}^{I} a_{ij}^n \phi_j^n \, \mathbf{y}_{(j)}^n + \frac{1}{\phi_i^{n+1}} \left(\nabla f_i\left(\mathbf{x}_{(i)}^{n+1}\right) - \nabla f_i\left(\mathbf{x}_{(i)}^n\right)\right); \tag{34}$$

`[S.4]` $n \longleftarrow n+1$, go to `[S.1]`

Note that the algorithm is distributed. Indeed, in Step 2, the optimization (26) is performed locally by each agent i, computing its own $\widetilde{\mathbf{x}}_{(i)}^n$. To do so, agent i needs to know the current $\mathbf{x}_{(i)}^n$ and $\mathbf{y}_{(i)}^n$, which are both available locally. There are then two consensus steps (Step 3) whereby agents transmit/receive information only to/from their out/in neighbors: one is on the optimization variables $\mathbf{x}_{(i)}^n$ (and the auxiliary scalars ϕ_i^n)—see (32)–(33)—and one is on the variables $\mathbf{y}_{(i)}^n$—see (34).

4.3 Convergence and complexity analysis of SONATA

To prove convergence, in addition to Assumptions A–E, one needs some conditions on the step-size α^n. Since line-search methods are not practical in a distributed environment, there are two other options, namely: i) a fixed (sufficiently small) step-size; and ii) a diminishing step-size. We prove convergence using either choices. Recalling the definition of the network parameters c_0, $\bar{B}$, $\rho_{\bar{B}}$, ϕ_{lb}, and ϕ_{ub} as given in Proposition 3 [see also (10)] and introducing the problem parameters [cf. Assumptions A, E]

$$
\begin{aligned}
&L \triangleq \sum_{i=1}^{I} L_i, \quad \tilde{L}_{\mathrm{mx}} \triangleq \max_{1\le i\le I} \tilde{L}_i + L_G, \quad L_{\mathrm{mx}} \triangleq \max_{1\le i\le I} L_i,\\
&c_\tau \triangleq \min_{1\le i\le I} \tau_i, \quad c_L \triangleq \left(L\sqrt{I} + L_{\mathrm{mx}} + \tilde{L}_{\mathrm{mx}}\right)/I,
\end{aligned}
\tag{35}
$$

the step-size can be chosen as follows.

Assumption F The step-size $\{\alpha^n\}_{n\in\mathbb{N}_+}$ satisfies either one of the following conditions:

F1. *(diminishing)* $(0, 1] \ni \alpha^n \downarrow 0$ and $\sum_{n=0}^{\infty} \alpha^n = \infty$;
F2. *(fixed)* $\alpha^n \equiv \alpha$, for all $n \in \mathbb{N}_+$, with

$$
\begin{aligned}
\alpha \le \min\Bigg\{ &\frac{\left(1-\rho_{\bar{B}}\right)\sigma}{\sqrt{2}\,c\,\bar{B}},\\
&\frac{2c_\tau\phi_{lb}}{I\phi_{ub}}\left(\frac{L+L_G}{I} + \frac{2c_L\bar{B}c}{1-\rho_{\bar{B}}}\sqrt{\frac{2}{1-\sigma^2}} + \frac{12L_{\mathrm{mx}}\phi_{lb}^{-1}\bar{B}^2c^2}{(1-\rho_{\bar{B}})^2}\sqrt{\frac{1}{1-\sigma^2}}\right)^{-1}\Bigg\},
\end{aligned}
\tag{36}
$$

where σ is an arbitrary constant $\sigma \in (0, 1)$ and $c = I\sqrt{2I}$.

In addition, if all $\mathbf{A}^n$ are double stochastic, the upper bound in (36) holds with $c = 1$, $\bar{B} = B$, $\phi_{lb} = \phi_{ub} = 1$, and $\rho_{\bar{B}} = \left(1 - \kappa/(2\,I^2)\right)^{1/2}$.

We can now state the convergence results of the proposed algorithm, postponing all the proofs to Sect. 6. Given $\{\mathbf{x}^n \triangleq (\mathbf{x}_{(i)}^n)_{i=1}^I\}_{n\in\mathbb{N}_+}$ generated by Algorithm 1, convergence is stated measuring the distance of the average sequence $\bar{\mathbf{x}}^n \triangleq (1/I)\cdot\sum_{i=1}^I \mathbf{x}_{(i)}^n$ from optimality and well as the consensus disagreement among the local variables $\mathbf{x}_{(i)}^n$'s. Distance from stationarity is measured by the following function:

$$
\begin{aligned}
&J(\bar{\mathbf{x}}^n) \triangleq \\
&\left\|\bar{\mathbf{x}}^n - \operatorname*{argmin}_{\mathbf{z}\in\mathcal{K}}\left\{\left(\nabla F(\bar{\mathbf{x}}^n) - \nabla G^-\left(\bar{\mathbf{x}}^n\right)\right)^\top(\mathbf{z}-\bar{\mathbf{x}}^n) + \frac{1}{2}\|\mathbf{z}-\bar{\mathbf{x}}^n\|^2 + G(\mathbf{z})^+\right\}\right\|.
\end{aligned}
\tag{37}
$$

Note that J is a valid measure of stationarity because it is continuous and $J(\bar{\mathbf{x}}^\infty) = 0$ if and only if $\bar{\mathbf{x}}^\infty$ is a d-stationary solution of Problem (P) [16]. The consensus disagreement at iteration n is defined as

$$D(\mathbf{x}^n) \triangleq \|\mathbf{x}^n - \mathbf{1}_I \otimes \bar{\mathbf{x}}^n\|.$$

Note that D is equal to zero if and only if all the $\mathbf{x}_{(i)}^n$'s are consensual. We combine the metrics J and D in a single merit function, defined as

$$M(\mathbf{x}^n) \triangleq \max\left\{J(\bar{\mathbf{x}}^n)^2,\ D(\mathbf{x}^n)^2\right\}.$$

We are now ready to state the main convergence results for Algorithm 1.

Theorem 4 (asymptotic convergence) *Given Problem (P) and Algorithm 1, suppose that Assumptions A–F are satisfied; and let* $\{\mathbf{x}^n\}_{n\in\mathbb{N}_+}$ *be the sequence generated by the algorithm. Then, there holds* $\lim_{n\to\infty} M(\mathbf{x}^n) = 0$.

Proof See Sect. 6. □

Under a constant step-size (Assumption F.2), the next theorem provides an upper bound on the number of iterations needed to decrease $M(\mathbf{x}^n)$ below a given accuracy $\epsilon > 0$.

Theorem 5 (complexity) *Suppose that Assumptions A–E are satisfied; and let* $\{\mathbf{x}^n\}_{n\in\mathbb{N}_+}$ *be the sequence generated by Algorithm 1, with a constant step-size* $\alpha^n = \alpha$, *satisfying Assumption F.2. Given* $\epsilon > 0$, *let* T_ϵ *be the first iteration* n *such that* $M(\mathbf{x}^n) \leq \epsilon$. *Then* $T_\epsilon = \mathcal{O}(1/\epsilon)$.

Proof See the companion paper [40]. □

Remark 6 (*generalizations*) Theorems 4 and 5 can be established with minor modifications under the setting wherein each agent i uses different constant step-size α_i. Also the assumption on the strongly convexity of the surrogate function $\widetilde{f}_i$ (Assumption E.2) can be weakened to just convexity, if the feasible set $\mathcal{K}$ is compact. With mild additional assumptions on G^+—see [12]—we can extend convergence results in Theorem 4 to the case wherein agents solve their subproblems (26) inexactly. We omit further details because of space limitation.

4.4 Discussion

Theorem 4 (resp. Theorem 5) provides the first convergence (resp. complexity) result of distributed algorithms for *constrained* and/or *composite* optimization problems over time-varying (undirected or directed) graphs, which significantly enlarges the class of convex and nonconvex problems which distributed algorithms can be applied to with convergence guarantees.

SONATA represents a gamut of algorithms, each of them corresponding to a specific choice of the surrogate function $\widetilde{f}_i$, step-size α^n, and matrices $\mathbf{A}^n$. Convergence is guaranteed under several choices of the free parameters of the algorithms, some of which are briefly discussed next.

- **On the choice of** $\widetilde{f}_i$. Examples of $\widetilde{f}_i$ satisfying Assumption E are
  - *Linearization* Linearize f_i and add a proximal regularization (to make $\widetilde{f}_i$ strongly convex), which leads to

$$\widetilde{f}_i\left(\mathbf{x}_{(i)};\mathbf{x}_{(i)}^n\right)=f_i\left(\mathbf{x}_{(i)}^n\right)+\nabla f_i\left(\mathbf{x}_{(i)}^n\right)^{\top}\left(\mathbf{x}_{(i)}-\mathbf{x}_{(i)}^n\right)+\frac{\tau_i}{2}\left\|\mathbf{x}_{(i)}-\mathbf{x}_{(i)}^n\right\|_2^2;$$

  - *Partial Linearization* Consider the case where f_i can be decomposed as $f_i(\mathbf{x}_{(i)})=f_i^{(1)}(\mathbf{x}_{(i)})+f_i^{(2)}(\mathbf{x}_{(i)})$, where $f_i^{(1)}$ is convex and $f_i^{(2)}$ is nonconvex with Lipschitz continuous gradient. Preserving the convex part of f_i while linearizing $f_i^{(2)}$ leads to the following valid surrogate

$$\begin{aligned}\widetilde{f}_i(\mathbf{x}_{(i)};\mathbf{x}_{(i)}^n)=f_i^{(1)}(\mathbf{x}_{(i)})+f_i^{(2)}(\mathbf{x}_{(i)}^n)+\frac{\tau_i}{2}\,\|\mathbf{x}_i-\mathbf{x}_{(i)}^n\|^2\\+\nabla f_i^{(2)}(\mathbf{x}_{(i)}^n)^{\top}(\mathbf{x}_i-\mathbf{x}_{(i)}^n);\end{aligned}$$

  - *Partial Convexification* Consider the case where $\mathbf{x}_{(i)}$ is partitioned as $(\mathbf{x}_{(i,1)},\mathbf{x}_{(i,2)})$, and f_i is convex in $\mathbf{x}_{(i,1)}$ but not in $\mathbf{x}_{(i,2)}$. Then, one can convexify only the nonconvex part of f_i, which leads to the surrogate:

$$\begin{aligned}\widetilde{f}_i(\mathbf{x}_{(i)};\mathbf{x}_{(i)}^n)=f_i(\mathbf{x}_{(i,1)},\mathbf{x}_{(i,2)}^n)+\frac{\tau_i}{2}\,\|\mathbf{x}_{(i,2)}-\mathbf{x}_{(i,2)}^n\|^2\\+\nabla^{(2)}f_i(\mathbf{x}_{(i)}^n)^{\top}(\mathbf{x}_{(i,2)}-\mathbf{x}_{(i,2)}^n),\end{aligned}$$

    where $\nabla^{(2)}f_i$ denotes the gradient of f_i with respect to $\mathbf{x}_{(i,2)}$. Other choices of surrogates can be obtained hinging on [15,16,36].

- **On the choice of the step-size** Several options are possible for the step-size sequence $\{\alpha^n\}_n$ satisfying the diminishing-rule in Assumption F.1; see, e.g., [2]. Two instances we found to be effective in our experiments are: i) $\alpha^n=\alpha_0/\,(n+1)^{\beta}$, with $\alpha_0>0$ and $0.5<\beta\le 1$; and ii) $\alpha^n=\alpha^{n-1}\left(1-\mu\alpha^{n-1}\right)$, with $\alpha^0\in(0,1]$, and $\mu\in(0,1)$.
- **On the choice of matrix $\mathbf{A}^n$** When dealing with digraphs, the key requirement of Assumption B is that each $\mathbf{A}^n$ is column stochastic. Such matrices can be built locally by the agents: each agent j can simply choose weight a_{ij}^n for $i\in\mathcal{N}_j^{\text{out}}[n]$ so that $\sum_{i\in\mathcal{N}_j^{\text{out}}[n]}a_{ij}^n=1$. As a special case, $\mathbf{A}^n$ can be set to be the following push-sum matrix [24]: $a_{ij}^n=1/d_j^n$, if $(j,i)\in\mathcal{E}^n$; and $a_{ij}^n=0$, otherwise; where d_i^n is the out-degree of agent i. In this case, the information mixing process in Step 2 becomes a broadcasting protocol, which requires from each agent only the knowledge of its out-degree.

When the digraphs $\mathcal{G}^n$ admit a double-stochastic matrix (e.g., they are *undirected*), as already observed in Sect. 3 (cf. Remark 1), one can choose $\mathbf{A}^n$ as double-stochastic; and the consensus and tracking protocols in Step 3 reduce respectively to

$$
\begin{aligned}
\mathbf{x}_{(i)}^{n+1} &= \sum_{j=1}^{I} a_{ij}^{n} \left(\mathbf{x}_{(j)}^{n} + \alpha^{n} (\widetilde{\mathbf{x}}_{(j)}^{n} - \mathbf{x}_{(j)}^{n}) \right) \\
\mathbf{y}_{(i)}^{n+1} &= \sum_{j=1}^{I} a_{ij}^{n} \mathbf{y}_{(j)}^{n} + \nabla f_i(\mathbf{x}_{(i)}^{n+1}) - \nabla f_i(\mathbf{x}_{(i)}^{n}).
\end{aligned}
\tag{38}
$$

Several choices have been proposed in the literature to build in a distributed way a double stochastic matrix $\mathbf{A}^n$, including: the Laplacian, Metropolis–Hastings, and maximum-degree weights; see, e.g., [51].

• **ATC/CAA updates.** In the case of unconstrained optimization, the information mixing step in Algorithm 1 can be performed following two alternative protocols, namely: i) the *Adapt-Then-Combine-based* (ATC) scheme; and ii) the *Combine-And-Adapt-based* (CAA) approach (termed "consensus strategy" in [35]). The former is the one used in (30)—each agent i first updates its local copy $\mathbf{x}_{(i)}^n$ along the direction $\widetilde{\mathbf{x}}_{(i)}^n - \mathbf{x}_{(i)}^n$, and then combines its new update with that of its neighbors via consensus. Alternatively, in the CAA update, agent i first mixes its own local copy $\mathbf{x}_{(i)}^n$ with that of its neighbors via consensus, and then performs its local optimization-based update using $\widetilde{\mathbf{x}}_{(i)}^n - \mathbf{x}_{(i)}^n$, that is

$$
\mathbf{x}_{(i)}^{n+1} = \frac{1}{\phi_i^{n+1}} \sum_{j=1}^{I} a_{ij} \phi_j^n \mathbf{x}_{(j)}^n + \frac{\phi_i^n}{\phi_i^{n+1}} \cdot \alpha^n (\widetilde{\mathbf{x}}_{(i)}^n - \mathbf{x}_{(i)}^n).
$$

It is not difficult to check that SONATA based on CAA updates converges under the same conditions as in Theorem 4.

5 SONATA and special cases

In this section, we contrast SONATA with related algorithms proposed in the literature [12–14,52] and very recent proposals [30,33,50] for *special* instances of Problem (P). We show that algorithms in [30,33,50,52] are all special cases of SONATA and NEXT, proposed in our earlier works [12–14,42].

We preliminarily rewrite Algorithm 1 in a matrix–vector form. Similarly to $\mathbf{x}^n$, define the concatenated vectors

$$
\widetilde{\mathbf{x}}^n \triangleq \left[\widetilde{\mathbf{x}}_{(1)}^{n\top}, \ldots, \widetilde{\mathbf{x}}_{(I)}^{n\top} \right]^\top, \tag{39}
$$

$$
\mathbf{y}^n \triangleq \left[\mathbf{y}_{(1)}^{n\top}, \ldots, \mathbf{y}_{(I)}^{n\top} \right]^\top, \tag{40}
$$

$$
\mathbf{g}^n \triangleq \left[\mathbf{g}_1^{n\top}, \ldots, \mathbf{g}_I^{n\top} \right]^\top, \quad \mathbf{g}_i^n \triangleq \nabla f_i(\mathbf{x}_{(i)}^n), \tag{41}
$$

$$
\Delta\mathbf{x}^n \triangleq \widetilde{\mathbf{x}}^n - \mathbf{x}^n, \tag{42}
$$

where $\widetilde{\mathbf{x}}^n_{(i)}$ and $\mathbf{y}^n_{(i)}$ are defined in (26) and (34), respectively. Using the above notation and the matrices introduced in (14a), SONATA [cf. (32)–(34)] can be written in compact form as

$$\boldsymbol{\phi}^{n+1} = \mathbf{A}^n \boldsymbol{\phi}^n \tag{43}$$

$$\mathbf{x}^{n+1} = \widehat{\mathbf{W}}^n (\mathbf{x}^n + \alpha^n \Delta \mathbf{x}^n) \tag{44}$$

$$\mathbf{y}^{n+1} = \widehat{\mathbf{W}}^n \mathbf{y}^n + (\widehat{\mathbf{D}}_{\boldsymbol{\phi}^{n+1}})^{-1} \left(\mathbf{g}^{n+1} - \mathbf{g}^n\right). \tag{45}$$

5.1 Preliminaries: NEXT and SONATA-L

Since [30,33,50,52] are applicable only to *unconstrained* ($\mathcal{K} = \mathbb{R}^m$), *smooth* ($G = 0$) and *convex* (each f_i is convex) multiagent problems, in the following, we consider only such an instance of Problem (P). Choose each $\widetilde{f}_i$ as first order approximation of f_i plus a proximal term, that is,

$$\widetilde{f}_i(\mathbf{x}_{(i)}; \mathbf{x}^n_{(i)}) = f_i(\mathbf{x}^n_{(i)}) + \nabla f_i(\mathbf{x}^n_{(i)})^\top (\mathbf{x}_{(i)} - \mathbf{x}^n_{(i)}) + \frac{\tau_i}{2} \|\mathbf{x}_{(i)} - \mathbf{x}^n_{(i)}\|^2,$$

and set $\tau_i = I$. Then, $\widetilde{\mathbf{x}}^n_{(i)}$ can be computed in closed form [cf. (26)]:

$$\begin{aligned} \widetilde{\mathbf{x}}^n_{(i)} &= \underset{\mathbf{x}_{(i)}}{\operatorname{argmin}} \, (I \cdot \mathbf{y}^n_{(i)})^\top (\mathbf{x}_{(i)} - \mathbf{x}^n_{(i)}) + \frac{I}{2} \|\mathbf{x}_{(i)} - \mathbf{x}^n_{(i)}\|^2 \\ &= \underset{\mathbf{x}_{(i)}}{\operatorname{argmin}} \, \frac{I}{2} \left\|\mathbf{x}_{(i)} - \mathbf{x}^n_{(i)} + \mathbf{y}^n_{(i)}\right\|^2 = \mathbf{x}^n_{(i)} - \mathbf{y}^n_{(i)}. \end{aligned} \tag{46}$$

Therefore, $\Delta \mathbf{x}^n_{(i)} = \widetilde{\mathbf{x}}^n_{(i)} - \mathbf{x}^n_{(i)} = \mathbf{y}^n_{(i)}$.

