Distributed Optimization Based on Gradient-tracking Revisited: Enhancing Convergence Rate via Surrogation

Ying Sun , Amir Daneshmand, and Gesualdo Scutari

ABSTRACT
We study distributed multiagent optimization over (directed, time-varying) graphs. We consider the minimization of $F + G$ subject to convex constraints, where F is the smooth strongly convex sum of the agent's losses and G is a nonsmooth convex function. We build on the SONATA algorithm: the algorithm employs the use of surrogate objective functions in the agents' subproblems (going thus beyond linearization, such as proximal-gradient) coupled with a perturbed (push-sum) consensus mechanism that aims to track locally the gradient of F. SONATA achieves precision $\epsilon > 0$ on the objective value in $\mathcal{O}(\kappa_g \log(1/\epsilon))$ gradient computations at each node and $\tilde{\mathcal{O}}\big(\kappa_g(1-\rho)^{-1/2}\log(1/\epsilon)\big)$ communication steps, where κ_g is the condition number of F and ρ characterizes the connectivity of the network. This is the first linear rate result for distributed composite optimization; it also improves on existing (non-accelerated) schemes just minimizing F, whose rate depends on much larger quantities than κ_g (e.g., the worst-case condition number among the agents). When considering in particular empirical risk minimization problems with statistically similar data across the agents, SONATA employing high-order surrogates achieves precision $\epsilon > 0$ in $\mathcal{O}\big((\beta/\mu)\log(1/\epsilon)\big)$ iterations and $\tilde{\mathcal{O}}\big((\beta/\mu)(1-\rho)^{-1/2}\log(1/\epsilon)\big)$ communication steps, where β measures the degree of similarity of the agents' losses and μ is the strong convexity constant of F. Therefore, when $\beta/\mu < \kappa_g$, the use of high-order surrogates yields provably faster rates than what achievable by first-order models; this is without exchanging any Hessian matrix over the network.

KEYWORDS
Distributed optimization, gradient tracking, linear rate, machine learning, statistical similarity, surrogate functions.

Submitted on May 2019; Revised on Oct. 2020

1. Introduction

We study distributed optimization over networks in the form:

$$\begin{aligned} \min_{\mathbf{x}} \quad & U(\mathbf{x}) \triangleq \underbrace{\frac{1}{m}\sum_{i=1}^{m} f_i(\mathbf{x})}_{F(\mathbf{x})} + G(\mathbf{x}) \\ \text{s.t.} \quad & \mathbf{x} \in \mathcal{K}, \end{aligned} \tag{P}$$

where $f_i : \mathbb{R}^d \to \mathbb{R}$ is the loss function of agent i, assumed to be smooth and convex while F is strongly convex on $\mathcal{K}$; $G : \mathbb{R}^d \to \mathbb{R}$ is a nonsmooth convex function on $\mathcal{K}$; and $\mathcal{K} \subseteq \mathbb{R}^d$ represents the set of common convex constraints. Each f_i is known to

The authors are with the School of Industrial Engineering, Purdue University, West-Lafayette, IN, USA. Emails: <sun578,adaneshm,gscutari>@purdue.edu.

Funding: This work has been supported by the USA National Science Foundation under Grants CIF 1719205, CIF 1564044, CMMI 1832688; and the Army Research Office under Grant W911NF1810238..

the associated agent only. Agents are connected through a communication network, modeled as a graph, possibly directed and/or time-varying. The goal is to cooperatively solve (P) by exchanging information only with their immediate neighbors.

Distributed 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. An instance of particular interest to this work is the distributed Empirical Risk Minimization (ERM) whereby the goal is to minimize the average loss over some dataset, distributed across the nodes of the network (cf. Sec. 2.1.2). Letting $\mathcal{D}^{(i)} = \{\mathbf{z}_1^{(i)}, \ldots, \mathbf{z}_n^{(i)}\}$ the dataset of n examples available at node i's side, the local empirical loss reads $f_i(\mathbf{x}) = 1/n \sum_{j=1}^{n} f(\mathbf{x}; \mathbf{z}_j^{(i)})$, where $f(\mathbf{x}; \mathbf{z}_j^{(i)})$ measures the fit between the parameter $\mathbf{x}$ and the sample $\mathbf{z}_j^{(i)}$. Data sets are usually large and high-dimensional, which makes routing local data to other agents (let alone to a centralized node) infeasible or highly inefficients. Given the cost of communications (especially if compared with the speed of local processing), the challenge in such a network setting is designing communication efficient distributed algorithms.

Motivated by the aforementioned applications, our focus pertains to such a design in two possible settings (one being a special case of the other) [2]: **1)** The scenario where no significant relationship can be assumed among the local functions f_i–this is what the literature of distributed optimization has extensively studied, and will be refereed to as the *unrelated* setting—and **2)** the case where the f_i's are *related*, e.g., because they reflect statistical similarity in the data residing at different nodes. For instance, in the distributed ERM problem above, when data are i.i.d. among machines, one can show that quantities such as the gradients and Hessian matrices of the local functions differ only by $\beta = \mathcal{O}(1/\sqrt{n})$, due to concentrations of measure effects [37, 57]–we will refer to this as *β-related* setting (cf. Sec. 2.1.2). If properly exploited in the algorithmic design, such similarity can speed up the optimization/learning process over general purpose optimization algorithms.

Centralized algorithms Problem (P) in the two settings above has been extensively studied in the centralized environment, including star-networks wherein there is a master node connected to all the other workers. Our interest is in the following (non-accelerated) algorithms:

1) Unrelated setting: (P) can be solved on star-networks employing the standard proximal gradient method: to reach precision $\epsilon > 0$ on the objective value, one needs $\mathcal{O}\big(\kappa_g \log(1/\epsilon)\big)$ iterations (which is also the number of communication rounds between the master and the workers), where κ_g is the condition number of F.

2) β-related setting: When the agents' functions f_i are sufficiently similar, a linear rate proportional to κ_g may be highly suboptimal. For instance, in the extreme case where all f_i's are identical ($\beta = 0$), the number of iterations/communications to an $\epsilon > 0$ solution would remain the same as for $\beta = \mathcal{O}(L)$. In fact, when $1 + \beta/\mu < \kappa_g$, faster rates can be obtained exploiting the similarity of the f_i's. Specifically, [37] proposed DANE: a mirror-descent type algorithm over star-networks, where each worker i replaces the quadratic term in its local proximal-gradient update with the Bregman divergence of the reference function $f_i + \beta/2\| \bullet \|^2$; and the master averages the solutions of the workers. DANE is applicable to (P) with $G = 0$: For *quadratic* losses, it achieves an ϵ-solution in $\mathcal{O}\big((\beta/\mu)^2 \cdot \log(1/\epsilon)\big)$ iterations/communications (it is assumed $\beta/\mu \geq 1$) while no improvement is proved over the proximal gradient if the f_i's are not quadratic. More recently, [7] proposed CEASE, which achieves DANE's rate for (P) with $G \neq 0$ and nonquadratic losses. Using recent results in [17], it is not difficult to check that the mirror-descent algorithm implemented at the master (thus without averaging workers' iterates) with the Bregman divergence of $f_1 + \beta/2\| \bullet \|^2$ (f_1 is the local function at the master) achieves an $\epsilon > 0$ solution in $\widetilde{\mathcal{O}}\big(\beta/\mu \cdot \log(1/\epsilon)\big)$ iterations/communications, improving thus on DANE/CEASE's rates.

A natural question is whether similar results–in particular the dependence of the rate on global optimization parameters as obtained on star-networks in the unrelated and β-related settings–are achievable over general network topologies, possibly time-varying and directed. The literature of distributed algorithms over general network topologies–albeit vast–do not provide a satisfactory answer, leaving a gap between rate results over star networks and what has been certified over general graphs–see Sec 1.2 for a review of the state of the art. In a nutshell, **(i)** there are no distributed schemes provably achieving linear rate for (P) with $G \neq 0$ and/or constraints (cf. Table 1). Furthermore, even considering the unconstrained minimization of F (i.e., $G = 0$ and $\mathcal{K} = \mathbb{R}^d$), **(ii)** linear convergence is certified at a rate depending on much larger quantities than the global condition number κ_g–see Table 1; and **(iii)** when $1 + \beta/\mu < \kappa_g$ (β-related setting), no rate improvement is provably achieved by existing distributed algorithms. These are much more pessimistic rate dependencies than what achieved over star-topologies. The goal of this paper is to close exactly this gap.

1.1. *Major contributions*

Our major results are summarized next.

(1) We provide the first linear convergence rate analysis of a distributed algorithm, SONATA (Successive cONvex Approximation algorithm over Time-varying digrAphs), applicable to the *composite, constrained* formulation (P) over (time-varying, directed) graphs. SONATA was earlier proposed in the companion paper [35] for nonconvex problems. It combines the use of surrogate functions in the agents' subproblems with a perturbed (push-sum) consensus mechanism that aims at locally tracking the gradient of F. Surrogate functions replace the more classical first order approximation of the local f_i's, which is the omnipresent choice in current distributed algorithms, offering the potential to better suit the geometry of the problem. For instance, (approximate) Newton-type subproblems or mirror descent-type updates naturally fit our surrogate models; they are the key enabler of provably faster rates in the β-related setting. We comment SONATA's rates below (cf. Table 3).