Substituting (46) into (44) and using either ATC or CAA mixing protocols, Algorithm 1 reduces to

$$\boldsymbol{\phi}^{n+1} = \mathbf{A}^n \, \boldsymbol{\phi}^n$$

$$\mathbf{x}^{n+1} = \begin{cases} \widehat{\mathbf{W}}^n \, (\mathbf{x}^n - \alpha^n \, \mathbf{y}^n) & \text{(ATC-based update)} \\ \widehat{\mathbf{W}}^n \mathbf{x}^n - \alpha^n \, \left(\widehat{\mathbf{D}}_{\boldsymbol{\phi}^{n+1}}\right)^{-1} \widehat{\mathbf{D}}_{\boldsymbol{\phi}^n} \mathbf{y}^n & \text{(CAA-based update)} \end{cases} \tag{47}$$

$$\mathbf{y}^{n+1} = \begin{cases} \widehat{\mathbf{W}}^n \left(\mathbf{y}^n + (\widehat{\mathbf{D}}_{\boldsymbol{\phi}^n})^{-1} (\mathbf{g}^{n+1} - \mathbf{g}^n)\right) & \text{(ATC-based update)} \\ \widehat{\mathbf{W}}^n \mathbf{y}^n + \left(\widehat{\mathbf{D}}_{\boldsymbol{\phi}^{n+1}}\right)^{-1} \left(\mathbf{g}^{n+1} - \mathbf{g}^n\right); & \text{(CAA-based update)} \end{cases} \tag{48}$$

which we will refer to as *(ATC/CAA-)SONATA-L* (L stands for “linearized”).

When the digraph $\mathcal{G}^n$ admits a *double-stochastic* matrix $\mathbf{A}^n$, and $\mathbf{A}^n$ in (43) is chosen so, the iterates (47) can be further simplified as reduces to

$$\mathbf{x}^{n+1} = \begin{cases} \widehat{\mathbf{W}}^n \left(\mathbf{x}^n - \alpha^n \mathbf{y}^n\right) & \text{(ATC-based update)} \\ \widehat{\mathbf{W}}^n \mathbf{x}^n - \alpha^n \mathbf{y}^n & \text{(CAA-based update)} \end{cases} \tag{49}$$

$$\mathbf{y}^{n+1} = \begin{cases} \widehat{\mathbf{W}}^n \left(\mathbf{y}^n + \mathbf{g}^{n+1} - \mathbf{g}^n\right) & \text{(ATC-based update)} \\ \widehat{\mathbf{W}}^n \mathbf{y}^n + \mathbf{g}^{n+1} - \mathbf{g}^n, & \text{(CAA-based update)} \end{cases} \tag{50}$$

where $\mathbf{W}^n = \mathbf{A}^n$ and thus $\widehat{\mathbf{W}}^n = \mathbf{W}^n \otimes \mathbf{I}_m$. The ATC-based updates coincide with our previous algorithm NEXT [based on the surrogate (46)], introduced in [12–14]. We will refer to (49) as *(ATC/CAA-)NEXT-L*.

5.2 Connection with current algorithms

We can now show that the algorithms recently studied in [30,33,50,52] are all special cases of SONATA and NEXT, earlier proposed in [12–14].

Aug-DGM [52] **and Algorithm in** [33]. Introduced in [52] for *undirected, time-invariant* graphs, the Aug-DGM algorithm reads

$$\begin{aligned} \mathbf{x}^{n+1} &= \widehat{\mathbf{W}} \left(\mathbf{x}^n - \text{Diag}\left(\boldsymbol{\alpha} \otimes \mathbf{1}_m\right) \mathbf{y}^n\right) \\ \mathbf{y}^{n+1} &= \widehat{\mathbf{W}} \left(\mathbf{y}^n + \mathbf{g}^{n+1} - \mathbf{g}^n\right) \end{aligned} \tag{51}$$

where $\widehat{\mathbf{W}} \triangleq \mathbf{W} \otimes \mathbf{I}_m$; $\mathbf{W}$ is a double stochastic matrix satisfying Assumption C, and $\boldsymbol{\alpha}$ is the vector of agents' step-sizes α_i's.

A similar algorithm was proposed independently in [33] (in the same networking setting of [52]), which reads

$$\begin{aligned} \mathbf{x}^{n+1} &= \widehat{\mathbf{W}} \left(\mathbf{x}^n - \alpha \mathbf{y}^n\right) \\ \mathbf{y}^{n+1} &= \widehat{\mathbf{W}} \mathbf{y}^n + \mathbf{g}^{n+1} - \mathbf{g}^n. \end{aligned} \tag{52}$$

6 Convergence Proof of SONATA

Clearly Aug-DGM [52] in (51) with the α_i's equal, and Algorithm [33] in (52) coincide with (ATC-)NEXT-L [cf. (49)].

(Push-)DIGing [30]. Appeared in [30] and applicable to *B-strongly connected undirected graphs*, the DIGing Algorithm reads

$$\begin{aligned} \mathbf{x}^{n+1} &= \widehat{\mathbf{W}}^n \mathbf{x}^n - \alpha \mathbf{y}^n \\ \mathbf{y}^{n+1} &= \widehat{\mathbf{W}}^n \mathbf{y}^n + \mathbf{g}^{n+1} - \mathbf{g}^n, \end{aligned} \tag{53}$$

where $\mathbf{W}^n$ is a double-stochastic matrix satisfying Assumption C. Clearly, DIGing coincides with (CAA-)NEXT-L [12–14][cf. (49)]. The push-DIGing algorithm, studied in the same paper [30], extends DIGing to B-strongly connected digraphs. It

Table 4 Connection of SONATA with current algorithms

Algorithms	Connection with SONATA	Instance of Problem (P)	Graph topology/ Weight matrix
NEXT [12]	special case of SONATA (38)	F nonconvex $G \neq 0$ $\mathcal{K} \subseteq \mathbb{R}^m$	time-varying digraph/ doubly-stochastic weights
Aug-DGM [33, 52]	ATC-NEXT-L ($\boldsymbol{\alpha} = \alpha \mathbf{1}_I$) (49)	F convex $G = 0$ $\mathcal{K} = \mathbb{R}^m$	static undirected graph/ doubly-stochastic weights
DIGing [30]	CAA-NEXT-L (49)	F convex $G = 0$ $\mathcal{K} = \mathbb{R}^m$	time-varying digraph/ doubly-stochastic weights
push-DIGing [30]	ATC-SONATA-L (47)	F convex $G = 0$ $\mathcal{K} = \mathbb{R}^m$	time-varying digraph/ column-stochastic weights
ADD-OPT [50]	ATC-SONATA-L (47)	F convex $G = 0$ $\mathcal{K} = \mathbb{R}^m$	static digraph/ column-stochastic weights

turns out that push-DIGing coincides with (ATC-)SONATA-L [cf. Eq. (47)] when $a_{ij}^n = 1/d_j^n$.

ADD-OPT [50]. Finally, we mention the ADD-OPT algorithm, proposed in [50] for *strongly connected static digraphs*, which takes the following form:

$$\begin{aligned}
\mathbf{z}^{n+1} &= \widehat{\mathbf{A}}\mathbf{z}^n - \alpha\widetilde{\mathbf{y}}^n \\
\boldsymbol{\phi}^{n+1} &= \mathbf{A}\,\boldsymbol{\phi}^n \\
\mathbf{x}^{n+1} &= \left(\widehat{\mathbf{D}}_{\boldsymbol{\phi}^{n+1}}\right)^{-1}\mathbf{z}^{n+1} \\
\widetilde{\mathbf{y}}^{n+1} &= \widehat{\mathbf{A}}\,\widetilde{\mathbf{y}}^n + \mathbf{g}^{n+1} - \mathbf{g}^n,
\end{aligned} \tag{54}$$

where $\mathbf{A}$ is a column stochastic matrix satisfying Assumption C, and $\widehat{\mathbf{A}} = \mathbf{A} \otimes \mathbf{I}_m$. Defining $\mathbf{y}^n = (\widehat{\mathbf{D}}_{\boldsymbol{\phi}^{n+1}})^{-1}\widetilde{\mathbf{y}}^n$, it is not difficult to check that (54) can be rewritten as

$$\begin{aligned}
\boldsymbol{\phi}^{n+1} &= \mathbf{A}\boldsymbol{\phi}^n, \quad \mathbf{W} = \left(\widehat{\mathbf{D}}_{\boldsymbol{\phi}^{n+1}}\right)^{-1}\widehat{\mathbf{A}}\,\widehat{\mathbf{D}}_{\boldsymbol{\phi}^n} \\
\mathbf{x}^{n+1} &= \widehat{\mathbf{W}}\mathbf{x}^n - \alpha\left(\widehat{\mathbf{D}}_{\boldsymbol{\phi}^{n+1}}\right)^{-1}\widehat{\mathbf{D}}_{\boldsymbol{\phi}^n}\,\mathbf{y}^n \\
\mathbf{y}^{n+1} &= \widehat{\mathbf{W}}\mathbf{y}^n + \left(\widehat{\mathbf{D}}_{\boldsymbol{\phi}^{n+1}}\right)^{-1}\left(\mathbf{g}^{n+1} - \mathbf{g}^n\right).
\end{aligned} \tag{55}$$

Comparing Eqs. (47) and (55), one can see that ADD-OPT coincides with (CAA-)SONATA-L.

We summarize the connections between the different versions of SONATA(-NEXT) and its special cases in Table 4.

In this section, we prove convergence of SONATA; because of space limitation we prove only Theorem 4; the proof of Theorem 5 can be found in the companion paper [40]. The proof consists in studying the dynamics of a suitably chosen Lyapunov function along the weighted average of the agents' local copies, and of the consensus disagreement and tracking errors. We begin introducing some convenient notation along with some preliminary results. For the sake of simplicity, all the results of the forthcoming subsections are stated under Assumptions A–F.

6.1 Notations and preliminaries

The weighted average and associated consensus disagreement are denoted by

$$\bar{\mathbf{x}}_{\boldsymbol{\phi}^n} \triangleq \frac{1}{I}\left(\boldsymbol{\phi}^{n\top} \otimes \mathbf{I}_m\right)\mathbf{x}^n \quad \text{and} \quad \mathbf{e}_x^n \triangleq \mathbf{x}^n - \mathbf{J}_{\boldsymbol{\phi}^n}\mathbf{x}^n, \tag{56}$$

respectively. Similar quantities are defined for the tracking variables $\mathbf{y}_{(i)}^n$:

$$\bar{\mathbf{y}}_{\boldsymbol{\phi}^n} \triangleq \frac{1}{I}\left(\boldsymbol{\phi}^{n\top} \otimes \mathbf{I}_m\right)\mathbf{y}^n \quad \text{and} \quad \mathbf{e}_y^n \triangleq \mathbf{y}^n - \mathbf{J}_{\boldsymbol{\phi}^n}\mathbf{y}^n. \tag{57}$$

Recalling (39), define the deviation of the local solution $\widetilde{\mathbf{x}}_{(i)}^n$ of each agent from the weighted average as

$$\Delta\widetilde{\mathbf{x}}_{(i),\boldsymbol{\phi}}^n \triangleq \widetilde{\mathbf{x}}_{(i)}^n - \bar{\mathbf{x}}_{\boldsymbol{\phi}^n}, \tag{58}$$

and the associated stacked vector

$$\Delta\widetilde{\mathbf{x}}_{\boldsymbol{\phi}}^n \triangleq \widetilde{\mathbf{x}}^n - \mathbf{J}_{\boldsymbol{\phi}^n}\mathbf{x}^n. \tag{59}$$

Note that $\Delta\mathbf{x}^n$ [cf. (39)] can be rewritten as

$$\Delta\mathbf{x}^n = \Delta\widetilde{\mathbf{x}}_{\boldsymbol{\phi}}^n - \mathbf{e}_x^n. \tag{60}$$

Using the above notation, the dynamics of $\bar{\mathbf{x}}_{\boldsymbol{\phi}^n}$ and $\bar{\mathbf{y}}_{\boldsymbol{\phi}^n}$ generated by Algorithm 1 are given by [cf. (44) and (45)]:

$$\bar{\mathbf{x}}_{\boldsymbol{\phi}^{n+1}} = \bar{\mathbf{x}}_{\boldsymbol{\phi}^n} + \frac{\alpha^n}{I}\left((\boldsymbol{\phi}^n)^\top \otimes \mathbf{I}_m\right)\Delta\widetilde{\mathbf{x}}^n_{\boldsymbol{\phi}} \tag{61a}$$

$$\bar{\mathbf{y}}_{\boldsymbol{\phi}^{n+1}} = \bar{\mathbf{y}}_{\boldsymbol{\phi}^n} + \bar{\mathbf{g}}^{n+1} - \bar{\mathbf{g}}^n. \tag{61b}$$

Note that, since $\mathbf{y}^0 = \mathbf{g}^0$ and $\phi_i^0 = 1$, we have $\bar{\mathbf{y}}_{\boldsymbol{\phi}^n} = \bar{\mathbf{g}}^n$, for all $n \in \mathbb{N}_+$.

Finally, we introduce the error-free local solution map of each agent i, denoted by $\widehat{\mathbf{x}}_{(i)} : \mathcal{K} \to \mathcal{K}$: Given $\mathbf{z} \in \mathcal{K}$ and $i = 1, \ldots, I$, let

$$\begin{aligned}\widehat{\mathbf{x}}_{(i)}(\mathbf{z}) \triangleq \underset{\mathbf{x}_{(i)}\in\mathcal{K}}{\operatorname{argmin}} \Big\{ & \widetilde{f}_i\left(\mathbf{x}_{(i)}; \mathbf{z}\right) - \nabla G^-(\mathbf{z})^\top\left(\mathbf{x}_{(i)} - \mathbf{z}\right) \\ & + \Big(\textstyle\sum_{j\neq i}\nabla f_j(\mathbf{z})\Big)^\top\left(\mathbf{x}_{(i)} - \mathbf{z}\right) + G^+(\mathbf{z})\Big\}.\end{aligned} \tag{62}$$

It is not difficult to check that $\widehat{\mathbf{x}}_{(i)}(\bullet)$ enjoys the following properties (the proof of the next lemma follows similar steps as in [16, Prop. 8] and thus is omitted).

Lemma 7 *Each $\widehat{\mathbf{x}}_{(i)}(\bullet)$ satisfies:*

i) *[Lipschitz continuity] $\widehat{\mathbf{x}}_{(i)}(\bullet)$ is $\hat{L}$-Lipschitz continuous on $\mathcal{K}$, that is, there exits a finite $\hat{L} > 0$ such that*

$$\left\|\widehat{\mathbf{x}}_{(i)}(\mathbf{z}) - \widehat{\mathbf{x}}_{(i)}(\mathbf{w})\right\| \leq \hat{L}\,\|\mathbf{z} - \mathbf{w}\|, \quad \forall \mathbf{z},\ \mathbf{w} \in \mathcal{K}; \tag{63}$$

ii) *[Fixed-points] The set of fixed points of $\widehat{\mathbf{x}}_{(i)}(\bullet)$ coincides with the set of d-stationary solutions of Problem (P).*

The next result shows that, as expected, the disagreement between agent i's solution $\widetilde{\mathbf{x}}^n_{(i)}$ and its error-free counterpart $\widehat{\mathbf{x}}_{(i)}(\mathbf{x}^n_{(i)})$ asymptotically vanishes if both the consensus error $\mathbf{e}^n_x$ and the tracking error $\mathbf{e}^n_y$ do so.

Lemma 8 $\widetilde{\mathbf{x}}^n_{(i)}$ *[cf. (26)] and $\widehat{\mathbf{x}}_i(\mathbf{x}^n_{(i)})$ [cf. (62)] satisfy:*

$$\left\|\widehat{\mathbf{x}}_i(\mathbf{x}^n_{(i)}) - \widetilde{\mathbf{x}}^n_{(i)}\right\| \leq \frac{I}{\tau_i}\left\|\mathbf{e}^n_y\right\| + \frac{2\,I\,L}{\tau_i}\|\mathbf{e}^n_x\|. \tag{64}$$

Therefore, $\|\mathbf{e}^n_x\|, \|\mathbf{e}^n_y\| \underset{n\to\infty}{\longrightarrow} 0 \Rightarrow \|\widehat{\mathbf{x}}_i(\mathbf{x}^n_{(i)}) - \widetilde{\mathbf{x}}^n_{(i)}\| \underset{n\to\infty}{\longrightarrow} 0.$

The last result of this section is a standard martingale-like result; the proof follows similar to that of [3, Lemma1] and thus is omitted.

Lemma 9 *Let $\{X^n\}_{n\in\mathbb{N}_+}$, $\{Y^n\}_{n\in\mathbb{N}_+}$ and $\{Z^n\}_{n\in\mathbb{N}_+}$ be three sequences such that X^n and Z^n are nonnegative, for all $n \in \mathbb{N}_+$. Suppose that*

$$\sum_{k=0}^{\bar{B}-1} Y^{n+\bar{B}+k} \leq \sum_{k=0}^{\bar{B}-1} Y^{n+k} - \sum_{k=0}^{\bar{B}-1} X^{n+k} + \sum_{k=0}^{\bar{B}-1} Z^{n+k}, \quad n = 0, 1, \ldots, \tag{65}$$

and that $\sum_{n=0}^{\infty} Z^n < +\infty$. Then, either $\sum_{k=0}^{\bar{B}-1} Y^{n+k} \to -\infty$, or else $\sum_{k=0}^{\bar{B}-1} Y^{n+k}$ converges to a finite value and $\sum_{n=0}^{\infty} X^n < +\infty$.

6.2 Average descent

We begin our analysis studying the dynamics of U along the trajectory of $\bar{\mathbf{x}}_{\boldsymbol{\phi}^n}$. We define the total energy of the optimization input $\alpha^n \Delta\widetilde{\mathbf{x}}_{\boldsymbol{\phi}}^n$ and consensus errors $\mathbf{e}_x^n$ and $\mathbf{e}_y^n$ in $\bar{B}$ consecutive iterations [$\bar{B}$ is defined in Lemma 3]:

$$E_{\Delta\widetilde{\mathbf{x}}}^n \triangleq \sum_{t=0}^{\bar{B}-1} \left(\alpha^{n+t}\right)^2 \left\| \Delta\widetilde{\mathbf{x}}_{\boldsymbol{\phi}}^{n+t} \right\|^2, \quad E_{\mathbf{x}_\perp}^n \triangleq \sum_{t=0}^{\bar{B}-1} \left\| \mathbf{e}_x^{n+t} \right\|^2, \quad E_{\mathbf{y}_\perp}^n \triangleq \sum_{t=0}^{\bar{B}-1} \left\| \mathbf{e}_y^{n+t} \right\|^2. \tag{66}$$

Recalling the definitions of c_τ, ϕ_{lb}, and ϕ_{ub} [see (8) and (35)], we have the following.

Lemma 10 *Let $\{(\mathbf{x}^n, \mathbf{y}^n)\}_{n\in\mathbb{N}_+}$ be the sequence generated by Algorithm 1. Then, there holds*

$$\begin{aligned}
&\sum_{k=0}^{\bar{B}-1} U\left(\bar{\mathbf{x}}_{\boldsymbol{\phi}^{n+\bar{B}+k}}\right) \\
&\quad\le \sum_{k=0}^{\bar{B}-1} U\left(\bar{\mathbf{x}}_{\boldsymbol{\phi}^{n+k}}\right) - \frac{c_\tau\,\phi_{lb}}{I} \cdot \sum_{k=0}^{\bar{B}-1}\sum_{t=0}^{\bar{B}-1} \alpha^{n+k+t} \left\| \Delta\widetilde{\mathbf{x}}_{\boldsymbol{\phi}}^{n+k+t} \right\|^2 \\
&\qquad + \frac{\phi_{ub}}{2}\left(\frac{L+L_G}{I}\,\phi_{ub} + c_L\epsilon_x + \epsilon_y\right)\sum_{k=0}^{\bar{B}-1} E_{\Delta\widetilde{\mathbf{x}}}^{n+k} + \underbrace{\frac{\phi_{ub}}{2}\sum_{k=0}^{\bar{B}-1}\left(c_L\epsilon_x^{-1} E_{\mathbf{x}_\perp}^{n+k} + \epsilon_y^{-1} E_{\mathbf{y}_\perp}^{n+k}\right)}_{\text{term iv}},
\end{aligned} \tag{67}$$

where $\epsilon_x > 0$ and $\epsilon_y > 0$ are arbitrary, finite constants.

Proof Denote for simplicity $\bar{F} \triangleq F - G^-$. Since $\widetilde{f}_i$ is strongly convex and G^+ is convex, by the first order optimality of $\widetilde{\mathbf{x}}_{(i)}^n$, we have

$$\begin{aligned}
&\left(\Delta\widetilde{\mathbf{x}}_{(i),\boldsymbol{\phi}}^n\right)^\top \left(I \cdot \mathbf{y}_{(i)}^n + \nabla \widetilde{f}_i(\bar{\mathbf{x}}_{\boldsymbol{\phi}^n}; \mathbf{x}_{(i)}^n) - \nabla G^-(\mathbf{x}_{(i)}^n) - \nabla f_i(\mathbf{x}_{(i)}^n)\right) \\
&\quad + G^+(\widetilde{\mathbf{x}}_{(i)}^n) - G^+(\bar{\mathbf{x}}_{\boldsymbol{\phi}^n}) \le -\tau_i \|\Delta\widetilde{\mathbf{x}}_{(i),\boldsymbol{\phi}}^n\|^2.
\end{aligned} \tag{68}$$

Since ∇f_i and ∇G^- are L_i and L_G-Lipschitz, respectively, ∇F is $(L + L_G)$-Lipschitz, where $L \triangleq \sum_{i=1}^I L_i$ [cf. def. (35)]. Applying the descent lemma to $\bar{F}$ and using (61a) yields

$$\begin{aligned}
&\bar{F}\left(\bar{\mathbf{x}}_{\boldsymbol{\phi}^{n+1}}\right) \\
&\quad\le \bar{F}\left(\bar{\mathbf{x}}_{\boldsymbol{\phi}^n}\right) + \frac{\alpha^n}{I}\nabla\bar{F}\left(\bar{\mathbf{x}}_{\boldsymbol{\phi}^n}\right)^\top \left((\boldsymbol{\phi}^n)^\top \otimes \mathbf{I}_m\right)\Delta\widetilde{\mathbf{x}}_{\boldsymbol{\phi}}^n \\
&\qquad + \frac{L+L_G}{2}\cdot\frac{(\alpha^n)^2}{I}\left\|\left((\boldsymbol{\phi}^n)^\top \otimes \mathbf{I}_m\right)\Delta\widetilde{\mathbf{x}}_{\boldsymbol{\phi}}^n\right\|^2
\end{aligned}$$

$$
\begin{aligned}
&\overset{(a)}{\leq} \bar{F}\left(\bar{\mathbf{x}}_{\boldsymbol{\phi}^n}\right) + \frac{L+L_G}{2} \cdot \frac{(\alpha^n)^2}{I}\, \phi_{ub}^2 \left\| \Delta \widetilde{\mathbf{x}}_{\boldsymbol{\phi}}^n \right\|^2 \\
&\quad - \frac{\alpha^n}{I} \sum_{i=1}^{I} \phi_i^n \left(\tau_i \left\| \Delta \widetilde{\mathbf{x}}_{(i),\boldsymbol{\phi}}^n \right\|^2 + G^+(\widetilde{\mathbf{x}}_{(i)}^n) - G^+(\bar{\mathbf{x}}_{\boldsymbol{\phi}^n}) \right) \\
&\quad + \frac{\alpha^n}{I} \sum_{i=1}^{I} \phi_i^n \left(\nabla \bar{F}\left(\bar{\mathbf{x}}_{\boldsymbol{\phi}^n}\right) + \nabla G^-(\mathbf{x}_{(i)}^n) - I \cdot \bar{\mathbf{y}}_{\boldsymbol{\phi}^n} + I \cdot \bar{\mathbf{y}}_{\boldsymbol{\phi}^n} - I \cdot \mathbf{y}_{(i)}^n \right)^\top \Delta \widetilde{\mathbf{x}}_{(i),\boldsymbol{\phi}}^n \\
&\quad + \frac{\alpha^n}{I} \sum_{i=1}^{I} \phi_i^n \left(\nabla f_i(\mathbf{x}_{(i)}^n) - \nabla f_i(\bar{\mathbf{x}}_{\boldsymbol{\phi}^n}) + \nabla \widetilde{f}_i(\bar{\mathbf{x}}_{\boldsymbol{\phi}^n}; \bar{\mathbf{x}}_{\boldsymbol{\phi}^n}) - \nabla \widetilde{f}_i(\bar{\mathbf{x}}_{\boldsymbol{\phi}^n}; \mathbf{x}_{(i)}^n) \right)^\top \Delta \widetilde{\mathbf{x}}_{(i),\boldsymbol{\phi}}^n \\
&\overset{(b)}{\leq} \bar{F}\left(\bar{\mathbf{x}}_{\boldsymbol{\phi}^n}\right) + \frac{L+L_G}{2} \cdot \frac{(\alpha^n)^2}{I}\, \phi_{ub}^2 \| \Delta \widetilde{\mathbf{x}}_{\boldsymbol{\phi}}^n \|^2 \\
&\quad - \frac{\alpha^n}{I} \sum_{i=1}^{I} \phi_i^n \left(\tau_i \| \Delta \widetilde{\mathbf{x}}_{(i),\boldsymbol{\phi}}^n \|^2 + G^+(\widetilde{\mathbf{x}}_{(i)}^n) - G^+(\bar{\mathbf{x}}_{\boldsymbol{\phi}^n}) \right) \\
&\quad + \frac{\alpha^n}{I} \sum_{i=1}^{I} \phi_i^n \left\| \nabla \bar{F}(\bar{\mathbf{x}}_{\boldsymbol{\phi}^n}) - \left(\sum_{j=1}^{I} \nabla f_j(\mathbf{x}_{(j)}^n) - \nabla G^-(\mathbf{x}_{(i)}^n) \right) \right\| \left\| \Delta \widetilde{\mathbf{x}}_{(i),\boldsymbol{\phi}}^n \right\| \\
&\quad + \alpha^n \sum_{i=1}^{I} \phi_i^n \| \bar{\mathbf{y}}_{\boldsymbol{\phi}^n} - \mathbf{y}_{(i)}^n \| \, \| \Delta \widetilde{\mathbf{x}}_{(i),\boldsymbol{\phi}}^n \| \\
&\quad + \frac{\alpha^n}{I} \sum_{i=1}^{I} \phi_i^n \left\| \nabla f_i(\mathbf{x}_{(i)}^n) - \nabla f_i\left(\bar{\mathbf{x}}_{\boldsymbol{\phi}^n}\right) \right\| \, \| \Delta \widetilde{\mathbf{x}}_{(i),\boldsymbol{\phi}}^n \| \\
&\quad + \frac{\alpha^n}{I} \sum_{i=1}^{I} \phi_i^n \left\| \nabla \widetilde{f}_i(\bar{\mathbf{x}}_{\boldsymbol{\phi}^n}; \bar{\mathbf{x}}_{\boldsymbol{\phi}^n}) - \nabla \widetilde{f}_i(\bar{\mathbf{x}}_{\boldsymbol{\phi}^n}; \mathbf{x}_{(i)}^n) \right\| \, \| \Delta \widetilde{\mathbf{x}}_{(i),\boldsymbol{\phi}}^n \| \\
&\overset{(c)}{\leq} \bar{F}\left(\bar{\mathbf{x}}_{\boldsymbol{\phi}^n}\right) + \frac{L+L_G}{2} \cdot \frac{(\alpha^n)^2}{I}\, \phi_{ub}^2 \| \Delta \widetilde{\mathbf{x}}_{\boldsymbol{\phi}}^n \|^2 \\
&\quad - \frac{\alpha^n}{I} \sum_{i=1}^{I} \phi_i^n \left(\tau_i \| \Delta \widetilde{\mathbf{x}}_{(i),\boldsymbol{\phi}}^n \|^2 + G^+(\widetilde{\mathbf{x}}_{(i)}^n) - G^+(\bar{\mathbf{x}}_{\boldsymbol{\phi}^n}) \right) \\
&\quad + \frac{\alpha^n}{I} \sum_{i=1}^{I} \phi_i^n \left(\sum_{j=1}^{I} L_j \| \bar{\mathbf{x}}_{\boldsymbol{\phi}^n} - \mathbf{x}_{(j)}^n \| + L_G \| \bar{\mathbf{x}}_{\boldsymbol{\phi}^n} - \mathbf{x}_{(i)}^n \| \right) \| \Delta \widetilde{\mathbf{x}}_{(i),\boldsymbol{\phi}}^n \| \\
&\quad + \alpha^n \sum_{i=1}^{I} \phi_i^n \| \bar{\mathbf{y}}_{\boldsymbol{\phi}^n} - \mathbf{y}_{(i)}^n \| \| \Delta \widetilde{\mathbf{x}}_{(i),\boldsymbol{\phi}}^n \| \\
&\quad + \frac{\alpha^n}{I} \sum_{i=1}^{I} \phi_i^n \left(L_i \| \mathbf{x}_{(i)}^n - \bar{\mathbf{x}}_{\boldsymbol{\phi}^n} \| \, \| \Delta \widetilde{\mathbf{x}}_{(i),\boldsymbol{\phi}}^n \| + \tilde{L}_i \| \mathbf{x}_{(i)}^n - \bar{\mathbf{x}}_{\boldsymbol{\phi}^n} \| \, \| \Delta \widetilde{\mathbf{x}}_{(i),\boldsymbol{\phi}}^n \| \right) \\
&\overset{(d)}{\leq} \bar{F}\left(\bar{\mathbf{x}}_{\boldsymbol{\phi}^n}\right) + \frac{L+L_G}{2} \cdot \frac{(\alpha^n)^2}{I}\, \phi_{ub}^2 \left\| \Delta \widetilde{\mathbf{x}}_{\boldsymbol{\phi}}^n \right\|^2
\end{aligned}
$$

$$
\begin{aligned}
&- \frac{\alpha^n}{I} \sum_{i=1}^{I} \phi_i^n \left(\tau_i \| \Delta \widetilde{\mathbf{x}}_{(i),\boldsymbol{\phi}}^n \|^2 + G^+(\widetilde{\mathbf{x}}_{(i)}^n) - G^+(\bar{\mathbf{x}}_{\boldsymbol{\phi}^n}) \right) \\
&+ \alpha^n c_L \phi_{ub} \left\| \mathbf{e}_x^n \right\| \left\| \Delta \widetilde{\mathbf{x}}_{\boldsymbol{\phi}}^n \right\| + \alpha^n \phi_{ub} \left\| \mathbf{e}_y^n \right\| \left\| \Delta \widetilde{\mathbf{x}}_{\boldsymbol{\phi}}^n \right\|, \qquad (69)
\end{aligned}
$$

where in (a) we used (68), Assumption E.1, and the bound (109) (along with some basic manipulations); in (b) we used $\bar{\mathbf{y}}_{\boldsymbol{\phi}^n} = \bar{\mathbf{g}}^n$ [cf. (61b)]; (c) follows from the L_i-Lipschitz continuity of ∇f_i, L_G-Lipschitz continuity of ∇G^-, and the uniformly $\tilde{L}_i$-Lipschitz continuity of $\nabla \widetilde{f}_i(\mathbf{x}; \bullet)$; and in (d) we used the inequality $\|\mathbf{x}\|_1 \leq \sqrt{n}\,\|\mathbf{x}\|$, and the definition of c_L [cf. (35)].

Invoking the convexity of G^+ and using (61a), we can write

$$
G^+\left(\bar{\mathbf{x}}_{\boldsymbol{\phi}^{n+1}}\right) \leq \left(1-\alpha^n\right) G^+\left(\bar{\mathbf{x}}_{\boldsymbol{\phi}^n}\right) + \frac{\alpha^n}{I} \sum_{i=1}^{I} \phi_i^n G^+(\widetilde{\mathbf{x}}_{(i)}^n),
$$

which combined with (69) yields

$$
\begin{aligned}
&U\left(\bar{\mathbf{x}}_{\boldsymbol{\phi}^{n+1}}\right) \\
&\leq U\left(\bar{\mathbf{x}}_{\boldsymbol{\phi}^n}\right) - \frac{\alpha^n}{I} \phi_{lb}\, c_\tau \left\| \Delta \widetilde{\mathbf{x}}_{\boldsymbol{\phi}}^n \right\|^2 + \frac{L+L_G}{2} \cdot \frac{(\alpha^n)^2}{I} \phi_{ub}^2 \left\| \Delta \widetilde{\mathbf{x}}_{\boldsymbol{\phi}}^n \right\|^2 \\
&\quad + \alpha^n c_L \phi_{ub} \left\| \mathbf{e}_x^n \right\|_2 \left\| \Delta \widetilde{\mathbf{x}}_{\boldsymbol{\phi}}^n \right\| + \alpha^n \phi_{ub} \left\| \mathbf{e}_y^n \right\| \left\| \Delta \widetilde{\mathbf{x}}_{\boldsymbol{\phi}}^n \right\| \\
&\leq U\left(\bar{\mathbf{x}}_{\boldsymbol{\phi}^n}\right) - \frac{\alpha^n}{I} \phi_{lb}\, c_\tau \left\| \Delta \widetilde{\mathbf{x}}_{\boldsymbol{\phi}}^n \right\|^2 + \frac{L+L_G}{2} \cdot \frac{(\alpha^n)^2}{I} \phi_{ub}^2 \left\| \Delta \widetilde{\mathbf{x}}_{\boldsymbol{\phi}}^n \right\|^2 \\
&\quad + \frac{\phi_{ub}}{2} \left(c_L \epsilon_x + \epsilon_y\right) \left(\alpha^n\right)^2 \left\| \Delta \widetilde{\mathbf{x}}_{\boldsymbol{\phi}}^n \right\|^2 + \frac{\phi_{ub}}{2} c_L\, \epsilon_x^{-1} \left\| \mathbf{e}_x^n \right\|^2 + \frac{\phi_{ub}}{2} \epsilon_y^{-1} \left\| \mathbf{e}_y^n \right\|^2,
\end{aligned}
$$

where the last inequality follows from the Young's inequality, with $\epsilon_x > 0$ and $\epsilon_y > 0$. Applying the above inequality recursively for $\bar{B}$ steps, with $\bar{B}$ defined in Lemma 3, yields

$$
\begin{aligned}
U\left(\bar{\mathbf{x}}_{\boldsymbol{\phi}^{n+\bar{B}}}\right) \leq U\left(\bar{\mathbf{x}}_{\boldsymbol{\phi}^n}\right) &- \frac{c_\tau\, \phi_{lb}}{I} \sum_{t=0}^{\bar{B}-1} \alpha^{n+t} \left\| \Delta \widetilde{\mathbf{x}}_{\boldsymbol{\phi}}^{n+t} \right\|^2 \\
&+ \frac{\phi_{ub}}{2} \left(\frac{L+L_G}{I} \cdot \phi_{ub} + c_L\, \epsilon_x + \epsilon_y \right) E_{\Delta \widetilde{\mathbf{x}}}^n \\
&+ \frac{\phi_{ub}}{2} \left(c_L \epsilon_x^{-1} E_{\mathbf{x}_\perp}^n + \epsilon_y^{-1} E_{\mathbf{y}_\perp}^n \right). \qquad (70)
\end{aligned}
$$

Summing up (70) over $\bar{B}$ consecutive iterations leads to the desired result. □

Since, for sufficiently small α^n, the negative term on the RHS of (67) dominates the positive third term, to prove convergence of $\{U(\bar{\mathbf{x}}_{\boldsymbol{\phi}^{n+\bar{B}+k}})\}_{n\in\mathbb{N}_+}$, descent-based techniques used in the literature of distributed gradient-based algorithms would call for the summability of the consensus error $\{E^n_{\mathbf{x}_\perp}\}_{n\in\mathbb{N}_+}$ and tracking error $\{E^n_{\mathbf{y}_\perp}\}_{n\in\mathbb{N}_+}$ sequences. However, under constant step-size or unbounded (sub-)gradient of U, it seems not possible to infer such a result by just studying the dynamics of $\{E^n_{\mathbf{x}_\perp}\}_{n\in\mathbb{N}_+}$ and $\{E^n_{\mathbf{y}_\perp}\}_{n\in\mathbb{N}_+}$ *independently* from the optimization error $\Delta\widetilde{\mathbf{x}}^n_{\boldsymbol{\phi}}$. Therefore, exploring the interplay between these quantities, we put forth a new analysis, based on the following steps:

- *Step 1* We first bound $E^n_{\mathbf{x}_\perp}$ and $E^n_{\mathbf{y}_\perp}$ [specifically, term iv in (67)] as a function of $E^n_{\Delta\widetilde{\mathbf{x}}}$ (and thus $\Delta\widetilde{\mathbf{x}}^n_{\boldsymbol{\phi}}$)—see Proposition 12 [cf. Sect. 6.3.1]. Using Proposition 12, we then prove that $\{E^n_{\mathbf{x}_\perp}\}_{n\in\mathbb{N}_+}$ and $\{E^n_{\mathbf{y}_\perp}\}_{n\in\mathbb{N}_+}$ are summable, if $\{E^n_{\Delta\widetilde{\mathbf{x}}}\}_{n\in\mathbb{N}_+}$ is so−see Proposition 14 [cf. Sect. 6.3.2].
- *Step 2* Using Propositions 12 and 14, we build a new Lyapunov function [cf. Sect. 6.4], whose convergence implies the summability of $\{E^n_{\Delta\widetilde{\mathbf{x}}}\}_{n\in\mathbb{N}_+}$ and thus convergence of all error sequences [cf. Sect. 6.5], as stated in Theorem 4.