(2) **Unrelated setting (Table 3):** When the network is sufficiently connected or it has a star-topology, SONATA reaches an ϵ-solution on the objective value in $\mathcal{O}\big(\kappa_g \log(1/\epsilon)\big)$ iterations/communications, which matches the rate of the centralized proximal-gradient algorithm. For arbitrary network connectivity, the same iteration complexity is achieved at the cost of $\mathcal{O}((1-\rho)^{-1/2})$ rounds of communications per iteration (employing Chebishev acceleration), where $\rho \in [0, 1)$ is the second largest eigenvalue modulus of the mixing matrix. Our rates improve on those of existing distributed algorithms which show a much more pessimistic dependence on the optimization parameters and are proved under more restrictive assumptions–contrast Table 1 with Table 3. Linear rates over time-varying digraphs are reported in Table 4 (cf. Sec. 4.2).

(3) **β-related setting (Table 3):** When the agents' functions are sufficiently similar (specifically, $1 + \beta/\mu < \kappa_g$), the use of a mirror descent-type surrogate over

Algorithms		[9, 10, 14, 15, 18, 20, 29, 38, 40]	[27, 46, 48, 56]	[19, 21, 25, 31, 55]	**SONATA**
Problem:	F **(smooth)**	each f_i scvx	each f_i scvx	F scvx	F scvx
	G **(nonsmooth)**				✓
	constraints $\mathcal{K}$				✓
Network:	**time-varying**	only [18]		only [21, 31]	✓
	digraph		✓	only [21, 31]	✓

Table 1. Existing linearly convergent distributed algorithms. SONATA is the only scheme achieving linear rate in the presence of G in (P) or constraints. The explicit expression of the rates of the above nonaccelerated schemes (for which is available) is reported in Table 1.

linearization of the f_i's provably yields faster rates, at higher computation costs. This improves on the rate of existing distributed algorithms, which are oblivious of function similarity (cf. Table 1). Notice that this is achieved without exchanging any Hessian matrix over the network but leveraging function homogeneity via surrogation. When customized over star-topologies, SONATA's rates improve on DANE/CEASE's ones too.

Algorithm	**Problem**	**Linear rate:** $\mathcal{O}(\delta\ \log(1/\epsilon))$
EXTRA [38]	F	$\delta = \mathcal{O}(\frac{\kappa_\ell^2}{1-\rho}), \quad \kappa_\ell = \frac{L_{\text{mx}}}{\mu_{\text{mn}}}$
DIGing [21, 25]	F	$\delta = \frac{\hat{\kappa}^{1.5}}{(1-\rho)^2}, \quad \hat{\kappa} \triangleq \frac{L_{\text{mx}}}{(1/m)\sum_i \mu_i}$
Harnessing [29]	F	$\delta = \frac{\kappa_\ell^2}{(1-\rho)^2}$
NIDS [14], ABC [12]	F	$\delta = \max\{\kappa_\ell,\ \frac{1}{1-\rho}\}$
Exact Diffusion [55]	F	$\delta = \frac{\bar{\kappa}^2}{1-\rho}, \quad \bar{\kappa} \triangleq \frac{L_{\text{mx}}}{\mu_{\text{mx}}}$
Augmented Lagrangian [9]	F	$\delta = \frac{\kappa_\ell}{1-\rho}$
ADMM [40]	F	$\frac{\kappa_\ell^4}{1-\rho}$

Table 2. Linear rate of existing non-accelerated algorithms over undirected graphs: communications rounds to reach $\epsilon > 0$ accuracy; L_i and μ_i are the smoothness and strong convexity constants of f_i's, respectively; $L_{\text{mx}} \triangleq max_i L_i$, $\mu_{\text{mn}} \triangleq \min_i \mu_i$; and $\rho \in [0, 1)$ is the second largest eigenvalue modulus of the mixing matrix [cf. (26)]. The rates above include the quantities κ_l, $\hat{\kappa}$, and $\breve{\kappa}$ rather than the much desirable global condition number $\kappa_g \triangleq L/\mu$ (L and μ are the smoothness and strong convexity constants of F, respectively). Furthermore, they are independent on β, implying that faster rates are not certified when $1 + \beta/\mu < \kappa_g$ (β-related setting).

1.2. *Related works*

Early works on distributed optimization aimed at decentralizing the (sub)gradient algorithm. The Distributed Gradient Descent (DGD) was introduced in [23] for unconstrained instances of (P) and in [16] for least squares, bot over undirected graphs. A refined convergence rate analysis of DGD [23] can be found in [54]. Subsequent variants of DGD include the projected (sub)gradient algorithm [24] and the push-sum gradient consensus algorithm [22], the latter implementable over digraphs. While different, the updates of the agents' variables in the above algorithms can be abstracted as a combination of one (or multiple) consensus step(s) (weighted average with neighbors variables) and a local (sub)gradient descent step, controlled by a step-size (in some schemes, followed by a proximal operation). A diminishing step-size is used to reach *exact* consensus on the solution, converging thus at a *sublinear rate*. With a fixed step-size α, linear rate of the iterates is achievable, but it can only converge to a $\mathcal{O}(\alpha)$-neighborhood of the solution [23, 54].

Surrogate	**Communication Rounds**	**Extra Averaging**	ρ **(network)**	β
linearization	$\mathcal{O}\left(\kappa_g \log\left(1/\epsilon\right)\right)$	✗	$\rho = \mathcal{O}(\kappa_g^{-1}(1+\frac{\beta}{L})^{-2})$ or star-networks	arbitrary
	$\widetilde{\mathcal{O}}\left(\frac{\kappa_g}{\sqrt{1-\rho}}\log(1/\epsilon)\right)$	✓	arbitrary	arbitrary
local f_i	$\mathcal{O}\left(1 \cdot \log\left(1/\epsilon\right)\right)$	✗	$\rho = \mathcal{O}\left(\left(1+\frac{\beta}{\mu}\right)^{-2}\left(\kappa_g+\frac{\beta}{\mu}\right)^{-2}\right)$ or star-networks	$\beta \leq \mu$
	$\widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{1-\rho}}\log(1/\epsilon)\right)$	✓	arbitrary	
	$\mathcal{O}\left(\frac{\beta}{\mu} \cdot \log\left(1/\epsilon\right)\right)$	✗	$\rho = \mathcal{O}\left(\left(1+\frac{L}{\beta}\right)^{-1}\left(\kappa_g+\frac{\beta}{\mu}\right)^{-1}\right)$ or star-networks	$\beta > \mu$
	$\widetilde{\mathcal{O}}\left(\frac{\beta/\mu}{\sqrt{1-\rho_0}} \cdot \log(1/\epsilon)\right)$	✓	arbitrary	

Table 3. Summary of convergence rates of SONATA over undirected graphs: number of communication rounds to reach ϵ-accuracy. In the table, β is the homogeneity parameter measuring the similarity of the loss functions f_i's (cf. Definition 2.1); the other quantities are defined as in Table 1. The extra averaging steps are performed using Chebyshev acceleration [32, 45]. The $\widetilde{\mathcal{O}}$ notation hides log dependence on κ_g and β/μ (see Sec. 3.4.2 for the exact expressions). Rates over time-varying directed graphs are summarized in Table 4 (cf. Sec. 4.2).

Several subsequent attempts have been proposed to cope with this speed-accuracy dilemma, leading to algorithms converging to the *exact* solution while employing a *constant* step-size. Based upon the mechanism put forth to cancel the steady state error in the individual gradient direction, existing proposals can be roughly organized in three groups, namely: i) primal-based distributed methods leveraging the idea of gradient tracking [4, 5, 21, 27–29, 46–48, 50–52]; ii) distributed schemes using ad-hoc corrections of the local optimization direction [3, 38, 56]; and iii) primal-dual-based methods [9, 15, 20, 32, 40]. We elaborate next on these works, focusing on schemes achieving linear rate– Table 1 organizes these schemes based upon the setting their convergence is established while Table 1 reports the explicit expression of the rates.

i) Gradient-tracking-based methods: In these schemes, each agent updates its own variables along a direction that tracks the global gradient ∇F. This idea was proposed independently in the NEXT algorithm [4, 5] for Problem (P) and in AUG-DGM [52] for strongly convex, smooth, unconstrained optimization. The work [42] introduced SONATA, extending NEXT over (time-varying) digraphs. A convergence rate analysis of [52] was later developed in [21, 29, 53], with [21] considering also (time-varying) digraphs. Other algorithms based on the idea of gradient tracking and implementable over digraphs are ADD-OPT [47] and [46]. Subsequent schemes, [48], the Push-Pull [27], and the $\mathcal{AB}$ [31] algorithms, relaxed previous conditions on the mixing matrices used in the consensus and gradient tracking steps over digraphs, which neither need to be row- nor column-stochastic. All the schemes above but NEXT and SONATA are applicable only to *smooth, unconstrained* instances of (P), with *each* f_i *strongly convex.* This latter assumption is restrictive in some applications, such as distributed machine learning, where not all f_i are strongly convex but F is so.

ii) Ad-hoc gradient correction-based methods: These methods developed specific corrections of the plain DGD direction. Specifically, EXTRA [38] and its variant over digraphs, EXTRA-PUSH [56], introduce two different weight matrices for any two consecutive iterations as well as leverage history of gradient information. They are applicable only to it smooth, unconstrained problems; when each f_i is strongly convex, they generate iterates that converge linearly to the minimizer of F. To deal with an additive convex nonsmooth term in the objective, [39] proposed PG-EXTRA, which is thus applicable to (P) over undirected graphs, possibly with different local nonsmooth functions. However, linear convergence is not certified. A different approach is to use a linearly increasing number of consensus steps rather than correcting directly the gradient direction; this has been studied in [3] for unconstrained minimization of smooth, strongly convex f_i's over undirected graphs.

iii) Primal-dual methods: A common theme of these schemes is employing a prima-dual reformulation of the original multiagent problem whereby dual variables associated to a properly defined (augmented) Lagrangian function serve the purpose of correcting the plain DGD local direction. Examples of such algorithms include: i) distributed ADMM methods [11, 40] and their inexact implementations [15, 19]; ii) distributed Augmented Lagrangian-based methods with randomized primal variable updates [9]; and iii) a distributed dual ascent method employing tracking of the average of the primal variable [18]. All these schemes are applicable only to *smooth, unconstrained* optimization over undirected graphs, with [18] handling time-varying graphs. The extension of these methods to digraphs seems not straightforward, because it is not clear how to enforce consensus via constraints over directed networks.