6.3 Interplay among $E^n_{\mathbf{x}_\perp}$, $E^n_{\mathbf{y}_\perp}$ and $E^n_{\Delta\widetilde{\mathbf{x}}}$

6.3.1 Bounding $E^n_{\mathbf{x}_\perp}$ and $E^n_{\mathbf{y}_\perp}$

We first study the dynamics of $\|\mathbf{e}^n_x\|$ and $\|\mathbf{e}^n_y\|$.

Lemma 11 *The disagreements $\|\mathbf{e}^n_x\|$ and $\|\mathbf{e}^n_y\|$ satisfy*

$$\|\mathbf{e}_x^{n+\bar{B}}\| \leq \rho_{\bar{B}}\left\|\mathbf{e}^n_x\right\| + c\sum_{t=0}^{\bar{B}-1}\alpha^{n+t}\left\|\Delta\mathbf{x}^{n+t}\right\|, \tag{71}$$

$$\|\mathbf{e}_y^{n+\bar{B}}\| \leq \rho_{\bar{B}}\left\|\mathbf{e}^n_y\right\| + c\,L_{\mathrm{mx}}\,\phi_{lb}^{-1}\sum_{t=0}^{\bar{B}-1}\left(2\left\|\mathbf{e}_x^{n+t}\right\| + \alpha^{n+t}\|\Delta\mathbf{x}^{n+t}\|\right), \tag{72}$$

where $c = I\sqrt{2I}$. Furthermore, if all $\mathbf{A}^n$ are double stochastic, then (71) and (72) hold with $\bar{B} = B$, $\rho_{\bar{B}} = \sqrt{1-\kappa/(2I^2)}$ and $c = 1$.

Proof See Appendix B. □

Using Lemma 11, we now study the dynamics of the weighted sum of the disagreements $\|\mathbf{e}^n_x\|$ and $\|\mathbf{e}^n_y\|$ over $\bar{B}$ consecutive iterations.

Proposition 12 *The sequences $\{\|\mathbf{e}^n_x\|^2\}_{n\in\mathbb{N}_+}$ and $\{\|\mathbf{e}^n_y\|^2\}_{n\in\mathbb{N}_+}$ satisfy*

$$\begin{aligned}\sum_{k=0}^{\bar{B}-1}\frac{k+1+(\bar{B}-k-1)\tilde{\rho}}{1-\tilde{\rho}}\left\|\mathbf{e}_x^{n+\bar{B}+k}\right\|^2 &\leq \sum_{k=0}^{\bar{B}-1}\frac{k+1+(\bar{B}-k-1)\tilde{\rho}}{1-\tilde{\rho}}\left\|\mathbf{e}_x^{n+k}\right\|^2\\ &\quad-\underbrace{\left(1-\left(\epsilon^{-1}+\bar{B}\right)\frac{2\bar{B}c^2}{1-\tilde{\rho}}\left(\alpha^n_{\mathrm{mx}}\right)^2\right)}_{\mu^n}\sum_{k=0}^{\bar{B}-1}E^{n+k}_{\mathbf{x}_\perp}+\underbrace{\left(\epsilon^{-1}+\bar{B}\right)\frac{2\bar{B}c^2}{1-\tilde{\rho}}}_{c_\Delta}\sum_{k=0}^{\bar{B}-1}E^{n+k}_{\Delta\widetilde{\mathbf{x}}},\end{aligned} \tag{73}$$

$$\sum_{k=0}^{\bar{B}-1} \frac{k+1+(\bar{B}-k-1)\tilde{\rho}}{1-\tilde{\rho}} \left\|\mathbf{e}_y^{n+\bar{B}+k}\right\|^2 \leq \sum_{k=0}^{\bar{B}-1} \frac{k+1+(\bar{B}-k-1)\tilde{\rho}}{1-\tilde{\rho}} \left\|\mathbf{e}_y^{n+k}\right\|^2$$
$$- \sum_{k=0}^{\bar{B}-1} E_{\mathbf{y}\perp}^{n+k} + \underbrace{\left(\epsilon^{-1}+\bar{B}\right) \frac{2\bar{B}c^2}{1-\tilde{\rho}} L_{\text{mx}}^2 \, \phi_{lb}^{-2}}_{c_\perp} \left(2+\alpha_{\text{mx}}^n\right)^2 \sum_{k=0}^{\bar{B}-1} E_{\mathbf{x}_\perp}^{n+k}$$
$$+ \underbrace{\left(\epsilon^{-1}+\bar{B}\right) \frac{2\bar{B}c^2}{1-\tilde{\rho}} L_{\text{mx}}^2 \, \phi_{lb}^{-2}}_{c_\perp} \sum_{k=0}^{\bar{B}-1} E_{\Delta\tilde{\mathbf{x}}}^{n+k}, \tag{74}$$

where $\alpha_{\text{mx}}^n \triangleq \max_{k=0,\ldots,2\bar{B}-2} \alpha^{n+k}$; $\tilde{\rho} \triangleq \rho_{\bar{B}}^2\left(1+\bar{B}\epsilon\right)$; *and* $\epsilon > 0$ *is any constant such that* $\tilde{\rho} < 1$.

Proof We prove only (73); (74) can be proved using similar steps. Squaring both sides of the inequality (71) leads to

$$\begin{aligned}
&\left\|\mathbf{e}_x^{n+\bar{B}}\right\|^2 \\
&\leq \rho_{\bar{B}}^2 \left\|\mathbf{e}_x^n\right\|^2 + \left(c \cdot \sum_{t=0}^{\bar{B}-1} \alpha^{n+t} \left\|\Delta\mathbf{x}^{n+t}\right\|\right)^2 + 2 \sum_{t=0}^{\bar{B}-1} c\, \rho_{\bar{B}}\, \alpha^{n+t} \left\|\mathbf{e}_x^n\right\| \left\|\Delta\mathbf{x}^{n+t}\right\| \\
&\overset{(a)}{\leq} \rho_{\bar{B}}^2 \left(1+\bar{B}\epsilon\right) \left\|\mathbf{e}_x^n\right\|^2 + \sum_{t=0}^{\bar{B}-1} \left(\frac{1}{\epsilon}+\bar{B}\right) c^2 \left(\alpha^{n+t}\right)^2 \left\|\Delta\mathbf{x}^{n+t}\right\|^2 \\
&\overset{(b)}{\leq} \tilde{\rho} \left\|\mathbf{e}_x^n\right\|^2 + \sum_{t=0}^{\bar{B}-1} \left(\frac{1}{\epsilon}+\bar{B}\right) 2\,c^2 \left(\alpha^{n+t}\right)^2 \left(\left\|\Delta\tilde{\mathbf{x}}_{\boldsymbol{\phi}}^{n+t}\right\|^2 + \left\|\mathbf{e}_x^{n+t}\right\|^2\right),
\end{aligned} \tag{75}$$

where (a) follows from the Young's inequality, with $\epsilon > 0$, and the Jensen's inequality; and in (b) we used (60). Note that, since $\rho_{\bar{B}} < 1$, $\tilde{\rho} = \rho_{\bar{B}}^2\left(1+\bar{B}\epsilon\right) < 1$, for all $\epsilon \in \left(0, \left(1-\rho_{\bar{B}}^2\right)/(\rho_{\bar{B}}^2 \bar{B})\right)$.

Denote $\tilde{\alpha}_{\text{mx}}^n \triangleq \max_{k=0,\ldots,\bar{B}-1} \alpha^{n+k}$. Multiplying (75) by $1/(1-\tilde{\rho})$ [resp. $\tilde{\rho}/(1-\tilde{\rho})$], adding $\|\mathbf{e}_x^n\|^2$ (resp. $\|\mathbf{e}_x^{n+\bar{B}}\|^2$) to both sides, and using the definitions of $E_{\Delta\tilde{\mathbf{x}}}^n$ and $E_{\mathbf{x}_\perp}^n$ [cf. (66)], yield

$$\begin{aligned}
&\frac{1}{1-\tilde{\rho}} \|\mathbf{e}_x^{n+\bar{B}}\|^2 + \|\mathbf{e}_x^n\|^2 \\
&\quad\leq \frac{\tilde{\rho}}{1-\tilde{\rho}} \left\|\mathbf{e}_x^n\right\|^2 + \left\|\mathbf{e}_x^n\right\|^2 + \frac{2\,c^2}{1-\tilde{\rho}} \left(\frac{1}{\epsilon}+\bar{B}\right) \left(E_{\Delta\tilde{\mathbf{x}}}^n + \left(\tilde{\alpha}_{\text{mx}}^n\right)^2 E_{\mathbf{x}_\perp}^n\right) \\
&\quad= \frac{1}{1-\tilde{\rho}} \left\|\mathbf{e}_x^n\right\|^2 + \frac{2\,c^2}{1-\tilde{\rho}} \left(\frac{1}{\epsilon}+\bar{B}\right) \left(E_{\Delta\tilde{\mathbf{x}}}^n + \left(\tilde{\alpha}_{\text{mx}}^n\right)^2 E_{\mathbf{x}_\perp}^n\right)
\end{aligned} \tag{76}$$

and

$$\begin{aligned}\frac{\tilde{\rho}}{1-\tilde{\rho}}\left\|\mathbf{e}_x^{n+\bar{B}}\right\|^2+\left\|\mathbf{e}_x^{n+\bar{B}}\right\|^2&=\frac{1}{1-\tilde{\rho}}\left\|\mathbf{e}_x^{n+\bar{B}}\right\|^2\\&\leq\frac{\tilde{\rho}}{1-\tilde{\rho}}\left\|\mathbf{e}_x^{n}\right\|^2+\frac{2\,c^2}{1-\tilde{\rho}}\left(\frac{1}{\epsilon}+\bar{B}\right)\left(E_{\Delta\tilde{\mathbf{x}}}^{n}+\left(\tilde{\alpha}_{\text{mx}}^{n}\right)^2E_{\mathbf{x}_\perp}^{n}\right),\end{aligned}\tag{77}$$

respectively.

We write now $\sum_{k=0}^{\bar{B}-1}E_{\mathbf{x}_\perp}^{n+k}$ as

$$\begin{aligned}\sum_{k=0}^{\bar{B}-1}E_{\mathbf{x}_\perp}^{n+k}=&\left(\left\|\mathbf{e}_x^{n+2\bar{B}-2}\right\|^2+2\left\|\mathbf{e}_x^{n+2\bar{B}-3}\right\|^2+\cdots+(\bar{B}-1)\left\|\mathbf{e}_x^{n+\bar{B}}\right\|^2\right)\\&+\left(\bar{B}\left\|\mathbf{e}_x^{n+\bar{B}-1}\right\|^2+(\bar{B}-1)\left\|\mathbf{e}_x^{n+\bar{B}-2}\right\|^2+\cdots+\left\|\mathbf{e}_x^{n}\right\|^2\right).\end{aligned}\tag{78}$$

Using (76) and (77) on the two terms in (78), we obtain the following bounds:

$$\begin{aligned}&\frac{\tilde{\rho}}{1-\tilde{\rho}}\left(\left\|\mathbf{e}_x^{n+2\bar{B}-2}\right\|^2+2\left\|\mathbf{e}_x^{n+2\bar{B}-3}\right\|^2+\cdots+(\bar{B}-1)\left\|\mathbf{e}_x^{n+\bar{B}}\right\|^2\right)\\&\quad+\left(\left\|\mathbf{e}_x^{n+2\bar{B}-2}\right\|^2+2\left\|\mathbf{e}_x^{n+2\bar{B}-3}\right\|^2+\cdots+(\bar{B}-1)\left\|\mathbf{e}_x^{n+\bar{B}}\right\|^2\right)\\&\leq\frac{\tilde{\rho}}{1-\tilde{\rho}}\left(\left\|\mathbf{e}_x^{n+\bar{B}-2}\right\|^2+2\left\|\mathbf{e}_x^{n+\bar{B}-3}\right\|^2+\cdots+(\bar{B}-1)\left\|\mathbf{e}_x^{n}\right\|^2\right)\\&\quad+\frac{2\,c^2}{1-\tilde{\rho}}\left(\frac{1}{\epsilon}+\bar{B}\right)\left[\left(E_{\Delta\tilde{\mathbf{x}}}^{n+\bar{B}-2}+\left(\tilde{\alpha}_{\text{mx}}^{n+\bar{B}-2}\right)^2E_{\mathbf{x}_\perp}^{n+\bar{B}-2}\right)+\cdots\right.\\&\quad\left.+(\bar{B}-1)\left(E_{\Delta\tilde{\mathbf{x}}}^{n}+\left(\tilde{\alpha}_{\text{mx}}^{n}\right)^2E_{\mathbf{x}_\perp}^{n}\right)\right],\end{aligned}\tag{79}$$

and

$$\begin{aligned}&\frac{1}{1-\tilde{\rho}}\left(\bar{B}\left\|\mathbf{e}_x^{n+2\bar{B}-1}\right\|^2+(\bar{B}-1)\left\|\mathbf{e}_x^{n+2\bar{B}-2}\right\|^2+\cdots+\left\|\mathbf{e}_x^{n+\bar{B}}\right\|^2\right)\\&\quad+\left(\bar{B}\left\|\mathbf{e}_x^{n+\bar{B}-1}\right\|^2+(\bar{B}-1)\left\|\mathbf{e}_x^{n+\bar{B}-2}\right\|^2+\cdots+\left\|\mathbf{e}_x^{n}\right\|^2\right)\\&\leq\frac{1}{1-\tilde{\rho}}\left(\bar{B}\left\|\mathbf{e}_x^{n+\bar{B}-1}\right\|^2+(\bar{B}-1)\left\|\mathbf{e}_x^{n+\bar{B}-2}\right\|^2+\cdots+\left\|\mathbf{e}_x^{n}\right\|^2\right)\\&\quad+\frac{2\,c^2}{1-\tilde{\rho}}\left(\frac{1}{\epsilon}+\bar{B}\right)\left[\bar{B}\left(E_{\Delta\tilde{\mathbf{x}}}^{n+\bar{B}-1}+\left(\tilde{\alpha}_{\text{mx}}^{n+\bar{B}-1}\right)^2E_{\mathbf{x}_\perp}^{n+\bar{B}-1}\right)+\cdots\right.\\&\quad\left.+\left(E_{\Delta\tilde{\mathbf{x}}}^{n}+\left(\tilde{\alpha}_{\text{mx}}^{n}\right)^2E_{\mathbf{x}_\perp}^{n}\right)\right].\end{aligned}\tag{80}$$

Summing (79) and (80) and rearranging terms while using (78), it is not difficult to check that

$$\sum_{k=0}^{\bar{B}-1} \frac{k+1+(\bar{B}-k-1)\tilde{\rho}}{1-\tilde{\rho}} \left\| \mathbf{e}_x^{n+\bar{B}+k} \right\|^2 + \sum_{k=0}^{\bar{B}-1} E_{\mathbf{x}_\perp}^{n+k}$$
$$\leq \sum_{k=0}^{\bar{B}-1} \frac{k+1+(\bar{B}-k-1)\tilde{\rho}}{1-\tilde{\rho}} \left\| \mathbf{e}_x^{n+k} \right\|^2 \tag{81}$$
$$+ \left(\frac{1}{\epsilon}+\bar{B}\right) \frac{2\bar{B}c^2}{1-\tilde{\rho}} \sum_{k=0}^{\bar{B}-1} E_{\Delta\tilde{\mathbf{x}}}^{n+k} + \left(\frac{1}{\epsilon}+\bar{B}\right) \frac{2\bar{B}c^2}{1-\tilde{\rho}} \left(\alpha_{\mathrm{mx}}^n\right)^2 \sum_{k=0}^{\bar{B}-1} E_{\mathbf{x}_\perp}^{n+k},$$

which leads to the desired result (73). □

We use now Proposition 12 in conjunction with Lemma 9 to prove the summability of $\{E_{\mathbf{x}_\perp}^n\}_{n\in\mathbb{N}_+}$ and $\{E_{\mathbf{y}_\perp}^n\}_{n\in\mathbb{N}_+}$, under that of $\{E_{\Delta\tilde{\mathbf{x}}}^n\}_{n\in\mathbb{N}_+}$. Let

$$\alpha_{\mathrm{mx}} \triangleq \sigma \cdot \sqrt{\frac{1-\tilde{\rho}}{2\bar{B}\left(\bar{B}+\epsilon^{-1}\right)c^2}} \tag{82}$$

with $\sigma \in (0, 1)$. This implies [recall the definition of μ^n in (73)]

$$\mu^n \geq \mu_{\min} \triangleq \left(1 - \left(\epsilon^{-1}+\bar{B}\right) \frac{2\bar{B}c^2}{1-\tilde{\rho}} \alpha_{\mathrm{mx}}^2\right) = 1-\sigma^2 > 0, \quad \forall \alpha_{\mathrm{mx}}^n \leq \alpha_{\mathrm{mx}}. \tag{83}$$

Proposition 13 *Suppose that i)* $\sum_{n=0}^{\infty}(\alpha^n)^2\|\Delta\tilde{\mathbf{x}}_{\boldsymbol{\phi}}^n\|^2 < \infty$*; and ii)* $\alpha^n \leq \alpha_{\mathrm{mx}}$*, for all but finite* $n \in \mathbb{N}_+$*. Then, the consensus and tracking disagreements satisfy* $\sum_{n=0}^{\infty}\|\mathbf{e}_x^n\|^2 < \infty$ *and* $\sum_{n=0}^{\infty}\|\mathbf{e}_y^n\|^2 < \infty$*, respectively.*

Proof It follows from (66) that it is sufficient to prove $\sum_{n=0}^{\infty} E_{\mathbf{x}_\perp}^n < \infty$ (for $\sum_{n=0}^{\infty}\|\mathbf{e}_x^n\|^2 < \infty$) and $\sum_{n=0}^{\infty} E_{\mathbf{y}_\perp}^n < \infty$ (for $\sum_{n=0}^{\infty}\|\mathbf{e}_y^n\|^2 < \infty$). We prove next only the former result.