To summarize, the above literature review shows that currently there exists no distributed algorithm for the general formulation (P) that provably converges at linear rate to the exact solution, in the presence of a nonsmooth function G or constraints (cf. Table 1); let alone mentioning digraphs. Furthermore, when it comes to the dependence of the rate on the optimization parameters, Table 3 shows that, even restricting to unconstrained, smooth minimization, SONATA's rates improve on existing ones–in particular, SONATA provably obtains fast convergence if the agents' objective functions (e.g., data) are sufficiently similar.

Concurrent works While our manuscript was under review and available on arXiv [41], a few other related technical reports appeared online [1, 13, 30], which we briefly discuss next. The authors in [1] studied a class of distributed proximal gradient-based methods to solve Problem (P) with $G \neq 0$, over undirected, static, graphs. The algorithms reach an ϵ-solution in $\mathcal{O}\big(\breve{\kappa}(1-\rho)^{-1}\log(1/\epsilon)\big)$ iterations/communications, where $\breve{\kappa} \triangleq L_{\text{mx}}/\mu$. The authors in [30] proposed an inexact distributed projected gradient descent method for the unconstraint minimization of F and proved a communication complexity of $\tilde{O}\big(\kappa_g\,(1-\rho)^{-1}\log^2(1/\epsilon)\big)$ ($\tilde{\mathcal{O}}$ hides a log-dependence on $L^2_{\max}/\mu^2$), which is determined by the global condition number κ_g; the algorithm runs over time-varying, undirected, graphs (as long as they are connected at each iteration). SONATA's rates compare favorably with those above. Furthermore, since both schemes [1] and [30] are gradient-type methods, unlike SONATA, their performance cannot benefit from function similarity, resulting in convergence rates independent on β. On the other hand, [13] explicitly considered the β-related setting, and proposed Network-DANE, a decentralization of the DANE algorithm. It turns out that Network-DANE is a special case of SONATA; there are however some important differences in the convergence analysis/results. First, convergence in [13] is established only for the *unconstrained* minimization of F ($G = 0$ and $\mathcal{K} = \mathbb{R}^d$) over *undirected* graphs, with *each* f_i assumed to be strongly convex. Second, convergence rates therein are more pessimistic than what predicted by our analysis. In fact, the best communication complexity of Network-DANE reads $\tilde{O}\big((1+(\beta/\mu)^2)(1-\rho)^{-1/2}\log(1/\epsilon)\big)$ for quadratic f_i's and worsens to $\tilde{O}\big(\kappa_\ell(1+\beta/\mu)(1-\rho)^{-1/2}\log(1/\epsilon)\big)$ for nonquadratic losses. Note that the latter is of the order of the worst-case rate of first-order methods, which do not benefit from function similarity. A direct comparison with Table 3, shows that SONATA' rates exhibit a better dependence on the optimization parameters (κ_g vs. κ_ℓ) and β/μ in all scenarios. In particular, in the β-related setting, SONATA retains faster rates, even when f_i's are nonquadratic.

1.3. *Paper organization*

Sec. 2 introduces the main assumptions on the optimization problem and network, along with some motivating examples from machine learning. The SONATA algorithm over undirected graphs is studied in Sec. 3; in particular, linear convergence is proved in Sec. 3.3, while a detailed discussion on the rate expression and its scalability properties is provided in Sec. 3.4. The case of time-varying, possibly directed, graphs is considered in Sec. 4. Finally, some numerical results supporting our theoretical findings are reported in Sec. 5. The study of SONATA when F is nonconvex can be found in the technical report [41].

2. Problem & Network Setting

This section summarizes the assumptions on the optimization problem and network setting. We also introduce a general learning problem over networks, which will be used as case study throughout the paper.

2.1. *Assumptions on Problem* (P)

Our algorithmic design and convergence results pertain to two problem settings, namely: i) the one where the local functions f_i are generic and unrelated (cf. Sec. 2.1.1), and ii) the case where they are related (cf. Sec. 2.1.2). These two settings are formally introduced below.

2.1.1. The unrelated setting

Consider the following standard assumption.

Assumption A (On Problem (P)). *A1 The set $\emptyset \neq \mathcal{K} \subseteq \mathbb{R}^d$ is closed and convex;*
A2 Each $f_i : \mathcal{O} \to \mathbb{R}$ is twice differentiable on the open set $\mathcal{O} \supseteq \mathcal{K}$ and convex;
A3 F satisfies

$$\mu \mathbf{I} \preceq \nabla^2 F(\mathbf{x}) \preceq L\mathbf{I}, \quad \forall \mathbf{x} \in \mathcal{K},$$

with $\mu > 0$ and $0 < L < \infty$;
A4 $G : \mathcal{K} \to \mathbb{R}$ is convex possibly nonsmooth.

Note that A3 together with A2 imply

$$\mu_i \mathbf{I} \preceq \nabla^2 f_i(\mathbf{x}) \preceq L_i\mathbf{I}, \quad \forall \mathbf{x} \in \mathcal{K}, \ \forall i \in [m], \tag{1}$$

for some $\mu_i \geq 0$ and $0 < L_i < \infty$. Unlike existing works (cf. Table 1), we do not require each f_i to be strongly convex but just F (cf. A3). Also, twice differentiability of f_i is not really necessary, but assumed here to simplify our derivations.

Under Assumption A, we define the global conditional number associated to (P):

$$\kappa_g \triangleq \frac{L}{\mu}. \tag{2}$$

Related quantities determining the (linear) convergence rate of existing distributed algorithms are (cf. Table 1):

$$\kappa_\ell \triangleq \frac{L_{\text{mx}}}{\mu_{\text{mn}}}, \quad \hat{\kappa} \triangleq \frac{L_{\text{mx}}}{(1/m)\sum_i \mu_i}, \quad \breve{\kappa} \triangleq \frac{L_{\text{mx}}}{\mu}, \quad \text{and} \quad \bar{\kappa} \triangleq \frac{L_{\text{mx}}}{\mu_{\text{mx}}}, \tag{3}$$

where

$$L_{\text{mx}} \triangleq \max_{i=1,\ldots,m} L_i, \quad \mu_{\text{mn}} \triangleq \min_{i=1,\ldots,m} \mu_i, \quad \text{and} \quad \mu_{\text{mx}} \triangleq \max_{i=1,\ldots,m} \mu_i. \tag{4}$$

When $\mu_i = 0$, we set $\kappa_\ell = \infty$. It is not difficult to check that κ_g can be much smaller than $\breve{\kappa}$, $\bar{\kappa}$, $\hat{\kappa}$ and κ_ℓ, as shown in the following example.
Example 1: Consider the following instance of Problem (P):

$$f_i(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \left(\mathtt{a}\mathbf{I} + m \cdot \mathtt{b}\, \text{diag}(\mathbf{e}_i)\right)\mathbf{x}, \quad F(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^m f_i(\mathbf{x}) = \frac{\mathtt{a}+\mathtt{b}}{2}\|\mathbf{x}\|^2,$$

$G = 0$, and $\mathcal{K} = \mathbb{R}^d$, where $\mathbf{e}_i$ is the i-th canonical vector, and $\mathtt{a}$, $\mathtt{b}$ are some positive constants. We have $\mu_i = \mathtt{a}$, $L_i = \mathtt{a} + m \cdot \mathtt{b}$, and $\mu = L = \mathtt{a} + \mathtt{b}$. Therefore,

$$\frac{\kappa_\ell}{\kappa_g} = \frac{\hat{\kappa}}{\kappa_g} = \frac{\bar{\kappa}}{\kappa_g} = 1 + m \cdot \frac{\mathtt{b}}{\mathtt{a}} \quad \text{and} \quad \frac{\breve{\kappa}}{\kappa_g} = \frac{1 + m \cdot \mathtt{b}/\mathtt{a}}{1 + \mathtt{b}/\mathtt{a}},$$

which all grow indefinitely as $\mathtt{b}/\mathtt{a}$ or m increase. □

In the setting above, our goal is to design linearly convergent distributed algorithms whose iterations complexity is proportional to κ_g, instead of the larger quantities in (3).

2.1.2. The β-related setting

This setting considers explicitly the case where the functions f_i are similar, in the sense defined below [2].

Definition 2.1 (β-related f_i's)**.** The local functions f_i's (satisfying Assumption A) are called β-related if $\left\|\nabla^2 F(\mathbf{x}) - \nabla^2 f_i(\mathbf{x})\right\|_2 \leq \beta$, for all $\mathbf{x} \in \mathcal{K}$ and some $\beta \geq 0$.

The more similar the f_i's, the smaller β. For arbitrary f_i's, β is of the order of

$$\beta \leq \max_{i=1,\ldots,m} \sup_{\mathbf{x}\in\mathcal{K},\, \|\mathbf{u}\|=1} \left|\mathbf{u}^\top \left(\nabla^2 F(\mathbf{x}) - \nabla^2 f_i(\mathbf{x})\mathbf{u}\right)\right| \leq \max_{i=1,\ldots,m} \max\left\{|L - \mu_i|,\, |\mu - L_i|\right\}. \tag{5}$$

The interesting case is when $1 + \beta/\mu << \kappa_g$; a specific example is discussed next.