By Assumption F and (83), there exists a sufficiently large n, say $\bar{n}$, such that $\mu^n \geq \mu_{\min} > 0$, for all $n \geq \bar{n}$. We assume, without loss of generality, that $\bar{n} = 0$. Applying Lemma 9 to (73) [cf. Proposition 12], we have $\sum_{n=0}^{\infty} E_{\Delta\tilde{\mathbf{x}}}^n < +\infty \implies \sum_{n=0}^{\infty} E_{\mathbf{x}_\perp}^n < +\infty$. It is then sufficient to prove that $\sum_{n=0}^{\infty}(\alpha^n)^2\|\Delta\tilde{\mathbf{x}}_{\boldsymbol{\phi}}^n\|^2 < \infty \implies \sum_{n=0}^{\infty} E_{\Delta\tilde{\mathbf{x}}}^n < +\infty$. This comes readily from the following chain of inequalities:

$$\sum_{k=0}^{n} E_{\Delta\tilde{\mathbf{x}}}^k = \sum_{k=0}^{n}\sum_{t=0}^{\bar{B}-1} \left(\alpha^{k+t}\right)^2 \left\|\Delta\tilde{\mathbf{x}}_{\boldsymbol{\phi}}^{k+t}\right\|^2 \leq \bar{B} \sum_{k=0}^{n+\bar{B}-1} \left(\alpha^k\right)^2 \left\|\Delta\tilde{\mathbf{x}}_{\boldsymbol{\phi}}^k\right\|^2.$$

□

6.3.2 Bounding term iv in (67)

We are now ready to bound term iv in (67), as stated next.

Proposition 14 *Suppose that $\alpha^n \leq \alpha_{\mathrm{mx}}$, then*

$$
\begin{aligned}
&\epsilon_y^{-1} \sum_{k=0}^{\bar{B}-1} \frac{k+1+(\bar{B}-k-1)\tilde{\rho}}{1-\tilde{\rho}} \left\| \mathbf{e}_y^{n+\bar{B}+k} \right\|^2 \\
&\quad + \frac{1}{\mu_{\min}} \left(c_L \epsilon_x^{-1} + \epsilon_y^{-1} c_\perp \left(2+\alpha_{\mathrm{mx}}\right)^2 \right) \sum_{k=0}^{\bar{B}-1} \frac{k+1+(\bar{B}-k-1)\tilde{\rho}}{1-\tilde{\rho}} \left\| \mathbf{e}_x^{n+\bar{B}+k} \right\|^2 \\
&\leq \epsilon_y^{-1} \sum_{k=0}^{\bar{B}-1} \frac{k+1+(\bar{B}-k-1)\tilde{\rho}}{1-\tilde{\rho}} \left\| \mathbf{e}_y^{n+k} \right\|^2 - \underbrace{\sum_{k=0}^{\bar{B}-1} \left(c_L \epsilon_x^{-1} E_{\mathbf{x}_\perp}^{n+k} + \epsilon_y^{-1} E_{\mathbf{y}_\perp}^{n+k} \right)}_{\text{term iv}} \\
&\quad + \frac{1}{\mu_{\min}} \left(c_L \epsilon_x^{-1} + \epsilon_y^{-1} c_\perp \left(2+\alpha_{\mathrm{mx}}\right)^2 \right) \sum_{k=0}^{\bar{B}-1} \frac{k+1+(\bar{B}-k-1)\tilde{\rho}}{1-\tilde{\rho}} \left\| \mathbf{e}_x^{n+k} \right\|^2 \\
&\quad + \left(\left(\epsilon_y^{-1} c_\perp \left(2+\alpha_{\mathrm{mx}}\right)^2 + c_L \epsilon_x^{-1} \right) \frac{c_\Delta}{\mu_{\min}} + \epsilon_y^{-1} c_\perp \right) \sum_{k=0}^{\bar{B}-1} E_{\Delta\tilde{\mathbf{x}}}^{n+k}
\end{aligned}
\tag{84}
$$

Proof Multiplying (74) by ϵ_y^{-1} on both sides we have

$$
\begin{aligned}
&\epsilon_y^{-1} \sum_{k=0}^{\bar{B}-1} \frac{k+1+(\bar{B}-k-1)\tilde{\rho}}{1-\tilde{\rho}} \left\| \mathbf{e}_y^{n+\bar{B}+k} \right\|^2 \\
&\leq \epsilon_y^{-1} \sum_{k=0}^{\bar{B}-1} \frac{k+1+(\bar{B}-k-1)\tilde{\rho}}{1-\tilde{\rho}} \left\| \mathbf{e}_y^{n+k} \right\|^2 \\
&\quad - \epsilon_y^{-1} \sum_{k=0}^{\bar{B}-1} E_{\mathbf{y}_\perp}^{n+k} + \epsilon_y^{-1} c_\perp \left(2+\alpha_{\mathrm{mx}}^n\right)^2 \sum_{k=0}^{\bar{B}-1} E_{\mathbf{x}_\perp}^{n+k} + \epsilon_y^{-1} c_\perp \sum_{k=0}^{\bar{B}-1} E_{\Delta\tilde{\mathbf{x}}}^{n+k} \\
&= \epsilon_y^{-1} \sum_{k=0}^{\bar{B}-1} \frac{k+1+(\bar{B}-k-1)\tilde{\rho}}{1-\tilde{\rho}} \left\| \mathbf{e}_y^{n+k} \right\|^2 \\
&\quad - \epsilon_y^{-1} \sum_{k=0}^{\bar{B}-1} E_{\mathbf{y}_\perp}^{n+k} - c_L \epsilon_x^{-1} \sum_{k=0}^{\bar{B}-1} E_{\mathbf{x}_\perp}^{n+k} \\
&\quad + \left(\epsilon_y^{-1} c_\perp \left(2+\alpha_{\mathrm{mx}}^n\right)^2 + c_L \epsilon_x^{-1} \right) \sum_{k=0}^{\bar{B}-1} E_{\mathbf{x}_\perp}^{n+k} + \epsilon_y^{-1} c_\perp \sum_{k=0}^{\bar{B}-1} E_{\Delta\tilde{\mathbf{x}}}^{n+k}.
\end{aligned}
\tag{85}
$$

Since $\alpha^n \leq \alpha_{\text{mx}}$, we have $\alpha^n_{\text{mx}} \leq \alpha_{\text{mx}}$ and $\mu^n \geq \mu_{\min}$. Equation (73) then implies

$$\begin{aligned}
&\sum_{k=0}^{\bar{B}-1} \frac{k+1+(\bar{B}-k-1)\tilde{\rho}}{1-\tilde{\rho}} \left\|\mathbf{e}_x^{n+\bar{B}+k}\right\|^2 \\
&\quad \leq \sum_{k=0}^{\bar{B}-1} \frac{k+1+(\bar{B}-k-1)\tilde{\rho}}{1-\tilde{\rho}} \left\|\mathbf{e}_x^{n+k}\right\|^2 - \mu_{\min} \sum_{k=0}^{\bar{B}-1} E_{\mathbf{x}_\perp}^{n+k} + c_\Delta \sum_{k=0}^{\bar{B}-1} E_{\Delta\tilde{\mathbf{x}}}^{n+k}.
\end{aligned}$$

Multiplying both sides of the above inequality by $(\epsilon_y^{-1} c_\perp\,(2+\alpha_{\text{mx}})^2 + c_L\epsilon_x^{-1})/\mu_{\min}$ and using the fact that $\alpha^n_{\text{mx}} \leq \alpha_{\text{mx}}$, we have

$$\begin{aligned}
&\frac{1}{\mu_{\min}}\left(c_L\epsilon_x^{-1} + \epsilon_y^{-1} c_\perp\,(2+\alpha_{\text{mx}})^2\right) \sum_{k=0}^{\bar{B}-1} \frac{k+1+(\bar{B}-k-1)\tilde{\rho}}{1-\tilde{\rho}} \left\|\mathbf{e}_x^{n+\bar{B}+k}\right\|^2 \\
&\quad \leq \frac{1}{\mu_{\min}}\left(c_L\epsilon_x^{-1} + \epsilon_y^{-1} c_\perp\,(2+\alpha_{\text{mx}})^2\right) \sum_{k=0}^{\bar{B}-1} \frac{k+1+(\bar{B}-k-1)\tilde{\rho}}{1-\tilde{\rho}} \left\|\mathbf{e}_x^{n+k}\right\|^2 \\
&\qquad - \left(c_L\epsilon_x^{-1} + \epsilon_y^{-1} c_\perp\,(2+\alpha_{\text{mx}})^2\right) \sum_{k=0}^{\bar{B}-1} E_{\mathbf{x}_\perp}^{n+k} \\
&\qquad + \left(c_L\epsilon_x^{-1} + \epsilon_y^{-1} c_\perp\,(2+\alpha_{\text{mx}})^2\right) \frac{c_\Delta}{\mu_{\min}} \sum_{k=0}^{\bar{B}-1} E_{\Delta\tilde{\mathbf{x}}}^{n+k}
\end{aligned} \tag{86}$$

Adding (86) to (85) leads to the desired result. □

6.4 Lyapunov-like function and its descent properties

We are now in the position to construct a function whose descent properties (every $\bar{B}$ iterations) will used to prove Theorem 4. Because of that, we will refer to such a function as Lyapunov-like function.

Adding (67) and (84) (multiplied by $\phi_{ub}/2$), yields

$$\begin{aligned}
&\sum_{k=0}^{\bar{B}-1} U\left(\bar{\mathbf{x}}_{\boldsymbol{\phi}^{n+\bar{B}+k}}\right) + \frac{\phi_{ub}}{2}\epsilon_y^{-1} \sum_{k=0}^{\bar{B}-1} \frac{k+1+(\bar{B}-k-1)\tilde{\rho}}{1-\tilde{\rho}} \left\|\mathbf{e}_y^{n+\bar{B}+k}\right\|^2 \\
&\quad + \frac{\phi_{ub}}{2\,\mu_{\min}}\left(c_L\epsilon_x^{-1} + \epsilon_y^{-1} c_\perp\,(2+\alpha_{\text{mx}})^2\right) \sum_{k=0}^{\bar{B}-1} \frac{k+1+(\bar{B}-k-1)\tilde{\rho}}{1-\tilde{\rho}} \left\|\mathbf{e}_x^{n+\bar{B}+k}\right\|^2 \\
&\leq \sum_{k=0}^{\bar{B}-1} U\left(\bar{\mathbf{x}}_{\boldsymbol{\phi}^{n+k}}\right) + \frac{\phi_{ub}}{2}\epsilon_y^{-1} \sum_{k=0}^{\bar{B}-1} \frac{k+1+(\bar{B}-k-1)\tilde{\rho}}{1-\tilde{\rho}} \left\|\mathbf{e}_y^{n+k}\right\|^2
\end{aligned}$$

$$+ \frac{\phi_{ub}}{2\,\mu_{\min}} \left(c_L \epsilon_x^{-1} + \epsilon_y^{-1} c_\perp \left(2 + \alpha_{\text{mx}}\right)^2 \right) \sum_{k=0}^{\bar{B}-1} \frac{k + 1 + (\bar{B} - k - 1)\tilde{\rho}}{1 - \tilde{\rho}} \left\| \mathbf{e}_x^{n+k} \right\|^2$$

$$- \frac{c_\tau}{I}\, \phi_{lb} \sum_{k=0}^{\bar{B}-1} \sum_{t=0}^{\bar{B}-1} \alpha^{n+k+t} \left\| \Delta \widetilde{\mathbf{x}}_{\boldsymbol{\phi}}^{n+k+t} \right\|^2$$

$$+ \frac{\phi_{ub}}{2} \left(\frac{L + L_G}{I} \cdot \phi_{ub} + c_L \epsilon_x + \epsilon_y + \epsilon_y^{-1} c_\perp \right) \sum_{k=0}^{\bar{B}-1} E_{\Delta \widetilde{\mathbf{x}}}^{n+k}$$

$$+ \frac{\phi_{ub}}{2\mu_{\min}} \left(c_L \epsilon_x^{-1} + \epsilon_y^{-1} c_\perp \left(2 + \alpha_{\text{mx}}\right)^2 \right) c_\Delta \sum_{k=0}^{\bar{B}-1} E_{\Delta \widetilde{\mathbf{x}}}^{n+k}. \tag{87}$$

Define

$$V^n \triangleq \sum_{k=0}^{\bar{B}-1} U\left(\bar{\mathbf{x}}_{\boldsymbol{\phi}^{n+k}}\right) + \frac{\phi_{ub}}{2} \epsilon_y^{-1} \sum_{k=0}^{\bar{B}-1} \frac{k + 1 + (\bar{B} - k - 1)\tilde{\rho}}{1 - \tilde{\rho}} \left\| \mathbf{e}_y^{n+k} \right\|^2$$
$$+ \frac{\phi_{ub}}{2\mu_{\min}} \left(c_L \epsilon_x^{-1} + \epsilon_y^{-1} c_\perp \left(2 + \alpha_{\text{mx}}\right)^2 \right) \sum_{k=0}^{\bar{B}-1} \frac{k + 1 + (\bar{B} - k - 1)\tilde{\rho}}{1 - \tilde{\rho}} \left\| \mathbf{e}_x^{n+k} \right\|^2, \tag{88}$$

and

$$\beta^n \triangleq \frac{c_\tau}{I} \phi_{lb} - \frac{\phi_{ub}}{2} \alpha^n \left(\frac{L + L_G}{I} \cdot \phi_{ub} + c_L \epsilon_x + \epsilon_y + \epsilon_y^{-1} c_\perp \right.$$
$$\left. + \frac{c_\Delta}{\mu_{\min}} \left(c_L \epsilon_x^{-1} + \epsilon_y^{-1} c_\perp \left(2 + \alpha_{\text{mx}}\right)^2 \right) \right). \tag{89}$$

Substituting (88) and (89) in (87), we obtain the desired descent property of V^n: for sufficiently large n, it holds

$$V^{n+\bar{B}} \leq V^n - \sum_{k=0}^{\bar{B}-1} \sum_{t=0}^{\bar{B}-1} \beta^{n+k+t} \alpha^{n+k+t} \left\| \Delta \widetilde{\mathbf{x}}_{\boldsymbol{\phi}}^{n+k+t} \right\|^2. \tag{90}$$

6.5 Proof of Theorem 4

The proof consists in two steps, namely:

- *Step 1* Leveraging the descent property of the Lyapunov-like function, we first show that $\lim_{n\to\infty} \|\Delta \widetilde{\mathbf{x}}_{\boldsymbol{\phi}}^n\| = 0$, either using a diminishing or constant step-size α^n (satisfying Assumption F); and
- *Step 2* Using the results in Step 1, we conclude the proof showing that i) $\lim_{n\to\infty} D(\mathbf{x}^n) = 0$ and ii) $\lim_{n\to\infty} J(\bar{\mathbf{x}}^n) = 0$

6.5.1 Step 1: $\lim_{n\to\infty}\|\Delta\widetilde{\mathbf{x}}_{\boldsymbol{\phi}}^n\| = 0$

Let us distinguish the two choices of step-size, namely: α^n is constant (satisfying Assumption F.1); or α^n is diminishing (satisfying Assumption F.2).

Case 1: constant step-size. Set $\alpha^n \equiv \alpha$ for all $n \in \mathbb{N}_+$. To obtain the desired descent on V^n [cf. (90)], α has to be chosen so that $\beta^n = \beta > 0$ [cf. (89)]. We show next that if α satisfies (36) [cf. Assumption F.2], then $\beta > 0$.

Recall that (90) holds under the assumption that $\alpha \le \alpha_{\text{mx}}$, with α_{mx} defined in (82). Substituting the expressions of α_{mx} and $\mu_{\min} = 1 - \sigma^2$ [cf. (83)] in (89) and using the definitions of c_Δ and $c_\perp$ [cf. Proposition 12], one can check that $\beta^n = \beta > 0$ [cf. (89)] if, in addition to $\alpha \le \alpha_{\text{mx}}$, α satisfies also

$$\alpha \le \frac{2\,c_\tau\,\phi_{lb}}{I\phi_{ub}}\left(\frac{L+L_G}{I}\cdot\phi_{ub} + c_L\epsilon_x + \epsilon_y + \frac{c_\Delta}{1-\sigma^2}\left(c_L\epsilon_x^{-1} + 9\,c_\perp\epsilon_y^{-1}\right)\right)^{-1} \quad (91)$$

where $\epsilon_x, \epsilon_y > 0$ are free parameters. The above upperbound is maximized by

$$\epsilon_x = \sqrt{\frac{c_\Delta}{1-\sigma^2}} = c\sqrt{\frac{2\bar{B}\left(\epsilon^{-1}+\bar{B}\right)}{(1-\tilde{\rho})(1-\sigma^2)}}$$
$$\epsilon_y = \sqrt{\frac{9\,c_\perp c_\Delta}{1-\sigma^2}} = 6\,L_{\text{mx}}\,\phi_{lb}^{-1}\,(\epsilon^{-1}+\bar{B})\,\frac{\bar{B}c^2}{1-\tilde{\rho}}\sqrt{\frac{1}{1-\sigma^2}}$$

Combining $\alpha \le \alpha_{\text{mx}}$ and (91), we get the following bound for α:

$$\alpha \le \min\left\{\sigma\sqrt{\frac{1-\tilde{\rho}}{2c^2\bar{B}\left(\bar{B}+\epsilon^{-1}\right)}},\ \frac{2c_\tau\phi_{lb}}{I\phi_{ub}}\left(\frac{L+L_G}{I}\cdot\phi_{ub} + 2c\cdot c_L\sqrt{\frac{2\bar{B}\left(\epsilon^{-1}+\bar{B}\right)}{(1-\tilde{\rho})(1-\sigma^2)}}\right.\right.$$
$$\left.\left. +12\,L_{\text{mx}}\,\phi_{lb}^{-1}\,(\epsilon^{-1}+\bar{B})\frac{\bar{B}c^2}{1-\tilde{\rho}}\sqrt{\frac{1}{1-\sigma^2}}\right)^{-1}\right\}, \quad (92)$$

where recall that $\epsilon < (1-\rho_{\bar{B}}^2)/(\rho_{\bar{B}}^2\bar{B})$ [cf. Proposition 12]. Since $(1-\tilde{\rho})/(\epsilon^{-1}+\bar{B})$ is maximized by $\epsilon = (1-\rho_{\bar{B}})/(\rho_{\bar{B}}\cdot\bar{B})$ with the corresponding value being $\left(1-\rho_{\bar{B}}\right)^2/\bar{B}$, we obtain from (92) the final bound (36).

Under (36), using (90) and Lemma 9 (recall that $\liminf_{n\to\infty} V^n > -\infty$, since U is bounded from below on $\mathcal{K}$) we get $\lim_{n\to\infty}\|\Delta\widetilde{\mathbf{x}}_{\boldsymbol{\phi}}^n\| = 0$ and, by Proposition 13, $\lim_{n\to\infty}\|\mathbf{e}_x^n\| = 0$ and $\lim_{n\to\infty}\|\mathbf{e}_y^n\| = 0$.

Case 2: diminishing step-size. Since α^n is diminishing, there exists a sufficiently large n_2 so that $\beta^n \ge \beta > 0$ for all $n \ge n_2$, implying

$$\sum_{n=0}^{\infty}\sum_{t=0}^{\bar{B}-1}\alpha^{n+t}\left\|\Delta\widetilde{\mathbf{x}}_{\boldsymbol{\phi}}^{n+t}\right\|^2 < \infty, \quad (93)$$

which together with $\sum_{n=0}^{\infty} \alpha^n = \infty$ and Proposition 13 yield

$$\liminf_{n\to\infty} \left\| \Delta\widetilde{\mathbf{x}}_{\boldsymbol{\phi}}^n \right\| = 0; \tag{94}$$

$$\lim_{n\to\infty} \left\| \mathbf{e}_x^n \right\| = 0 \quad \text{and} \quad \lim_{n\to\infty} \left\| \mathbf{e}_y^n \right\| = 0. \tag{95}$$

We prove next that $\limsup_{n\to\infty} \|\Delta\widetilde{\mathbf{x}}_{\boldsymbol{\phi}}^n\| = 0$, which together with (94) implies $\lim_{n\to\infty} \|\Delta\widetilde{\mathbf{x}}_{\boldsymbol{\phi}}^n\| = 0$. Suppose that $\limsup_{n\to\infty} \|\Delta\widetilde{\mathbf{x}}_{\boldsymbol{\phi}}^n\| > 0$. This, together with $\liminf_{n\to\infty} \|\Delta\widetilde{\mathbf{x}}_{\boldsymbol{\phi}}^n\| = 0$, implies that there exists an infinite set of indices $\mathcal{N}$ such that for all $n \in \mathcal{N}$, one can find an integer $i_n > n$ such that:

$$\left\| \Delta\widetilde{\mathbf{x}}_{\boldsymbol{\phi}}^n \right\| < \eta, \quad \|\Delta\widetilde{\mathbf{x}}_{\boldsymbol{\phi}}^{i_n}\| > 2\eta \tag{96}$$

$$\eta \leq \|\Delta\widetilde{\mathbf{x}}_{\boldsymbol{\phi}}^j\| \leq 2\eta, \quad n < j < i_n. \tag{97}$$

Denote $\widehat{\mathbf{x}}_{(i)}^n \triangleq \widehat{\mathbf{x}}_i(\mathbf{x}_{(i)}^n)$ and $\widehat{\mathbf{x}}^n \triangleq [\widehat{\mathbf{x}}_{(1)}^{n\top}, \ldots, \widehat{\mathbf{x}}_{(I)}^{n\top}]^\top$. We have:

$$\begin{aligned}
\eta &\leq \left\| \Delta\widetilde{\mathbf{x}}_{\boldsymbol{\phi}}^{i_n} \right\| - \left\| \Delta\widetilde{\mathbf{x}}_{\boldsymbol{\phi}}^n \right\| \leq \left\| \Delta\widetilde{\mathbf{x}}_{\boldsymbol{\phi}}^{i_n} - \Delta\widetilde{\mathbf{x}}_{\boldsymbol{\phi}}^n \right\| \\
&\leq \left\| \widetilde{\mathbf{x}}^{i_n} - \widetilde{\mathbf{x}}^n \right\| + \left\| \mathbf{J}_{\boldsymbol{\phi}^{i_n}} \mathbf{x}^{i_n} - \mathbf{J}_{\boldsymbol{\phi}^n} \mathbf{x}^n \right\| \\
&\leq \left\| \widehat{\mathbf{x}}^{i_n} - \widehat{\mathbf{x}}^n \right\| + \underbrace{\left\| \widetilde{\mathbf{x}}^{i_n} - \widehat{\mathbf{x}}^{i_n} \right\| + \left\| \widetilde{\mathbf{x}}^n - \widehat{\mathbf{x}}^n \right\|}_{e_1^n} + \left\| \mathbf{J}_{\boldsymbol{\phi}^{i_n}} \mathbf{x}^{i_n} - \mathbf{J}_{\boldsymbol{\phi}^n} \mathbf{x}^n \right\| \\
&\overset{(a)}{\leq} \hat{L} \left\| \mathbf{x}^{i_n} - \mathbf{x}^n \right\| + \left\| \mathbf{J}_{\boldsymbol{\phi}^{i_n}} \mathbf{x}^{i_n} - \mathbf{J}_{\boldsymbol{\phi}^n} \mathbf{x}^n \right\| + e_1^n \\
&\leq \hat{L} \left(\left\| \mathbf{x}^{i_n} - \mathbf{J}_{\boldsymbol{\phi}^{i_n}} \mathbf{x}^{i_n} \right\| + \left\| \mathbf{x}^n - \mathbf{J}_{\boldsymbol{\phi}^n} \mathbf{x}^n \right\| + \sqrt{I} \left\| \bar{\mathbf{x}}_{\boldsymbol{\phi}^{i_n}} - \bar{\mathbf{x}}_{\boldsymbol{\phi}^n} \right\| \right) \\
&\quad + \sqrt{I} \left\| \bar{\mathbf{x}}_{\boldsymbol{\phi}^{i_n}} - \bar{\mathbf{x}}_{\boldsymbol{\phi}^n} \right\| + e_1^n \\
&\leq \left(\hat{L} + 1 \right) \sqrt{I} \left\| \bar{\mathbf{x}}_{\boldsymbol{\phi}^{i_n}} - \bar{\mathbf{x}}_{\boldsymbol{\phi}^n} \right\| + \underbrace{\hat{L} \left(\left\| \mathbf{e}_x^{i_n} \right\| + \left\| \mathbf{e}_x^n \right\| \right)}_{e_2^n} + e_1^n \\
&\overset{(b)}{\leq} \left(\hat{L} + 1 \right) \sqrt{I} \sum_{t=n}^{i_n - 1} \alpha^t \left\| \frac{1}{I} \left((\boldsymbol{\phi}^t)^\top \otimes \mathbf{I}_m \right) \Delta\widetilde{\mathbf{x}}_{\boldsymbol{\phi}}^t \right\| + e_2^n + e_1^n \\
&\leq \left(\hat{L} + 1 \right) \sqrt{I} \sum_{t=n+1}^{i_n - 1} \alpha^t \left\| \Delta\widetilde{\mathbf{x}}_{\boldsymbol{\phi}}^t \right\| + \underbrace{(\hat{L} + 1)\sqrt{I}\alpha^n \left\| \Delta\widetilde{\mathbf{x}}_{\boldsymbol{\phi}}^n \right\|}_{e_3^n} + e_2^n + e_1^n \\
&\overset{(c)}{\leq} (\hat{L} + 1)\sqrt{I}\, \eta^{-1} \sum_{t=n+1}^{i_n - 1} \alpha^t \left\| \Delta\widetilde{\mathbf{x}}_{\boldsymbol{\phi}}^t \right\|^2 + e_3^n + e_2^n + e_1^n,
\end{aligned} \tag{98}$$

where in (a) we used (63) [cf. Lemma 7]; (b) follows from (61a); and in (c) we used the lower bound in (97).

Since (i) $\lim_{n\to\infty}\|\mathbf{e}_x^n\| = 0$ and $\lim_{n\to\infty}\|\mathbf{e}_y^n\| = 0$ [cf. (95)]; (ii) $\lim_{n\to\infty}\|\widetilde{\mathbf{x}}^n - \widehat{\mathbf{x}}^n\| = 0$ [cf. Lemma 8]; iii) and $\sum_{n=0}^{\infty}\sum_{t=0}^{\bar{B}-1}\alpha^{n+t}\|\Delta\widetilde{\mathbf{x}}_{\boldsymbol{\phi}}^{n+t}\|^2 < \infty$ [cf. (93)], there exists a sufficiently large n_3 such that the right-hand-side of (98) is less than η, for all $n > n_3$, which leads to a contradiction. Therefore, $\limsup_{n\to\infty}\|\Delta\widetilde{\mathbf{x}}_{\boldsymbol{\phi}}^n\| = 0$.

6.5.2 Step 2: $\lim_{n\to\infty} M(\mathbf{x}^n) = 0$

Recall that in the previous subsection we proved that i)$\lim_{n\to\infty}\|\Delta\widetilde{\mathbf{x}}_{\boldsymbol{\phi}}^n\| = 0$; ii) $\lim_{n\to\infty}\|\mathbf{e}_x^n\| = 0$; and iii) $\lim_{n\to\infty}\|\mathbf{e}_y^n\| = 0$, using either a constant step-size $\alpha^n \equiv \alpha$, with α satisfying (36), or a diminishing one. The statement $\lim_{n\to\infty} D(\mathbf{x}^n) = 0$ follows readily from point ii) and

$$\begin{aligned}\lim_{n\to\infty}\|\mathbf{x}_{(i)}^n - \bar{\mathbf{x}}^n\| &\le \lim_{n\to\infty}\|\mathbf{x}_{(i)}^n - \bar{\mathbf{x}}_{\boldsymbol{\phi}^n}\| + \lim_{n\to\infty}\left\|\bar{\mathbf{x}}_{\boldsymbol{\phi}^n} - \bar{\mathbf{x}}^n\right\| \\ &\le \lim_{n\to\infty}\|\mathbf{x}_{(i)}^n - \bar{\mathbf{x}}_{\boldsymbol{\phi}^n}\| + \lim_{n\to\infty}\frac{1}{I}\sum_{j=1}^{I}\|\mathbf{x}_{(j)}^n - \bar{\mathbf{x}}_{\boldsymbol{\phi}^n}\| = 0.\end{aligned} \tag{99}$$

Next we show $\lim_{n\to\infty} J(\bar{\mathbf{x}}^n) = 0$. Recall the definition $J(\bar{\mathbf{x}}^n) \triangleq \|\bar{\mathbf{x}}(\bar{\mathbf{x}}^n) - \bar{\mathbf{x}}^n\|$, where for notation simplicity, we set

$$\bar{\mathbf{x}}(\bar{\mathbf{x}}^n) \triangleq \underset{\mathbf{z}\in\mathcal{K}}{\operatorname{argmin}}\ \left\{\left(\nabla F(\bar{\mathbf{x}}^n) - \nabla G^-\left(\bar{\mathbf{x}}^n\right)\right)^\top(\mathbf{z} - \bar{\mathbf{x}}^n) + \frac{1}{2}\|\mathbf{z} - \bar{\mathbf{x}}^n\|^2 + G(\mathbf{z})^+\right\}. \tag{100}$$

Since

$$J(\bar{\mathbf{x}}^n) \le \|\widehat{\mathbf{x}}_i(\bar{\mathbf{x}}^n) - \bar{\mathbf{x}}^n\| + \|\bar{\mathbf{x}}(\bar{\mathbf{x}}^n) - \widehat{\mathbf{x}}_i(\bar{\mathbf{x}}^n)\|, \tag{101}$$

it is sufficient to show that the two terms on the right hand side are asymptotically vanishing, which is proved below.

- $\lim_{n\to\infty}\|\widehat{\mathbf{x}}_i(\bar{\mathbf{x}}^n) - \bar{\mathbf{x}}^n\| = 0$. We bound $\|\widehat{\mathbf{x}}_i(\bar{\mathbf{x}}^n) - \bar{\mathbf{x}}^n\|$ as

$$\begin{aligned}\|\widehat{\mathbf{x}}_i(\bar{\mathbf{x}}^n) - \bar{\mathbf{x}}^n\| &\le \|\widehat{\mathbf{x}}_i(\bar{\mathbf{x}}_{\boldsymbol{\phi}^n}) - \bar{\mathbf{x}}_{\boldsymbol{\phi}^n}\| + \|\bar{\mathbf{x}}_{\boldsymbol{\phi}^n} - \bar{\mathbf{x}}^n\| + \|\widehat{\mathbf{x}}_i(\bar{\mathbf{x}}^n) - \widehat{\mathbf{x}}_i(\bar{\mathbf{x}}_{\boldsymbol{\phi}^n})\| \\ &\overset{(a)}{\le} \|\widehat{\mathbf{x}}_i(\bar{\mathbf{x}}_{\boldsymbol{\phi}^n}) - \bar{\mathbf{x}}_{\boldsymbol{\phi}^n}\| + (1+\hat{L})\|\bar{\mathbf{x}}_{\boldsymbol{\phi}^n} - \bar{\mathbf{x}}^n\|,\end{aligned} \tag{102}$$

  where (a) follows from Lemma 7. From (99) we know $\lim_{n\to\infty}\|\bar{\mathbf{x}}_{\boldsymbol{\phi}^n} - \bar{\mathbf{x}}^n\| = 0$. To show $\|\widehat{\mathbf{x}}_i(\bar{\mathbf{x}}_{\boldsymbol{\phi}^n}) - \bar{\mathbf{x}}_{\boldsymbol{\phi}^n}\|$ is asymptotically vanishing, we bound it as

$$\begin{aligned}\|\widehat{\mathbf{x}}_i(\bar{\mathbf{x}}_{\boldsymbol{\phi}^n}) - \bar{\mathbf{x}}_{\boldsymbol{\phi}^n}\| \le{}& \|\widehat{\mathbf{x}}_i(\bar{\mathbf{x}}_{\boldsymbol{\phi}^n}) - \widehat{\mathbf{x}}_i(\mathbf{x}_{(i)}^n)\| + \|\widehat{\mathbf{x}}_i(\mathbf{x}_{(i)}^n) - \widetilde{\mathbf{x}}_{(i)}^n\| \\ &+ \|\widetilde{\mathbf{x}}_{(i)}^n - \bar{\mathbf{x}}_{\boldsymbol{\phi}^n}\|.\end{aligned} \tag{103}$$

The result $\lim_{n\to\infty} \|\widehat{\mathbf{x}}_i(\bar{\mathbf{x}}_{\boldsymbol{\phi}^n}) - \bar{\mathbf{x}}_{\boldsymbol{\phi}^n}\| = 0$ follows from Lemma 7, Lemma 8 and points (i)–(iii).
From (102) and (103) we conclude

$$\lim_{n\to\infty} \|\widehat{\mathbf{x}}_i(\bar{\mathbf{x}}^n) - \bar{\mathbf{x}}^n\| = 0. \tag{104}$$

- We prove $\lim_{n\to\infty} \|\bar{\mathbf{x}}(\bar{\mathbf{x}}^n) - \widehat{\mathbf{x}}_i(\bar{\mathbf{x}}^n)\| = 0$. Using the first order optimality conditions of $\bar{\mathbf{x}}(\bar{\mathbf{x}}^n)$ and $\widehat{\mathbf{x}}_i(\bar{\mathbf{x}}^n)$, we can bound their difference as

$$\begin{aligned}\|\bar{\mathbf{x}}(\bar{\mathbf{x}}^n) - \widehat{\mathbf{x}}_i(\bar{\mathbf{x}}^n)\| &\leq \|\nabla \widetilde{f}_i\big(\widehat{\mathbf{x}}_i(\bar{\mathbf{x}}^n); \bar{\mathbf{x}}^n\big) - \nabla f_i(\bar{\mathbf{x}}^n) - \widehat{\mathbf{x}}_i(\bar{\mathbf{x}}^n) + \bar{\mathbf{x}}^n\| \\ &\leq \|\nabla \widetilde{f}_i\big(\widehat{\mathbf{x}}_i(\bar{\mathbf{x}}^n); \bar{\mathbf{x}}^n\big) - \nabla f_i(\widehat{\mathbf{x}}_i(\bar{\mathbf{x}}^n))\| \\ &\quad + \|\nabla f_i(\widehat{\mathbf{x}}_i(\bar{\mathbf{x}}^n)) - \nabla f_i(\bar{\mathbf{x}}^n)\| + \|\widehat{\mathbf{x}}_i(\bar{\mathbf{x}}^n) - \bar{\mathbf{x}}^n\| \\ &\leq (\tilde{L}_i + L_i + 1)\|\widehat{\mathbf{x}}_i(\bar{\mathbf{x}}^n) - \bar{\mathbf{x}}^n\|.\end{aligned} \tag{105}$$

Using (104) we have

$$\lim_{n\to\infty} \|\bar{\mathbf{x}}(\bar{\mathbf{x}}^n) - \widehat{\mathbf{x}}_i(\bar{\mathbf{x}}^n)\| = 0. \tag{106}$$

The proof is completed just combining (101), (104) and (106).