Example 2: Convex-Lipschitz-bounded learning problems over networks Consider a stochastic learning setting whereby the ultimate goal is to minimize some population objective

$$\mathbf{x}^\star \in \underset{\mathbf{x}\in\mathcal{H}}{\text{argmin}}\, F(\mathbf{x}), \quad \text{with} \quad F(\mathbf{x}) \triangleq \mathbb{E}_{\mathbf{z}\sim\mathcal{P}}\left[f(\mathbf{x};\mathbf{z})\right], \tag{6}$$

where $f : \mathcal{O} \times \mathcal{Z} \to \mathbb{R}$ is the loss function, assumed to be C^2, convex (but not strongly convex), and L-smooth on the open set $\mathcal{O} \supset \mathcal{H}$, for all $\mathbf{z} \in \mathcal{Z}$; $\mathcal{H} \subseteq \mathbb{R}^d$ is the set of hypothesis classes, assumed to be convex and closed; $\mathcal{Z}$ is the set of examples; and $\mathcal{P}$ is the (unknown) distributed of $\mathbf{z} \in \mathcal{Z}$. Furthermore, we assume that any $\mathbf{x}^\star \in \mathcal{B}_B \triangleq \{\mathbf{x} : \|\mathbf{x}\| \le B\}$, for some $0 < B < \infty$. This setting includes, for example, supervised generalized linear models, where $\mathbf{z} = (\mathbf{w}, y)$ and $f(\mathbf{x}; (\mathbf{w}, y)) = \ell(\boldsymbol{\phi}(\mathbf{w})^\top \mathbf{x}; y)$, for some (strongly) convex loss $\ell(\bullet; y)$ and feature mapping $\boldsymbol{\phi}$. For instance, in linear regression, $f(\mathbf{x}; (\mathbf{w}, y)) = (y - \boldsymbol{\phi}(\mathbf{w})^\top \mathbf{x})^2$, with $\boldsymbol{\phi}(\mathbf{w}) \in \mathbb{R}^d$ and $y \in \mathbb{R}$; for logistic regression, we have $f(\mathbf{x}; (\mathbf{w}, y)) = \log(1 + \exp(-y(\boldsymbol{\phi}(\mathbf{w})^\top \mathbf{x})))$, with $\mathbf{w} \in \mathbb{R}^d$ and $y \in \{-1, 1\}$.

To solve (6), the m agents have access only to a finite number, say $N = nm$, of i.i.d. samples from the distribution $\mathcal{P}$, evenly and randomly distributed over the network. Using the notation introduced in Sec. 1, the ERM problem reads:

$$\widehat{\mathbf{x}} \triangleq \underset{\mathbf{x} \in \mathcal{H}}{\operatorname{argmin}}\, \widehat{F}(\mathbf{x}) \triangleq \frac{1}{m} \sum_{i=1}^{m} f_i(\mathbf{x}; \mathcal{D}^{(i)}), \qquad f_i(\mathbf{x}; \mathcal{D}^{(i)}) = \frac{1}{n} \sum_{j=1}^{n} f(\mathbf{x}; \mathbf{z}_j^{(i)}) + \frac{\lambda}{2} \|\mathbf{x}\|^2, \tag{7}$$

where f_i is regularized empirical loss of agent i, λ-strongly convex. Clearly (7) is an instance of (P), satisfying Assumption A.

For the ERM problems (7) we derive next the associated β/μ and contrasts with κ_g. $\widehat{F}$ is λ-strongly convex; therefore, we can set $\mu = \lambda$. The optimal choice of λ is the one minimizing the statistical error resulting in using $\widehat{\mathbf{x}}$ as proxy for $\mathbf{x}^\star$. We have [36, Th. 7], with high probability, $F(\widehat{\mathbf{x}}) - F(\mathbf{x}^\star) \le \frac{\lambda}{2} \|\boldsymbol{\theta}^\star\|^2 + \mathcal{O}\big(\frac{G_f^2}{\lambda N}\big) \le \mathcal{O}\big(\lambda\, B^2 + \frac{G_f^2}{\lambda N}\big)$, where G_f is the Lipschitz constant of $f(\bullet; \mathbf{z})$ on $\mathcal{H} \bigcap \mathcal{B}_B$, for all $\mathbf{z} \in \mathcal{Z}$. The optimal choice of λ and resulting minimum error rate are then

$$\lambda = \mathcal{O}\left(\sqrt{\frac{G^2}{B^2\, N}}\right) \quad \Rightarrow \quad F(\widehat{\mathbf{x}}) - F(\mathbf{x}^\star) \le \mathcal{O}\left(\sqrt{\frac{G^2 B^2}{N}}\right). \tag{8}$$

An estimate of β can be obtained exploring the statistical similarity of the local empirical losses f_i in (7). Under the additional assumption that $\nabla^2 f(\bullet; \mathbf{z})$ is M-Lipchitz on $\mathcal{H}$, for all $\mathbf{z} \in \mathcal{Z}$, a minor modification of [58, Lemma 6] applied to (6)-(7), yields: with high probability,

$$\sup_{\mathbf{x} \in \mathcal{B}_B} \left\| \nabla^2 f_i(\mathbf{x}; \mathbf{z}) - \nabla^2 \hat{F}(\mathbf{x}) \right\| \le \beta, \quad \forall \mathbf{z} \in \mathcal{Z},\, i \in [m],$$

with

$$\beta = \begin{cases} \widetilde{\mathcal{O}}\left(\sqrt{\frac{L^2}{n}}\right), & \text{if } M = 0; \\ \widetilde{\mathcal{O}}\left(\sqrt{\frac{L^2\, d}{n}}\right), & \text{otherwise,} \end{cases} \tag{9}$$

where $\widetilde{\mathcal{O}}$ hides the log-factor dependence. Note that when $f(\bullet; \mathbf{z})$ is quadratic (i.e., $M = 0$), β scales favorably with the dimension d.

Based on (8)-(9), an estimate of β/μ and κ_g for (7) reads:

$$1 + \frac{\beta}{\mu} = 1 + \widetilde{\mathcal{O}}\left(L\sqrt{d\, m}\right) \quad \text{and} \quad \kappa_g = 1 + \widetilde{\mathcal{O}}\left(L\sqrt{d\, m\, n}\right). \tag{10}$$

Note that κ_g increases with the local sample size n while β/μ *does not* (neglecting log-factors). It turns out that algorithms converging at a rate depending on κ_g exhibit a speed-accuracy dilemma: small statistical errors in (8) (larger n) are achieved at the cost of more iterations (larger κ_g). In this setting, it is thus desirable to design distributed algorithms whose rate depends on β/μ rather than κ_g.

2.2. *Network setting*

We will consider separately two network settings: i) the case where the underlying communication graph is fixed and undirected; and ii) the more general setting of time-varying directed graphs.

Undirected, static graphs: When the network of the agent is modeled as a fixed, undirected graph, we write $\mathcal{G} \triangleq (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} \triangleq \{1, \ldots, m\}$ denotes the vertex set–the set of agents–while $\mathcal{E} \triangleq \{(i, j) \,|\, i, j \in \mathcal{V}\}$ represents the set of edges–the communication links; $(i, j) \in \mathcal{E}$ iff there exists a communication link between agent i and j. We make the following standard assumption on the graph connectivity.

Assumption B (On the network)**.** *The graph $\mathcal{G}$ is connected.*

Directed, time-varying graphs In this setting, communication network is modeled as a time-varying digraph: time is slotted, and at time-frame ν, the digraph reads $\mathcal{G}^\nu = (\mathcal{V}, \mathcal{E}^\nu)$, where the set of edges $\mathcal{E}^\nu$ represents the agents' communication links: $(i, j) \in \mathcal{E}^\nu$ there is a link going from agent i to agent j. We make the following standard assumption on the "long-term" connectivity property of the graphs.

Assumption B$'$ (On the network). *The graph sequence $\{\mathcal{G}^\nu\}$, $\nu = 0, 1, \ldots,$ is B-strongly connected, i.e., there exists a finite integer $B > 0$ such that the graph with edge set $\cup_{t=\nu B}^{(\nu+1)B-1} \mathcal{E}^t$ is strongly connected, for all $\nu = 0, 1, \ldots$.*

The network setting covers, as special case, star-networks, i.e., architectures with a centralized node (a.k.a. master node) connected to all the others (a.k.a. workers). This is the typical computational architecture of several federated learning systems.

3. The SONATA algorithm over undirected graphs

We recall here the SONATA/NEXT algorithm [5, 35], customized to undirected, static, graphs. Each agent i maintains and updates iteratively a local copy $\mathbf{x}_i \in \mathbb{R}^d$ of the global variable $\mathbf{x}$, along with the auxiliary variable $\mathbf{y}_i \in \mathbb{R}^d$, which estimates the gradient of F. Denoting by $\mathbf{x}_i^\nu$ (resp. $\mathbf{y}_i^\nu$) the values of $\mathbf{x}_i$ (resp. $\mathbf{y}_i$) at iteration $\nu = 0, 1, \ldots,$ the SONATA algorithms is described in Algorithm 1. In words, each agent i, given the current iterates $\mathbf{x}_i^\nu$ and $\mathbf{y}_i^\nu$, first solves a strongly convex optimization problem wherein $\widetilde{F}_i$ is an approximation of the sum-cost F at $\mathbf{x}_i^\nu$; $\widetilde{f}_i$ in (11a) is a strongly convex function, which plays the role of a surrogate of f_i (cf. Assumption C below) while $\mathbf{y}_i^\nu$ acts as approximation of the gradient of F at $\mathbf{x}_i^\nu$, that is, $\nabla F(\mathbf{x}_i^\nu) \approx \mathbf{y}_i^\nu$ (see discussion below). Then, agent i updates $\mathbf{x}_i^\nu$ along the local direction $\mathbf{d}_i^\nu$ [cf. (11b)], using the step-size $\alpha \in (0, 1]$; the resulting point $\mathbf{x}_i^{\nu+1/2}$ is broadcast to its neighbors. The update $\mathbf{x}_i^{\nu+1/2} \to \mathbf{x}_i^{\nu+1}$ is obtained via the consensus step (11c) while the y-variables are updated via the perturbed consensus (11d), aiming at tracking $\nabla F(\mathbf{x}_i^\nu)$.