7 Numerical results

7.1 Sparse regression

In this section, we test the performance of SONATA on the sparse linear regression problem (1) [cf. Sect. 2.1]. We generated the data set as follows. The ground truth signal $\mathbf{x}^\star \in \mathbb{R}^{500}$ is built by first drawing randomly a vector from the normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$, then thresholding the smallest 80% of its elements to zero. The underlying linear model is $\mathbf{b}_i = \mathbf{A}_i\mathbf{x}^\star + \mathbf{n}_i$, where the observation matrix $\mathbf{A}_i \in \mathbb{R}^{20\times 500}$ is generated by first drawing i.i.d. elements from the distribution $\mathcal{N}(0, 1)$, and then normalizing the rows to unit norm; and $\mathbf{n}_i$ is the additive noise, with i.i.d. entries from $\mathcal{N}(0, 0.1)$. We simulated 100 Monte Carlo trials, generating in each trial new $\mathbf{A}_i$'s and $\mathbf{n}_i$'s. We considered a time-varying digraph, composed of $I = 30$ agents. In every time slot, a new digraph is generated according to the following procedure: each agent i has two out-neighbors, one of them belonging to a chain connecting all the agents and the other one picked uniformly at random. To promote sparsity we use the (nonconvex) log function $G(\mathbf{x}) = \lambda \cdot \sum_i \log(1 + \theta\,|x_i|)/\log(1 + \theta)$, where the parameter θ controls the tightness of the approximation of the ℓ_0 function. We set $\lambda = 0.1$ and $\theta = 2$. It is convenient to rewrite $G(\mathbf{x})$ in the DC form $G(\mathbf{x}) = G^+(\mathbf{x}) - G^-(\mathbf{x})$, with $G^+(\mathbf{x}) = \|\mathbf{x}\|_1 \cdot (\theta/\log(1 + \theta))$. It is not difficult to check that such G^+ and G^- satisfy Assumption A.3; see, e.g., [1].

We run SONATA considering two alternative choices of $\widetilde{f}_i$, namely:

- `SONATA-PL` (PL stands for *partial linearization*) Since $f_i = \|\mathbf{b}_i - \mathbf{A}_i \mathbf{x}\|^2$ is convex, one can keep f_i unaltered and set in (28) $\widetilde{f}_i(\mathbf{x}_{(i)}) = f_i(\mathbf{x}_{(i)}) + (\tau_{PL}/2) \cdot \|\mathbf{x}_{(i)} - \mathbf{x}_{(i)}^n\|^2$. We set $\tau_{PL} = 1.5$. The unique solution $\widetilde{\mathbf{x}}_{(i)}^n$ of the resulting subproblem (26) is computed using the FLEXA algorithm, with the following tuning (see [16] for details): the initial point is selected randomly; the proximal parameter in the subproblems solved by FLEXA is set to be 2; and the step-size of FLEXA is chosen according to the diminishing rule $\gamma^r = \gamma^{r-1}\left(1 - \mu\gamma^{r-1}\right)$, with $\gamma^0 = 0.5$ and $\mu = 0.01$, with r denoting the (inner) iteration index. We terminate FLEXA when $J_{(i)}^r \leq 10^{-8}$, with $J_{(i)}^r \triangleq \|\mathbf{x}_{(i)}^{n,r} - \mathcal{S}_{\eta\lambda}(\mathbf{x}_{(i)}^{n,r} - 2\mathbf{A}_i^\top(\mathbf{A}_i\mathbf{x}_{(i)}^{n,r} - \mathbf{b}_i) - \tau_{PL} \cdot (\mathbf{x}_{(i)}^{n,r} - \mathbf{x}_{(i)}^n) + \widetilde{\boldsymbol{\pi}}_i^n + \lambda\nabla G^-(\mathbf{x}_{(i)}^n))\|_\infty$, where $\mathbf{x}_{(i)}^{n,r}$ denotes the value of $\mathbf{x}_{(i)}$ at the n-th outer and the r-th inner iteration, and $\mathcal{S}_\beta(\mathbf{x}) \triangleq \text{sign}(\mathbf{x}) \cdot \max\{|\mathbf{x}| - \lambda\mathbf{1}, \mathbf{0}\}$ is the soft-thresholding operator (intended to be applied to $\mathbf{x}$ component-wise).
- `SONATA-L` (L stands for *linearization*) To obtain a closed form expression for $\widetilde{\mathbf{x}}_{(i)}^n$ in (28), one can choose $\widetilde{f}_i$ as linearization of f_i (plus the proximal term), that is, $\widetilde{f}_i(\mathbf{x}_{(i)}) = 2\mathbf{A}_i^\top(\mathbf{A}_i\mathbf{x}_{(i)}^n - \mathbf{b}_i + (\tau_L/2) \cdot \|\mathbf{x}_{(i)} - \mathbf{x}_{(i)}^n\|^2$. We set $\tau_L = 1.5$.

The solution $\widetilde{\mathbf{x}}_{(i)}^n$ of the resulting subproblem (28) has the following closed form expression $\widetilde{\mathbf{x}}_{(i)}^n = \mathcal{S}_{\eta\lambda/\tau_L}(\mathbf{x}_{(i)}^n - \frac{1}{\tau_L}(2\mathbf{A}_i^\top(\mathbf{A}_i\mathbf{x}_{(i)}^n - \mathbf{b}_i) + \widetilde{\boldsymbol{\pi}}_i^n - \lambda\nabla G^-(\mathbf{x}_{(i)}^n)))$.

As benchmark, we also simulated the subgradient-push algorithm [27] with diminishing step-size. Note that there is no proof of convergence for such a scheme, when applied to the nonconvex, nonsmooth problem (1). For all the algorithms, we use the same step-size rule: $\alpha^n = \alpha^{n-1}\left(1 - \mu\alpha^{n-1}\right)$, with $\alpha^0 = 0.5$ and $\mu = 0.01$. Also, for all algorithms, we set $\mathbf{x}_{(i)}^0 = \mathbf{0}$, for all i.

We monitor the progresses of the algorithms towards stationarity and consensus using respectively the following two functions: i) $J^n \triangleq \|\bar{\mathbf{x}}^n - \mathcal{S}_{\eta\lambda}(\bar{\mathbf{x}}^n - 2\sum_i \mathbf{A}_i^\top\left(\mathbf{A}_i\bar{\mathbf{x}}^n - \mathbf{b}_i) + \lambda\nabla G^-(\bar{\mathbf{x}}^n)\right)\|_\infty$; and ii) $D^n \triangleq \|\mathbf{x}^n - \mathbf{J}\mathbf{x}^n\|_\infty$.

It is not difficult to check that J^n is a valid distance of the average iterates $\mathbf{J}\mathbf{x}^n$ from stationarity: it is continuous and zero if and only if its argument is a stationary solution of (1). We also use the normalized mean squared error (NMSE), defined as $\text{NMSE}^n \triangleq \|\mathbf{x}^n - (\mathbf{1} \otimes \mathbf{I})\,\mathbf{x}^\star\|^2/(I \cdot \|\mathbf{x}^\star\|^2)$.

In Fig. 1, we plot $\log_{10} J^n$ and $\log_{10} D^n$ [subplot (a)] and the NMSE [subplot (b)] versus the number of agents' message exchanges, averaged over 100 Monte-Carlo trials (we applied the $\log_{10}$ transform to J^n and D^n so that their distribution is closer to the normal one). The figures show that both versions of SONATA are much faster than the distributed gradient algorithm. This seems mainly due to the gradient tracking mechanism put forth by the proposed scheme. Under the same tuning, SONATA-PL converges faster than SONATA-L. According to our intensive simulations (not reported here), SONATA-PL becomes up to one order of magnitude faster than SONATA-L when τ_{PL} is reduced whereas reducing τ_L slows down SONATA-L.

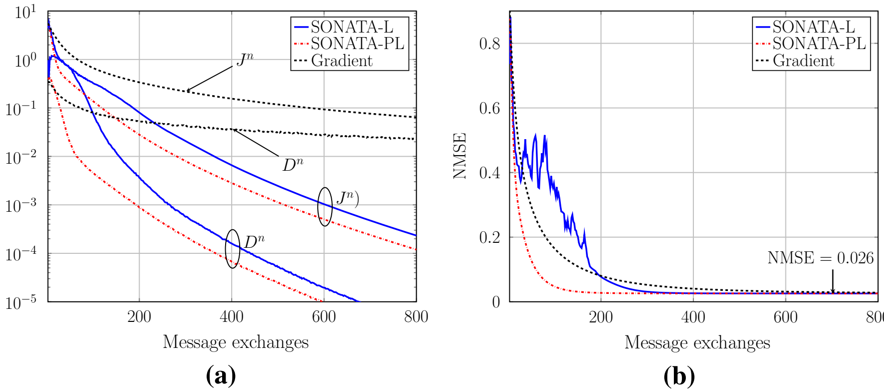


Fig. 1 Sparse regression problem (1) with log regularizer: SONATA-PL, SONATA-L, and subgradient-push; average of $\log_{10} J^n$ and $\log_{10} D^n$ versus agent's message exchange [subplot (**a**)]; average of NMSE versus agent's message exchange [subplot (**b**)]

7.2 Distributed PCA

Our second application is the distributed PCA problem

$$\min_{\|\mathbf{x}\|_2 \leq 1} \quad F(\mathbf{x}) \triangleq -\sum_{i=1}^{I} \|\mathbf{D}_i \mathbf{x}\|^2, \tag{107}$$

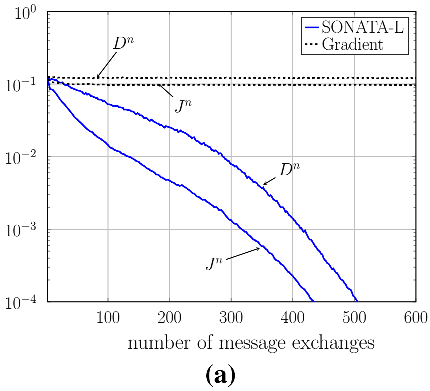
with $I = 30$.

Each agent i locally owns a data matrix $\mathbf{D}_i \in \mathbb{R}^{d_i \times m}$ and communicate via a time-varying digraph generated in the same way as the previous sparse regression example (cf. Sect. 7.1).

Since $f_i(\mathbf{x}) \triangleq -\|\mathbf{D}_i \mathbf{x}\|^2$ is concave, to apply SONATA we construct $\widetilde{f}_i$ by linearizing f_i, which leads to $\widetilde{F}_i(\mathbf{x}_{(i)}; \mathbf{x}^n_{(i)}) = (I \cdot \mathbf{y}^n_{(i)})^\top (\mathbf{x}_{(i)} - \mathbf{x}^n_{(i)}) + (\tau/2) \cdot \|\mathbf{x}_{(i)} - \mathbf{x}^n_{(i)}\|^2$. The solution $\widetilde{\mathbf{x}}^n_{(i)}$ of the resulting subproblem has the closed form solution $\widetilde{\mathbf{x}}^n_{(i)} = \mathcal{P}_{\|\mathbf{x}_{(i)}\| \leq 1}(\mathbf{x}^n_{(i)} - I \cdot \mathbf{y}^n_{(i)}/\tau)$, where $\mathcal{P}$ denotes the Euclidean projection onto the set $\{\mathbf{x}_{(i)} : \|\mathbf{x}_{(i)}\| \leq 1\}$. As benchmark, we implemented also the gradient projection algorithm [4], adapted to time-varying network. Note that there is no formal proof of this algorithm in the simulated setting. The performance of the algorithms is tested on both synthetic and real data sets, as detailed next.

7.2.1 Synthetic data

Each agent i locally owns a data matrix $\mathbf{D}_i \in \mathbb{R}^{30 \times 500}$, whose rows are i.i.d., drawn by the $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. The covariance matrix $\boldsymbol{\Sigma}$, whose eigendecomposition is $\boldsymbol{\Sigma} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$, is generated as follows: we synthesize $\mathbf{U}$ by first generating a square matrix whose entries follow the i.i.d. standard normal distribution, then perform the QR decomposition to obtain its orthonormal basis; and the eigenvalues diag$(\boldsymbol{\Lambda})$ are i.i.d. uniformly distributed in [0, 1].



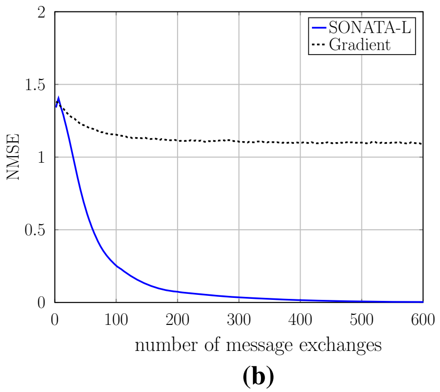


Fig. 2 Distributed PCA problem (1) on synthetic data set: SONATA and gradient projection algorithm; average of $\log_{10} J^n$ and $\log_{10} D^n$ versus agent's message exchange [subplot (**a**)]; average of NMSE versus agent's message exchange [subplot (**b**)]

The algorithms are tuned as follows: $\mathbf{x}_{(i)}^0$ is generated with i.i.d elements drawn by the standard Normal distribution. The step-size α^n is chosen according to the diminishing rule used in the previous example, where we set $\alpha^0 = 1$ and $\mu = 10^{-3}$ for SONATA and $\alpha^0 = 1$ and $\mu = 10^{-2}$ for the gradient algorithm. The proximal parameter τ for SONATA is set to be 1. The distance of $\bar{\mathbf{x}}^n$ from stationarity is measured by $J^n \triangleq \|\bar{\mathbf{x}}^n - \mathcal{P}_{\|\mathbf{x}_{(i)}\|\leq 1}(\bar{\mathbf{x}}^n - \nabla F(\bar{\mathbf{x}}^n))\|_\infty$,

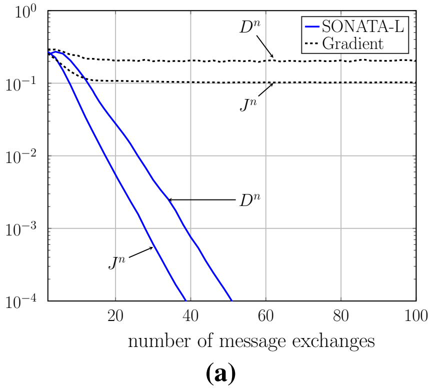
while the consensus disagreement D^n and the NMSE^n are defined as in the previous example; in the definition of NMSE^n the ground truth signal $\mathbf{x}^\star$ is now the leading eigenvector of matrix $\sum_{i=1}^I \mathbf{D}_i^\top \mathbf{D}_i$.

In Fig. 2, we plot $\log_{10} J^n$ and $\log_{10} D^n$ [subplot (a)] and the NMSE [subplot (b)] versus the number of agents' message exchanges, averaged over 100 Monte-Carlo trials. In each trial, $\boldsymbol{\Sigma}$ is fixed while the $\mathbf{D}_i$'s are randomly generated. Figure 2a clearly shows that SONATA can find a stationary point efficiently while the gradient algorithm progresses very slowly. More interestingly, Fig. 2b shows that SONATA always find the leading eigenvector whereas the gradient algorithm fails to achieve a small NMSE value.

7.2.2 Gene expression data

This second experiment tests SONATA on a real-world data set. Specifically, we used the breast cancer gene expression data set [5], which consists of $d = 158$ samples and $m = 12{,}625$ genes per sample. We first uniformly randomly permute the order the samples and then equally divided the samples among the $I = 30$ agents. To avoid the issue that d is not divisible by I, we let the first $I - 1$ agents owing $d_i = \lfloor d/I \rfloor$ samples each, while the I-th agent owning the remaining samples. The samples are preprocessed by subtracting the mean from all of them. Note that this can be achieved distributively by running an average consensus algorithm beforehand.

The rest of the setting and tuning of the algorithms are the same of those described in Sect. 7.2.1. In Fig. 3, we plot $\log_{10} J^n$ and $\log_{10} D^n$ [subplot (a)] and the NMSE



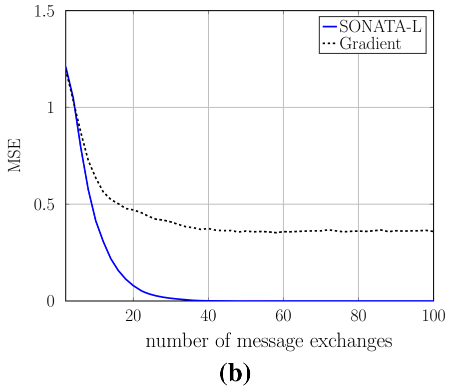


Fig. 3 Distributed PCA problem (1) on gene expression data set: SONATA and gradient projection algorithm; average of $\log_{10} J^n$ and $\log_{10} D^n$ versus agent's message exchange [subplot (**a**)]; average of NMSE versus agent's message exchange [subplot (**b**)]

[subplot (b)] versus the number of agents' message exchanges, averaged over 100 Monte-Carlo trials. In each trial, samples are randomly partitioned among the agents. From the figure we can see that the behavior of the algorithms on the gene expression data set is similar to that on synthetic data set. Moreover, SONATA converges quite fast even though the variable dimension of the real data set we adopted is massive.

8 Appendix

A Proof of Lemma 3

We begin introducing the following intermediate result.

Lemma 15 *In the setting of Lemma 3, the following hold:*

(i) *The elements of* $\mathbf{A}^{n:0}$, $n \in \mathbb{N}_+$, *can be bounded as*

$$\inf_{t \in \mathbb{N}_+} \left(\min_{1 \le i \le I} \left(\mathbf{A}^{t:0} \mathbf{1} \right)_i \right) \ge \phi_{lb}, \tag{108}$$

$$\sup_{t \in \mathbb{N}_+} \left(\max_{1 \le i \le I} \left(\mathbf{A}^{t:0} \mathbf{1} \right)_i \right) \le \phi_{ub}, \tag{109}$$

where ϕ_{lb} *and* ϕ_{ub} *are defined in* (8)*;*

(ii) *For any given* $n, k \in \mathbb{N}_+$, $n \ge k$, *there exists a stochastic vector* $\boldsymbol{\xi}^k \triangleq [\xi_1^k, \ldots \xi_I^k]^\top$ (*i.e.,* $\boldsymbol{\xi}^k > \mathbf{0}$ *and* $\mathbf{1}^\top \boldsymbol{\xi}^k = 1$) *such that*

$$\left| \mathbf{W}_{ij}^{n:k} - \xi_j^k \right| \le c_0 \, (\rho)^{\left\lfloor \frac{n-k+1}{(I-1)B} \right\rfloor}, \qquad \forall i, j \in [I], \tag{110}$$

where c_0 *and* ρ *are defined in* (10).

The proof Lemma 15 follows similar steps as those in [31, Lemma 2, Lemma 4] and thus is omitted, although the results in [31] are established under a stronger condition on $\mathcal{G}^n$ than Assumption B.

We prove now Lemma 3. Let $\mathbf{z} \in \mathbb{R}^{I\cdot m}$ be an arbitrary vector. For each $\ell = 1, \ldots, m$, define $\mathbf{z}_\ell \triangleq (\mathbf{I}_I \otimes \mathbf{e}_\ell^\top)\,\mathbf{z}$, where $\mathbf{e}_\ell$ is the ℓ-th canonical vector; we denote by $z_{\ell,j}$ the j-th component of $\mathbf{z}_\ell$, with $j \in [I]$. We have

$$\left\| \left(\widehat{\mathbf{W}}^{n:k} - \mathbf{J}_{\boldsymbol{\phi}^k} \right) \mathbf{z} \right\|_2 \leq \sqrt{I \cdot \sum_{\ell=1}^{m} \left\| \left(\mathbf{W}^{n:k} - \frac{1}{I}\mathbf{1}\,(\boldsymbol{\phi}^k)^\top \right) \mathbf{z}_\ell \right\|_\infty^2}. \tag{111}$$

We bound next the above term. Given $\boldsymbol{\xi}^k$ as in Lemma 15 [cf. (110)], define $\mathbf{E}^{n:k} \triangleq \mathbf{W}^{n:k} - \mathbf{1}(\boldsymbol{\xi}^k)^\top$, whose ij-th element is denoted by $E_{ij}^{n:k}$. We have

$$\begin{aligned}
&\left\| \left(\mathbf{W}^{n:k} - \frac{1}{I}\mathbf{1}(\boldsymbol{\phi}^k)^\top \right) \mathbf{z}_\ell \right\|_\infty \stackrel{(15)}{=} \left\| \left(\mathbf{W}^{n:k} - \frac{1}{I}\mathbf{1}\left(\boldsymbol{\phi}^{n+1}\right)^\top \mathbf{W}^{n:k} \right) \mathbf{z}_\ell \right\|_\infty \\
&\quad = \left\| \left(\mathbf{I} - \frac{1}{I}\mathbf{1}\left(\boldsymbol{\phi}^{n+1}\right)^\top \right) \mathbf{E}^{n:k}\, \mathbf{z}_\ell \right\|_\infty \\
&\quad \leq \max_{1\leq i \leq I} \left(\left(1 - \frac{\phi_i^{n+1}}{I} \right) \sum_{j=1}^{I} \left| E_{ij}^{n:k} \right| |z_{\ell,j}| + \sum_{j' \neq i}^{I} \frac{\phi_{j'}^{n+1}}{I} \sum_{j=1}^{I} \left| E_{j'j}^{n:k} \right| |z_{\ell,j}| \right) \\
&\quad \leq 2\, c_0\, (\rho)^{\left\lfloor \frac{n-k+1}{(I-1)B} \right\rfloor} \|\mathbf{z}_\ell\|_1 \leq 2\, c_0\, (\rho)^{\left\lfloor \frac{n-k+1}{(I-1)B} \right\rfloor} \sqrt{I}\ \|\mathbf{z}_\ell\|_2 .
\end{aligned} \tag{112}$$

Combining (111) and (112) we obtain

$$\left\| \widehat{\mathbf{W}}^{n:k} - \mathbf{J}_{\boldsymbol{\phi}^k} \right\|_2 \leq 2c_0 I\, (\rho)^{\left\lfloor \frac{n-k+1}{(I-1)B} \right\rfloor}. \tag{113}$$

Moreover, the matrix difference above can be alternatively uniformly bounded as follows:

$$\left\| \widehat{\mathbf{W}}^{n:k} - \mathbf{J}_{\boldsymbol{\phi}^k} \right\| = \|(\mathbf{I} - \mathbf{J}_{\boldsymbol{\phi}^{n+1}})\widehat{\mathbf{W}}^{n:k}\| \leq \|\mathbf{I} - \mathbf{J}_{\boldsymbol{\phi}^{n+1}}\| \|\widehat{\mathbf{W}}^{n:k}\| \stackrel{(a)}{\leq} \sqrt{2I} \cdot \sqrt{I},$$

where (a) follows from (25) and $\|\widehat{\mathbf{W}}^{n:k}\| \leq \sqrt{I}$. This completes the proof. $\square$

B Proof of Lemma 11

Recall the SONATA update written in vector–matrix form in (43)–(45). Note that the x-update therein is a special case of the perturbed condensed push-sum algorithm (16), with perturbation $\boldsymbol{\delta}^{n+1} = \alpha^n \widehat{\mathbf{W}}^n \mathbf{x}^n$. We can then apply Proposition 1 and readily obtain (71).

To prove (72), we follow a similar approach: noticing that the y-update in (45) is a special case of (16), with perturbation $\boldsymbol{\delta}^{n+1} = (\widehat{\mathbf{D}}_{\boldsymbol{\phi}^{n+1}})^{-1}\left(\mathbf{g}^{n+1} - \mathbf{g}^{n}\right)$, we can write

$$\begin{aligned}
\|\mathbf{e}_y^{n+\bar{B}}\| &\leq \rho_{\bar{B}}\|\mathbf{e}_y^n\| + \sqrt{2}I\sum_{t=0}^{\bar{B}-1}\|(\widehat{\mathbf{D}}_{\boldsymbol{\phi}^{n+t+1}})^{-1}\left(\mathbf{g}^{n+t+1} - \mathbf{g}^{n+t}\right)\| \\
&\leq \rho_{\bar{B}}\left\|\mathbf{e}_y^n\right\| + \sqrt{2}I\,L_{\mathrm{mx}}\phi_{lb}^{-1}\sum_{t=0}^{\bar{B}-1}\left\|\widehat{\mathbf{W}}^{n+t}(\mathbf{x}^{n+t} + \alpha^{n+t}\Delta\mathbf{x}^{n+t}) - \mathbf{x}^{n+t}\right\| \\
&\leq \rho_{\bar{B}}\left\|\mathbf{e}_y^n\right\| + \sqrt{2}I\,L_{\mathrm{mx}}\phi_{lb}^{-1}\sum_{t=0}^{\bar{B}-1}\left(\left\|\widehat{\mathbf{W}}^{n+t}\mathbf{e}_x^{n+t}\right\| + \left\|\mathbf{e}_x^{n+t}\right\| + \alpha^{n+t}\left\|\widehat{\mathbf{W}}^{n+t}\Delta\mathbf{x}^{n+t}\right\|\right) \\
&\leq \rho_{\bar{B}}\left\|\mathbf{e}_y^n\right\| + \sqrt{2}I\,L_{\mathrm{mx}}\phi_{lb}^{-1}\sum_{t=0}^{\bar{B}-1}\left((\sqrt{I}+1)\left\|\mathbf{e}_x^{n+t}\right\| + \alpha^{n+t}\sqrt{I}\left\|\Delta\mathbf{x}^{n+t}\right\|\right) \\
&\leq \rho_{\bar{B}}\left\|\mathbf{e}_y^n\right\| + I\sqrt{2I}\,L_{\mathrm{mx}}\phi_{lb}^{-1}\sum_{t=0}^{\bar{B}-1}\left(2\left\|\mathbf{e}_x^{n+t}\right\| + \alpha^{n+t}\left\|\Delta\mathbf{x}^{n+t}\right\|\right).
\end{aligned}$$

This completes the proof. □

References

1. Ahn, M., Pang, J., Xin, J.: Difference-of-convex learning: directional stationarity, optimality, and sparsity. SIAM J. Optim. **27**(3), 1637–1665 (2017). https://doi.org/10.1137/16M1084754
2. Bertsekas, D.P.: Nonlinear Programming, 2nd edn. Athena Scientific, Belmont (1999)
3. Bertsekas, D.P., Tsitsiklis, J.N.: Gradient convergence in gradient methods with errors. SIAM J. Optim. **10**(3), 627–642 (2000)
4. Bianchi, P., Jakubowicz, J.: Convergence of a multi-agent projected stochastic gradient algorithm for non-convex optimization. IEEE Trans. Autom. Control **58**(2), 391–405 (2013)
5. Bild, A.H., et al.: Oncogenic pathway signatures in human cancers as a guide to targeted therapies. Nature **439**(7074), 353 (2006)
6. Bottou, L., Curtis, F.E., Nocedal, J.: Optimization methods for large-scale machine learning. SIAM Rev. **60**(2), 223–311 (2018)
7. Bradley, P.S., Mangasarian, O.L.: Feature selection via concave minimization and support vector machines. In: Proceedings of the Fifteenth International Conference on Machine Learning (ICML 1998), vol. 98, pp. 82–90 (1998)
8. Cattivelli, F.S., Sayed, A.H.: Diffusion LMS strategies for distributed estimation. IEEE Trans. Signal Process. **58**(3), 1035–1048 (2010)
9. Chang, T.H.: A proximal dual consensus ADMM method for multi-agent constrained optimization. IEEE Trans. Signal Process. **64**(14), 3719–3734 (2014)
10. Chang, T.H., Hong, M., Wang, X.: Multi-agent distributed optimization via inexact consensus ADMM. IEEE Trans. Signal Process. **63**(2), 482–497 (2015)
11. Chen, J., Sayed, A.H.: Diffusion adaptation strategies for distributed optimization and learning over networks. IEEE Trans. Signal Process. **60**(8), 4289–4305 (2012)
12. Di Lorenzo, P., Scutari, G.: NEXT: in-network nonconvex optimization. IEEE Trans. Signal Inf. Process. Netw. **2**(2), 120–136 (2016)
13. Di Lorenzo, P., Scutari, G.: Distributed nonconvex optimization over networks. In: Proceedings of the IEEE 6th International Workshop on Computational Advances in Multi-sensor Adaptive Processing (CAMSAP 2015), Cancun, Mexico (2015)

14. Di Lorenzo, P., Scutari, G.: Distributed nonconvex optimization over time-varying networks. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 16), Shanghai (2016)
15. Facchinei, F., Lampariello, L., Scutari, G.: Feasible methods for nonconvex nonsmooth problems with applications in green communications. Math. Program. **164**(1–2), 55–90 (2017)
16. Facchinei, F., Scutari, G., Sagratella, S.: Parallel selective algorithms for nonconvex big data optimization. IEEE Trans. Signal Process. **63**(7), 1874–1889 (2015)
17. Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Stat. Assoc. **96**(456), 1348–1360 (2001)
18. Friedman, J., Hastie, T., Tibshirani, R.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Series in Statistics, vol. 1. Springer, New York (2009)
19. Fu, W.J.: Penalized regressions: the bridge versus the lasso. J. Comput. Graph. Stat. **7**(3), 397–416 (1998)
20. Gharesifard, B., Cortés, J.: When does a digraph admit a doubly stochastic adjacency matrix? In: Proceedings of the 2010 American Control Conference, pp. 2440–2445 (2010)
21. Hong, M., Hajinezhad, D., Zhao, M.: Prox-PDA: the proximal primal–dual algorithm for fast distributed nonconvex optimization and learning over networks. In: Proceedings of the 34th International Conference on Machine Learning (ICML 2017), vol. 70, pp. 1529–1538 (2017)
22. Jakovetic, D., Xavier, J., Moura, J.M.: Cooperative convex optimization in networked systems: augmented Lagrangian algorithms with directed gossip communication. IEEE Trans. Signal Process. **59**(8), 3889–3902 (2011)
23. Jakovetić, D., Xavier, J., Moura, J.M.: Fast distributed gradient methods. IEEE Trans. Autom. Control **59**(5), 1131–1146 (2014)
24. Kempe, D., Dobra, A., Gehrke, J.: Gossip-based computation of aggregate information. In: Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science, Cambridge, MA, USA, pp. 482–491 (2003)
25. Mokhtari, A., Shi, W., Ling, Q., Ribeiro, A.: DQM: decentralized quadratically approximated alternating direction method of multipliers. arXiv:1508.02073 (2015)
26. Mokhtari, A., Shi, W., Ling, Q., Ribeiro, A.: A decentralized second-order method with exact linear convergence rate for consensus optimization. IEEE Trans. Signal Inf. Process. Netw. **2**(4), 507–522 (2016)
27. Nedic, A., Olshevsky, A.: Distributed optimization over time-varying directed graphs. IEEE Trans. Autom. Control **60**(3), 601–615 (2015)
28. Nedić, A., Ozdaglar, A., Parrilo, P.A.: Constrained consensus and optimization in multi-agent networks. IEEE Trans. Autom. Control **55**(4), 922–938 (2010)
29. Nedich, A., Olshevsky, A., Ozdaglar, A., Tsitsiklis, J.N.: On distributed averaging algorithms and quantization effects. IEEE Trans. Autom. Control **54**(11), 2506–2517 (2009)
30. Nedich, A., Olshevsky, A., Shi, W.: Achieving geometric convergence for distributed optimization over time-varying graphs. SIAM J. Optim. **27**(4), 2597–2633 (2017)
31. Nedich, A., Ozdaglar, A.: Distributed subgradient methods for multi-agent optimization. IEEE Trans. Autom. Control **54**(1), 48–61 (2009)
32. Palomar, D.P., Chiang, M.: Alternative distributed algorithms for network utility maximization: framework and applications. IEEE Trans. Autom. Control **52**(12), 2254–2269 (2007)
33. Qu, G., Li, N.: Harnessing smoothness to accelerate distributed optimization. arXiv:1605.07112 (2016)
34. Rao, B.D., Kreutz-Delgado, K.: An affine scaling methodology for best basis selection. IEEE Trans. Signal Process. **47**(1), 187–200 (1999)
35. Sayed, A.H., et al.: Adaptation, learning, and optimization over networks. Found. Trends Mach. Learn. **7**(4–5), 311–801 (2014)
36. Scutari, G., Facchinei, F., Lampariello, L.: Parallel and distributed methods for constrained nonconvex optimization. Part I: theory. IEEE Trans. Signal Process. **65**(8), 1929–1944 (2017)
37. Scutari, G., Facchinei, F., Song, P., Palomar, D.P., Pang, J.S.: Decomposition by partial linearization: parallel optimization of multi-agent systems. IEEE Trans. Signal Process. **62**(3), 641–656 (2014)
38. Shi, W., Ling, Q., Wu, G., Yin, W.: EXTRA: an exact first-order algorithm for decentralized consensus optimization. SIAM J. Optim. **25**(2), 944–966 (2015)
39. Shi, W., Ling, Q., Wu, G., Yin, W.: A proximal gradient algorithm for decentralized composite optimization. IEEE Trans. Signal Process. **63**(22), 6013–6023 (2015)

40. Sun, Y., Daneshmand, A., Scutari, G.: Convergence rate of distributed convex and nonconvex optimization methods based on gradient tracking. Technical report, Purdue University (2018)
41. Sun, Y., Scutari, G.: Distributed nonconvex optimization for sparse representation. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4044–4048 (2017)
42. Sun, Y., Scutari, G., Palomar, D.: Distributed nonconvex multiagent optimization over time-varying networks. In: Proceedings of the Asilomar Conference on Signals, Systems, and Computers (2016). Appeared on arXiv on July 1, (2016)
43. Tatarenko, T., Touri, B.: Non-convex distributed optimization. arXiv:1512.00895 (2016)
44. Thi, H.L., Dinh, T.P., Le, H., Vo, X.: DC approximation approaches for sparse optimization. Eur. J. Oper. Res. **244**(1), 26–46 (2015)
45. Wai, H.T., Lafond, J., Scaglione, A., Moulines, E.: Decentralized Frank–Wolfe algorithm for convex and non-convex problems. arXiv:1612.01216 (2017)
46. Wei, E., Ozdaglar, A.: On the $o(1/k)$ convergence of asynchronous distributed alternating direction method of multipliers. In: Proceedings of the IEEE Global Conference on Signal and Information Processing (GlobalSIP 2013), Austin, TX, USA, pp. 551–554 (2013)
47. Weston, J., Elisseeff, A., Schölkopf, B., Tipping, M.: Use of the zero-norm with linear models and kernel methods. J. Mach. Learn. Res. **3**, 1439–1461 (2003)
48. Wright, S.J.: Coordinate descent algorithms. Math. Program. **151**(1), 3–34 (2015)
49. Xi, C., Khan, U.A.: On the linear convergence of distributed optimization over directed graphs. arXiv:1510.02149 (2015)
50. Xi, C., Khan, U.A.: ADD-OPT: accelerated distributed directed optimization. arXiv:1607.04757 (2016). Appeared on arXiv on July 16 (2016)
51. Xiao, L., Boyd, S., Lall, S.: A scheme for robust distributed sensor fusion based on average consensus. In: Proceedings of the 4th International Symposium on Information Processing in Sensor Networks, Los Angeles, CA, pp. 63–70 (2005)
52. Xu, J., Zhu, S., Soh, Y.C., Xie, L.: Augmented distributed gradient methods for multi-agent optimization under uncoordinated constant stepsizes. In: Proceedings of the 54th IEEE Conference on Decision and Control (CDC 2015), Osaka, Japan, pp. 2055–2060 (2015)
53. Zhang, S., Xin, J.: Minimization of transformed L_1 penalty: theory, difference of convex function algorithm, and robust application in compressed sensing. arXiv:1411.5735 (2014)
54. Zhu, M., Martínez, S.: An approximate dual subgradient algorithm for multi-agent non-convex optimization. IEEE Trans. Autom. Control **58**(6), 1534–1539 (2013)