The main assumptions underlying the convergence of SONATA are discussed next.

Algorithm 1 SONATA over undirected graphs

```
Data: $\mathbf{x}_i^0 \in \mathcal{K}$ and $\mathbf{y}_i^0 = \nabla f_i(\mathbf{x}_i^0)$, $i \in [m]$.
Iterate: $\nu = 1, 2, \ldots$
[S.1] [Distributed Local Optimization] Each agent $i$ solves
    $\widehat{\mathbf{x}}_i^\nu \triangleq \underset{\mathbf{x}_i \in \mathcal{K}}{\operatorname{argmin}} \; \underbrace{\widetilde{f}_i(\mathbf{x}_i; \mathbf{x}_i^\nu) + \big(\mathbf{y}_i^\nu - \nabla f_i(\mathbf{x}_i^\nu)\big)^\top (\mathbf{x}_i - \mathbf{x}_i^\nu)}_{\widetilde{F}_i(\mathbf{x}_i; \mathbf{x}_i^\nu)} + G(\mathbf{x}_i), \qquad (11a)$
    and updates
    $\mathbf{x}_i^{\nu+\frac{1}{2}} = \mathbf{x}_i^\nu + \alpha \cdot \mathbf{d}_i^\nu, \quad \text{with} \quad \mathbf{d}_i^\nu \triangleq \widehat{\mathbf{x}}_i^\nu - \mathbf{x}_i^\nu; \qquad (11b)$
[S.2] [Information Mixing] Each agent $i$ computes
    (a) Consensus
        $\mathbf{x}_i^{\nu+1} = \sum_{j=1}^m w_{ij} \mathbf{x}_j^{\nu+\frac{1}{2}}, \qquad (11c)$
    (b) Gradient tracking
        $\mathbf{y}_i^{\nu+1} = \sum_{j=1}^m w_{ij} \big(\mathbf{y}_j^\nu + \nabla f_j(\mathbf{x}_j^{\nu+1}) - \nabla f_j(\mathbf{x}_j^\nu)\big). \qquad (11d)$
end
```

• *On the subproblem (11a) and surrogate functions* $\widetilde{\boldsymbol{f}}_{\boldsymbol{i}}$ The surrogate functions satisfy the following conditions.

Assumption C. *Each* $\widetilde{f}_i : \mathcal{O} \times \mathcal{O} \to \mathbb{R}$ *is* C^2 *and satisfies*

(i) $\nabla \widetilde{f}_i(\mathbf{x}; \mathbf{x}) = \nabla f_i(\mathbf{x})$, *for all* $\mathbf{x} \in \mathcal{K}$;

(ii) $\nabla \widetilde{f}_i(\bullet; \mathbf{x})$ *is* $\widetilde{L}_i$*-Lipschitz continuous on* $\mathcal{K}$, *for all* $\mathbf{x} \in \mathcal{K}$;

(iii) $\widetilde{f}_i(\bullet; \mathbf{x})$ *is* $\widetilde{\mu}_i$*-strongly convex on* $\mathcal{K}$, *for all* $\mathbf{x} \in \mathcal{K}$;

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

The assumption states that $\widetilde{f}_i$ should be regarded as a surrogate of f_i that preserves at each iterate $\mathbf{x}_i^\nu$ the first order properties of f_i. Conditions (i)-(iii) are certainly satisfied if one uses the classical linearization of f_i, that is,

$$\widetilde{f}_i(\mathbf{x}_i; \mathbf{x}_i^\nu) = \nabla f_i(\mathbf{x}_i^\nu)^\top (\mathbf{x}_i - \mathbf{x}_i^\nu) + \frac{\tau_i}{2} \|\mathbf{x}_i - \mathbf{x}_i^\nu\|^2, \tag{12}$$

with $\tau_i > 0$, which leads to the standard proximal-gradient update for $\widehat{\mathbf{x}}_i$. Note that if, in addition, $G = 0$ and $\mathcal{K} = \mathbb{R}^d$, (11a)–(11c) reduces to the standard (ATC) consensus/gradient-tracking step (setting $\alpha = 1$ and absorbing $1/\tau_i$ into the common stepsize γ): $\mathbf{x}_i^{\nu+1} = \sum_j w_{ij}(\mathbf{x}_i^\nu - \gamma \, \mathbf{y}_i^\nu)$ [21, 29, 52]. However, Assumption C allows us to cover a much wider array of approximations that better suit the geometry of the problem at hand, enhancing convergence speed. For instance, on the opposite side of (12), we have a surrogate retaining all the structure of f_i, such as

$$\widetilde{f}_i(\mathbf{x}_i; \mathbf{x}_i^\nu) = f_i(\mathbf{x}_i) + \frac{\tau_i}{2} \|\mathbf{x}_i - \mathbf{x}_i^\nu\|^2, \tag{13}$$

with $\tau_i > 0$. Using (13), one can rewrite (11a) as:

$$\widehat{\mathbf{x}}_i^\nu = \underset{\mathbf{x}_i \in \mathcal{K}}{\operatorname{argmin}} \Big(1 \cdot \underbrace{(\nabla f_i(\mathbf{x}_i^\nu) + \mathbf{y}_i^\nu)}_{\nabla g(\mathbf{x}_i^\nu)} - \underbrace{(\nabla f_i(\mathbf{x}_i^\nu) + \tau_i \mathbf{x}_i^\nu)}_{\nabla \omega(\mathbf{x}_i^\nu)}\Big)^\top \mathbf{x}_i + \underbrace{\Big(f_i(\mathbf{x}_i) + \frac{\tau_i}{2}\|\mathbf{x}_i\|^2\Big)}_{\omega(\mathbf{x}_i)} + G(\mathbf{x}_i), \tag{14}$$

which can be interpreted as a mirror-descent update (with step-size one) for the composite minimization of $g(\mathbf{x}_i) \triangleq f_i(\mathbf{x}_i) + (\mathbf{y}_i^\nu)^\top(\mathbf{x}_i - \mathbf{x}_i^\nu)$, based on the Bregman distance associated with the reference function $\omega(\mathbf{x}_i) \triangleq f_i(\mathbf{x}_i) + \tau_i/2\|\mathbf{x}_i\|^2$.

We refer the reader to [6, 33, 34] as good sources of examples of nonlinear surrogates satisfying Assumption C; here we only anticipate that, when the f_i's are sufficiently similar, higher order models such as (13) yield indeed faster rates of SONATA than those achievable using linear surrogates (12). Further intuition is provided next.

Under Assumption C, it is not difficult to check that, for every $i \in [m]$, there exist constants D_i^ℓ and D_i^u, $D_i^\ell \leq D_i^u$, such that

$$D_i^\ell \mathbf{I} \preceq \nabla^2 \widetilde{f}_i(\mathbf{x}, \mathbf{y}) - \nabla^2 F(\mathbf{x}) \preceq D_i^u \mathbf{I}, \quad \forall \mathbf{x}, \mathbf{y} \in \mathcal{K}; \qquad \text{let } \; D_i \triangleq \max\{|D_i^\ell|, |D_i^u|\}. \tag{15}$$

For instance, (15) holds with $D_i = \max\{|\widetilde{\mu}_i - L|, |\widetilde{L}_i - \mu|\}$. Roughly speaking, the smaller D_i the better $\widetilde{F}_i$ in (11a) approximates F. To see this, compare F and $\widetilde{F}_i$ up to the second order: there exist $\theta_1, \theta_2 \in (0,1)$ such that

$$\begin{aligned} \widetilde{F}_i(\mathbf{x}_i; \mathbf{x}_i^\nu) &= \widetilde{f}_i(\mathbf{x}_i^\nu; \mathbf{x}_i^\nu) + \big(\mathbf{y}_i^\nu - \nabla f_i(\mathbf{x}_i^\nu) + \nabla \widetilde{f}_i(\mathbf{x}_i^\nu; \mathbf{x}_i^\nu)\big)^\top (\mathbf{x}_i - \mathbf{x}_i^\nu) \\ &\quad + \frac{1}{2}(\mathbf{x}_i - \mathbf{x}_i^\nu)^\top \nabla^2 \widetilde{f}_i\big(\mathbf{x}_i^\nu + \theta_1(\mathbf{x}_i - \mathbf{x}_i^\nu); \mathbf{x}_i^\nu\big)(\mathbf{x}_i - \mathbf{x}_i^\nu) \\ F(\mathbf{x}_i) &= F(\mathbf{x}_i^\nu) + \nabla F(\mathbf{x}_i^\nu)^\top (\mathbf{x}_i - \mathbf{x}_i^\nu) \\ &\quad + \frac{1}{2}(\mathbf{x}_i - \mathbf{x}_i^\nu)^\top \nabla^2 F\big(\mathbf{x}_i^\nu + \theta_2(\mathbf{x}_i - \mathbf{x}_i^\nu); \mathbf{x}_i^\nu\big)(\mathbf{x}_i - \mathbf{x}_i^\nu). \end{aligned} \tag{16}$$

Noting that $\nabla \widetilde{f}_i(\mathbf{x}_i^\nu; \mathbf{x}_i^\nu) = \nabla f_i(\mathbf{x}_i^\nu)$ [Assumption C(i)] and $\nabla \widetilde{F}_i(\mathbf{x}_i^\nu; \mathbf{x}_i^\nu) = \mathbf{y}_i^\nu$, and anticipating $\|\nabla F(\mathbf{x}_i^\nu) - \mathbf{y}_i^\nu\| \to 0$ as $\nu \to \infty$ (see discussion below), it follows that $\widetilde{F}_i$ approximates F asymptotically, up to the first order. A better match, is achieved when D_i is sufficiently small. One can then expect that, if the local functions are sufficiently similar (β is small), surrogates $\widetilde{f}_i$ exploiting higher order information of f_i, such as (13), may be more effective than mere linearization. Our theoretical findings confirm the above intuition–see Sec. 3.4.

- *Consensus and gradient tracking steps (11c)-(11d)* In the consensus and tracking steps, the weights w_{ij}'s satisfy the following standard assumption.

Assumption D. *The weight matrix $\mathbf{W} \triangleq (w_{ij})_{i,j=1}^m$ has a sparsity pattern compliant with $\mathcal{G}$, that is*

D1 $w_{ii} > 0$, for all $i = 1, \ldots, m$;
D2 $w_{ij} > 0$, if $(i,j) \in \mathcal{E}$; and $w_{ij} = 0$ otherwise;

Furthermore, $\mathbf{W}$ is doubly stochastic, that is, $\mathbf{1}^\top \mathbf{W} = \mathbf{1}^\top$ and $\mathbf{W}\mathbf{1} = \mathbf{1}$.

Several rules have been proposed in the literature compliant with Assumption D, such as the Laplacian, the Metropolis-Hasting, and the maximum-degree weights rules [49].

Finally, we comment the anticipated gradient tracking property of the y-variables, that is, $\|\nabla F(\mathbf{x}_i^\nu) - \mathbf{y}_i^\nu\| \to 0$ as $\nu \to \infty$. Define the average processes

$$\bar{\mathbf{y}}^\nu \triangleq \frac{1}{m}\sum_{i=1}^m \mathbf{y}_i^\nu \quad \text{and} \quad \overline{\nabla \mathbf{f}}^\nu \triangleq \frac{1}{m}\sum_{i=1}^m \nabla f_i(\mathbf{x}_i^\nu). \tag{17}$$

Summing (11d) over $i \in [m]$ and invoking the doubly stochasticity of $\mathbf{W}$; we have

$$\bar{\mathbf{y}}^{\nu+1} = \bar{\mathbf{y}}^\nu + \overline{\nabla \mathbf{f}}^{\nu+1} - \overline{\nabla \mathbf{f}}^\nu. \tag{18}$$

Applying (18) inductively and using the initial condition $\mathbf{y}_i^0 = \nabla f_i(\mathbf{x}_i^0)$, $i \in [m]$, yield

$$\bar{\mathbf{y}}^\nu = \overline{\nabla \mathbf{f}}^\nu, \quad \forall \nu = 0, 1, \ldots . \tag{19}$$

That is, the average of all the $\mathbf{y}_i^\nu$'s in the network is equal to that of the $\nabla f_i(\mathbf{x}_i^\nu)$'s, at every iteration ν. Assuming that consensus on $\mathbf{x}_i^\nu$'s and $\mathbf{y}_i^\nu$'s is asymptotically achieved, that is, $\|\mathbf{x}_i^\nu - \mathbf{x}_j^\nu\| \underset{\nu\to\infty}{\longrightarrow} 0$ and $\|\mathbf{y}_i^\nu - \mathbf{y}_j^\nu\| \underset{\nu\to\infty}{\longrightarrow} 0$, $i \neq j$, (19) would imply the desired gradient tracking property $\|\nabla F(\mathbf{x}_i^\nu) - \mathbf{y}_i^\nu\| \to 0$ as $\nu \to \infty$, for all $i \in [m]$.

3.1. *A special instance: SONATA on star-networks*

Although the main focus of the paper is the study of SONATA over meshed-networks, it is worth discussing here its special instance over star networks. Specifically, consider a star (unidirected) graph with m nodes, where one of them (the master node) connects with all the others (workers). The workers still own only one function f_i of the sum-cost F. Two common approaches developed in the literature to solve (P) in this setting are: (i) based upon receiving the gradients ∇f_i from the workers, the master solves (P) and broadcasts the updated vector variables to the workers; (ii) based upon receiving the full gradient ∇F and the current iterate from the master, all the workers solve locally an instance of (P) and send their outcomes to the master that averages them out, producing then the new iterate. Here we follow the latter approach; the algorithm is described in Algorithm 2, which corresponds to SONATA (up to a proper initialization), with weight matrix $\mathbf{W} = [\mathbf{1}, \mathbf{0}_{m,m-1}]\,[\mathbf{1}/m, \mathbf{0}_{m,m-1}]^\top$.

Connection with existing schemes SONATA-star, employing linear surrogates [cf. (12)] and $\alpha = 1$, reduces to the proximal gradient algorithm. When the surrogates (13) are used (and still $\alpha = 1$), SONATA-star coincides with the DANE algorithm [37] if $G = 0$ and to the CEASE (with averaging) algorithm [7] if $G \neq 0$. Nevertheless, our convergence rates improve on those of DANE and CEASE–see Sec. 3.4.1.

3.2. *Intermediate definitions*

We conclude this section introducing some quantities that will be used in the rest of the paper. We define the optimality gap as

$$p^\nu \triangleq \sum_{i=1}^m \big(U(\mathbf{x}_i^\nu) - U(\mathbf{x}^\star)\big), \tag{20}$$

where $\mathbf{x}^\star$ is the unique solution of Problem (P).

Algorithm 2 SONATA on Star-Networks (SONATA-Star)

```
Data: $\mathbf{x}^0 \in \mathcal{K}$.
Iterate: $\nu = 1, 2, ...$
  [S.1] Each worker $i$ evaluates $\nabla f_i(\mathbf{x}^\nu)$ and sends it to the master node;
  [S.2] The master broadcasts $\nabla F(\mathbf{x}^\nu) = 1/m \sum_{i=1}^m \nabla f_i(\mathbf{x}^\nu)$ to the workers;
  [S.3] Each worker $i$ computes
          $\widehat{\mathbf{x}}_i^\nu \triangleq \underset{\mathbf{x}_i \in \mathcal{K}}{\operatorname{argmin}} \widetilde{f}_i(\mathbf{x}_i; \mathbf{x}^\nu) + \big(\nabla F(\mathbf{x}^\nu) - \nabla f_i(\mathbf{x}_i^\nu)\big)^\top (\mathbf{x}_i - \mathbf{x}^\nu) + G(\mathbf{x}_i),$
        and sends $\widehat{\mathbf{x}}_i^\nu$ to the master;
  [S.4] The master computes
          $\mathbf{x}^{\nu+1} = \mathbf{x}^\nu + \alpha \left( \frac{1}{m} \sum_{i=1}^m \widehat{\mathbf{x}}_i^\nu - \mathbf{x}^\nu \right),$
        and sends it back to the workers.
end
```

We stack the local variables and gradients in the column vectors

$$\mathbf{x}^\nu \triangleq [\mathbf{x}_1^{\nu\top}, \ldots, \mathbf{x}_m^{\nu\top}]^\top, \ \mathbf{y}^\nu \triangleq [\mathbf{y}_1^{\nu\top}, \ldots, \mathbf{y}_m^{\nu\top}]^\top, \ \nabla \mathbf{f}^\nu \triangleq [\nabla f_1(\mathbf{x}_1^\nu)^\top, \ldots, \nabla f_m(\mathbf{x}_m^\nu)^\top]^\top. \tag{21}$$

The average of each of the vectors above is defined as $\bar{\mathbf{x}}^\nu \triangleq (1/m) \cdot \sum_{i=1}^m \mathbf{x}_i^\nu$. The consensus disagreements on $\mathbf{x}_i^\nu$'s and $\mathbf{y}_i^\nu$'s are

$$\mathbf{x}_\perp^\nu \triangleq \mathbf{x}^\nu - \mathbf{1}_m \otimes \bar{\mathbf{x}}^\nu \quad \text{and} \quad \mathbf{y}_\perp^\nu \triangleq \mathbf{y}^\nu - \mathbf{1}_m \otimes \bar{\mathbf{y}}^\nu, \tag{22}$$

respectively, while the gradient tracking error is defined as

$$\boldsymbol{\delta}^\nu \triangleq [\boldsymbol{\delta}_1^{\nu\top}, \ldots, \boldsymbol{\delta}_m^{\nu\top}]^\top, \quad \text{with} \quad \boldsymbol{\delta}_i^\nu \triangleq \nabla F(\mathbf{x}_i^\nu) - \mathbf{y}_i^\nu, \quad i = 1, \ldots, m. \tag{23}$$

Recalling L_i, $\widetilde{L}_i$, $\widetilde{\mu}_i$, D_i^ℓ and D_i as given in Assumptions A and C and (15), we introduce the following algorithm-dependent parameters

$$\begin{aligned} \widetilde{\mu}_{\text{mn}} &\triangleq \min_{i\in[m]} \widetilde{\mu}_i, \quad & \widetilde{L}_{\text{mx}} &\triangleq \max_{i\in[m]}, \widetilde{L}_i, \\ D_{\text{mn}}^\ell &\triangleq \min_{i\in[m]} D_i^\ell, \quad & D_{\text{mx}} &\triangleq \max_{i\in[m]} D_i. \end{aligned} \tag{24}$$

Finally, given the weight matrix $\mathbf{W}$, we define

$$\widehat{\mathbf{W}} \triangleq \mathbf{W} \otimes \mathbf{I}_d, \quad \text{and} \quad \mathbf{J} \triangleq \frac{1}{m} \mathbf{1}_m \mathbf{1}_m^\top \otimes \mathbf{I}_d. \tag{25}$$

Under Assumptions B and D, it is well known that (see, e.g., [44])

$$\rho \triangleq \sigma(\widehat{\mathbf{W}} - \mathbf{J}) < 1, \tag{26}$$

where $\sigma(\bullet)$ denotes the largest singular value of its argument.

3.3. *Linear convergence rate*

Our proof of linear rate of SONATA passes through the following steps. **Step 1:** We begin showing that the optimality gap p^ν converges linearly up to an error of the order of $\mathcal{O}(\|\mathbf{x}_\perp^\nu\|^2+\|\mathbf{y}_\perp^\nu\|^2)$, see Proposition 3.4. **Step 2** proves that $\|\mathbf{x}_\perp^\nu\|$ and $\|\mathbf{y}_\perp^\nu\|$ are also linearly convergent up to an error $\mathcal{O}(\|\mathbf{d}^\nu\|)$, see Proposition 3.5. In **Step 3** we close the loop establishing $\|\mathbf{d}^\nu\| = \mathcal{O}(\sqrt{p^\nu}+\|\mathbf{y}_\perp^\nu\|)$, see Proposition 3.6. Finally, in **Step 4**, we properly chain together the above inequalities (cf. Proposition 3.8), so that linear rate is proved for the sequences $\{p^\nu\}$, $\{\|\mathbf{x}_\perp^\nu\|^2\}$, $\{\|\mathbf{y}_\perp^\nu\|^2\}$, and $\{\|\mathbf{d}^\nu\|^2\}$–see Theorems 3.9 and 3.10. We will tacitly assume that Assumptions A, B, C, and D are satisfied.

3.3.1. Step 1:p^ν converges linearly up to $\mathcal{O}(\|\mathbf{x}_\perp^\nu\|^2+\|\mathbf{y}_\perp^\nu\|^2)$

Invoking the convexity of U and the doubly stochasticity of $\mathbf{W}$, we can bound $p^{\nu+1}$ as

$$p^{\nu+1} \leq \sum_{i=1}^m\sum_{j=1}^m w_{ij}\Big(U\big(\mathbf{x}_j^{\nu+\frac{1}{2}}\big)-U(\mathbf{x}^\star)\Big) = \sum_{i=1}^m \Big(U(\mathbf{x}_i^{\nu+\frac{1}{2}})-U(\mathbf{x}^\star)\Big). \tag{27}$$

We can now bound $U(\mathbf{x}_j^{\nu+\frac{1}{2}})$, regarding the local optimization (11a)-(11b) as a perturbed descent on the objective, whose perturbation is due to the tracking error $\boldsymbol{\delta}^\nu$. In fact, Lemma 3.1 below shows that, for sufficiently small α, the local update (11b) will decrease the objective value U up to some error, related to $\boldsymbol{\delta}_i^\nu$.

Lemma 3.1. *Let $\{\mathbf{x}_i^\nu\}$ be the sequence generated by SONATA; there holds:*

$$U(\mathbf{x}_i^{\nu+\frac{1}{2}}) \leq U(\mathbf{x}_i^\nu) - \alpha\left(\left(1-\frac{\alpha}{2}\right)\widetilde{\mu}_i + \frac{\alpha}{2}\cdot D_i^\ell\right)\|\mathbf{d}_i^\nu\|^2 + \alpha\|\mathbf{d}_i^\nu\|\|\boldsymbol{\delta}_i^\nu\|, \tag{28}$$

with D_i^ℓ and $\boldsymbol{\delta}_i^\nu$ are defined in (15) *and* (23)*, respectively.*

Proof. Consider the Taylor expansion of F:

$$\begin{aligned} F(\mathbf{x}_i^{\nu+\frac{1}{2}}) &= F(\mathbf{x}_i^\nu) + \nabla F(\mathbf{x}_i^\nu)^\top(\alpha\mathbf{d}_i^\nu) + (\alpha\mathbf{d}_i^\nu)^\top\mathbf{H}(\alpha\mathbf{d}_i^\nu),\\ &\overset{(23)}{=} F(\mathbf{x}_i^\nu) + \big(\boldsymbol{\delta}_i^\nu\big)^\top(\alpha\mathbf{d}_i^\nu) + \big(\mathbf{y}_i^\nu\big)^\top(\alpha\mathbf{d}_i^\nu) + (\alpha\mathbf{d}_i^\nu)^\top\mathbf{H}(\alpha\mathbf{d}_i^\nu), \end{aligned} \tag{29}$$

where $\mathbf{H} \triangleq \int_0^1(1-\theta)\nabla^2F(\theta\mathbf{x}_i^{\nu+\frac{1}{2}}+(1-\theta)\mathbf{x}_i^\nu)d\theta$.

Invoking the optimality of $\widehat{\mathbf{x}}_i^\nu$ and defining $\widetilde{\mathbf{H}}_i \triangleq \int_0^1\nabla^2\widetilde{f}_i(\theta\,\widehat{\mathbf{x}}_i^\nu+(1-\theta)\,\mathbf{x}_i^\nu;\mathbf{x}_i^\nu)d\theta$, we have

$$G(\mathbf{x}_i^\nu)-G(\widehat{\mathbf{x}}_i^\nu) \geq (\mathbf{d}_i^\nu)^\top\big(\nabla\widetilde{f}_i(\widehat{\mathbf{x}}_i^\nu;\mathbf{x}_i^\nu)+\mathbf{y}_i^\nu-\nabla f_i(\mathbf{x}_i^\nu)\big) = (\mathbf{d}_i^\nu)^\top\big(\mathbf{y}_i^\nu+\widetilde{\mathbf{H}}_i\mathbf{d}_i^\nu\big), \tag{30}$$

where the equality follows from $\nabla\widetilde{f}_i(\mathbf{x}_i^\nu;\mathbf{x}_i^\nu) = \nabla f_i(\mathbf{x}_i^\nu)$ and the integral form of the mean value theorem. Substituting (30) in (29) and using the convexity of G yield

$$
\begin{aligned}
&F(\mathbf{x}_i^{\nu+\frac{1}{2}})\\
&\leq F(\mathbf{x}_i^{\nu}) + (\boldsymbol{\delta}_i^{\nu})^{\top}(\alpha\mathbf{d}_i^{\nu}) + (\alpha\mathbf{d}_i^{\nu})^{\top}\mathbf{H}(\alpha\mathbf{d}_i^{\nu}) + \alpha\left(G(\mathbf{x}_i^{\nu}) - G(\widehat{\mathbf{x}}_i^{\nu}) - (\mathbf{d}_i^{\nu})^{\top}\widetilde{\mathbf{H}}_i\mathbf{d}_i^{\nu}\right)\\
&\leq F(\mathbf{x}_i^{\nu}) + (\boldsymbol{\delta}_i^{\nu})^{\top}(\alpha\mathbf{d}_i^{\nu}) + \alpha\left(-(\mathbf{d}_i^{\nu})^{\top}\widetilde{\mathbf{H}}_i\mathbf{d}_i^{\nu} + (\alpha\mathbf{d}_i^{\nu})^{\top}\mathbf{H}(\mathbf{d}_i^{\nu})\right) + G(\mathbf{x}_i^{\nu}) - G(\mathbf{x}_i^{\nu+\frac{1}{2}}).
\end{aligned}
\tag{31}
$$

It remains to bound $\alpha\mathbf{H} - \widetilde{\mathbf{H}}_i$. We proceed as follows:

$$
\begin{aligned}
&\alpha\mathbf{H} - \widetilde{\mathbf{H}}_i\\
&= \alpha\int_0^1 (1-\theta)\nabla^2 F(\theta\mathbf{x}_i^{\nu+\frac{1}{2}} + (1-\theta)\mathbf{x}_i^{\nu})d\theta - \int_0^1 \nabla^2\widetilde{f}_i(\theta\widehat{\mathbf{x}}_i^{\nu} + (1-\theta)\mathbf{x}_i^{\nu};\mathbf{x}_i^{\nu})d\theta\\
&\overset{(11b)}{=} \int_0^{\alpha}(1-\theta/\alpha)\nabla^2 F(\theta\widehat{\mathbf{x}}_i^{\nu} + (1-\theta)\mathbf{x}_i^{\nu})d\theta - \int_0^1 \nabla^2\widetilde{f}_i(\theta\widehat{\mathbf{x}}_i^{\nu} + (1-\theta)\mathbf{x}_i^{\nu};\mathbf{x}_i^{\nu})d\theta\\
&\overset{(a)}{\preceq} -\int_0^{\alpha}(1-\theta/\alpha)\cdot(D_i^{\ell})\,\mathbf{I}\,d\theta - \int_0^{\alpha}(\theta/\alpha)\nabla^2\widetilde{f}_i(\theta\widehat{\mathbf{x}}_i + (1-\theta)\mathbf{x}_i^{\nu};\mathbf{x}_i^{\nu})d\theta\\
&\quad - \int_{\alpha}^1 \nabla^2\widetilde{f}_i(\theta\,\widehat{\mathbf{x}}_i^{\nu} + (1-\theta)\,\mathbf{x}_i^{\nu};\mathbf{x}_i^{\nu})d\theta\\
&\overset{(b)}{\preceq} -\frac{1}{2}\alpha\,(D_i^{\ell})\,\mathbf{I} - \left(1-\frac{\alpha}{2}\right)\widetilde{\mu}_i\,\mathbf{I},
\end{aligned}
\tag{32}
$$

where in (a) we used $\nabla^2 F(\theta\widehat{\mathbf{x}}_i^{\nu} + (1-\theta)\mathbf{x}_i^{\nu}) \preceq -(D_i^{\ell})\mathbf{I} + \nabla^2\widetilde{f}_i(\theta\widehat{\mathbf{x}}_i^{\nu} + (1-\theta)\mathbf{x}_i^{\nu};\mathbf{x}_i^{\nu})$ [cf. (15)] while (b) follows from Assumption C(iii). Substituting (32) into (31) completes the proof □

We can now substitute (28) into (27) and get

$$
p^{\nu+1} \leq p^{\nu} + \sum_{i=1}^{m}\left\{\alpha\|\mathbf{d}_i^{\nu}\|\|\boldsymbol{\delta}_i^{\nu}\| - \alpha\left(1-\frac{\alpha}{2}\right)\widetilde{\mu}_i\|\mathbf{d}_i^{\nu}\|^2 - \frac{D_i^{\ell}}{2}\alpha^2\|\mathbf{d}_i^{\nu}\|^2\right\}
\tag{33a}
$$

$$
\overset{(a)}{\leq} p^{\nu} - \left(\left(1-\frac{\alpha}{2}\right)\widetilde{\mu}_{\text{mn}} + \frac{\alpha D_{\text{mn}}^{\ell}}{2} - \frac{1}{2}\epsilon_{opt}\right)\alpha\|\mathbf{d}^{\nu}\|^2 + \frac{1}{2}\epsilon_{opt}^{-1}\,\alpha\cdot\|\boldsymbol{\delta}^{\nu}\|^2,
\tag{33b}
$$

where in (a) we used Young's inequality, with $\epsilon_{opt} > 0$ satisfying

$$
\left(1-\frac{\alpha}{2}\right)\widetilde{\mu}_{\text{mn}} + \frac{\alpha D_{\text{mn}}^{\ell}}{2} - \frac{1}{2}\epsilon_{opt} > 0;
\tag{34}
$$

and D_{mn}^{ℓ} is defined in (24).

Next we lower bound $\|\mathbf{d}^{\nu}\|^2$ in terms of the optimality gap.

Lemma 3.2. *The following lower bound holds for* $\|\mathbf{d}^{\nu}\|^2$*:*

$$
\alpha\,\|\mathbf{d}^{\nu}\|^2 \geq \frac{\mu}{D_{\text{mx}}^2}\left(p^{\nu+1} - (1-\alpha)p^{\nu} - \frac{\alpha}{\mu}\|\boldsymbol{\delta}^{\nu}\|^2\right),
\tag{35}
$$

where D_{mx} *is defined in* (24).

Proof. Invoking the optimality condition of $\widehat{\mathbf{x}}_i^\nu$, yields

$$G(\mathbf{x}^\star) - G(\widehat{\mathbf{x}}_i^\nu) \geq -(\mathbf{x}^\star - \widehat{\mathbf{x}}_i^\nu)^\top \Big(\nabla \widetilde{f}_i(\widehat{\mathbf{x}}_i^\nu; \mathbf{x}_i^\nu) + \mathbf{y}_i^\nu - \nabla f_i(\mathbf{x}_i^\nu)\Big). \tag{36}$$

Using the μ-strong convexity of F, we can write

$$\begin{aligned}
U(\mathbf{x}^\star) &\geq U(\widehat{\mathbf{x}}_i^\nu) + G(\mathbf{x}^\star) - G(\widehat{\mathbf{x}}_i^\nu) + \nabla F(\widehat{\mathbf{x}}_i^\nu)^\top(\mathbf{x}^\star - \widehat{\mathbf{x}}_i^\nu) + \frac{\mu}{2}\|\mathbf{x}^\star - \widehat{\mathbf{x}}_i^\nu\|^2 \\
&\overset{(36)}{\geq} U(\widehat{\mathbf{x}}_i^\nu) + \Big(\nabla F(\widehat{\mathbf{x}}_i^\nu) - \nabla \widetilde{f}_i(\widehat{\mathbf{x}}_i^\nu; \mathbf{x}_i^\nu) - \big(\mathbf{y}_i^\nu - \nabla f_i(\mathbf{x}_i^\nu)\big)\Big)^\top (\mathbf{x}^\star - \widehat{\mathbf{x}}_i^\nu) + \frac{\mu}{2}\|\mathbf{x}^\star - \widehat{\mathbf{x}}_i^\nu\|^2 \\
&= U(\widehat{\mathbf{x}}_i^\nu) + \frac{\mu}{2}\left\|\mathbf{x}^\star - \widehat{\mathbf{x}}_i^\nu + \frac{1}{\mu}\Big(\nabla F(\widehat{\mathbf{x}}_i^\nu) - \nabla \widetilde{f}_i(\widehat{\mathbf{x}}_i^\nu; \mathbf{x}_i^\nu) - \big(\mathbf{y}_i^\nu - \nabla f_i(\mathbf{x}_i^\nu)\big)\Big)\right\|^2 \\
&\quad - \frac{1}{2\mu}\left\|\nabla F(\widehat{\mathbf{x}}_i^\nu) - \nabla \widetilde{f}_i(\widehat{\mathbf{x}}_i^\nu; \mathbf{x}_i^\nu) - \big(\mathbf{y}_i^\nu - \nabla f_i(\mathbf{x}_i^\nu)\big)\right\|^2 \\
&\geq U(\widehat{\mathbf{x}}_i^\nu) - \frac{1}{2\mu}\left\|\nabla F(\widehat{\mathbf{x}}_i^\nu) \pm \nabla F(\mathbf{x}_i^\nu) - \nabla \widetilde{f}_i(\widehat{\mathbf{x}}_i^\nu; \mathbf{x}_i^\nu) - \big(\mathbf{y}_i^\nu - \nabla f_i(\mathbf{x}_i^\nu)\big)\right\|^2 \\
&\geq U(\widehat{\mathbf{x}}_i^\nu) - \frac{1}{\mu}\left\|\nabla F(\widehat{\mathbf{x}}_i^\nu) - \nabla F(\mathbf{x}_i^\nu) + \nabla f_i(\mathbf{x}_i^\nu) - \nabla \widetilde{f}_i(\widehat{\mathbf{x}}_i^\nu; \mathbf{x}_i^\nu)\right\|^2 - \frac{1}{\mu}\|\boldsymbol{\delta}_i^\nu\|^2 \\
&= U(\widehat{\mathbf{x}}_i^\nu) - \frac{1}{\mu}\left\|\int_0^1 \Big(\nabla^2 F(\theta\widehat{\mathbf{x}}_i^\nu + (1-\theta)\mathbf{x}_i^\nu) - \nabla^2 \widetilde{f}_i(\theta\widehat{\mathbf{x}}_i^\nu + (1-\theta)\mathbf{x}_i^\nu; \mathbf{x}_i^\nu)\Big)\,(\mathbf{d}_i^\nu)\,\mathrm{d}\theta\right\|^2 - \frac{1}{\mu}\|\boldsymbol{\delta}_i^\nu\|^2 \\
&\geq U(\widehat{\mathbf{x}}_i^\nu) - \frac{D_i^2}{\mu}\,\|\mathbf{d}_i^\nu\|^2 - \frac{1}{\mu}\|\boldsymbol{\delta}_i^\nu\|^2.
\end{aligned}$$

Rearranging the terms and summing over $i \in [m]$, yields

$$\|\mathbf{d}^\nu\|^2 \geq \frac{\mu}{D_{\mathrm{mx}}^2}\left(\sum_{i=1}^m \big(U(\widehat{\mathbf{x}}_i^\nu) - U(\mathbf{x}^\star)\big) - \frac{1}{\mu}\|\boldsymbol{\delta}^\nu\|^2\right). \tag{37}$$

Using (27) in conjunction with $U(\mathbf{x}_i^{\nu+\frac{1}{2}}) \leq \alpha U(\widehat{\mathbf{x}}_i^\nu) + (1-\alpha)U(\mathbf{x}_i^\nu)$ leads to

$$\alpha \sum_{i=1}^m \left(U(\widehat{\mathbf{x}}_i^\nu) - U(\mathbf{x}^\star)\right) \geq p^{\nu+1} - (1-\alpha)p^\nu. \tag{38}$$

Combining (37) with (38) provides the desired result (35). □

As last step, we upper bound $\|\boldsymbol{\delta}^\nu\|^2$ in (33) in terms of the consensus errors $\|\mathbf{x}_\perp^\nu\|^2$ and $\|\mathbf{y}_\perp^\nu\|^2$.

Lemma 3.3. *The following upper bound holds for the tracking error $\|\boldsymbol{\delta}^\nu\|^2$:*

$$\|\boldsymbol{\delta}^\nu\|^2 \leq 4L_{\mathrm{mx}}^2\|\mathbf{x}_\perp^\nu\|^2 + 2\|\mathbf{y}_\perp^\nu\|^2, \tag{39}$$

where L_{mx} is defined in (4).

Proof.

$$\begin{aligned}
\|\boldsymbol{\delta}^\nu\|^2 &\overset{(23)}{=} \sum_{i=1}^m \|\nabla F(\mathbf{x}_i^\nu) \pm \bar{\mathbf{y}}^\nu - \mathbf{y}_i^\nu\|^2 \\
&\overset{(17)}{=} \frac{1}{m^2} \sum_{i=1}^m \left\| \sum_{j=1}^m \nabla f_j(\mathbf{x}_i^\nu) - \sum_{j=1}^m \nabla f_j(\mathbf{x}_j^\nu) + m \cdot \bar{\mathbf{y}}^\nu - m \cdot \mathbf{y}_i^\nu \right\|^2 \\
&\overset{(1),(4)}{\leq} \frac{1}{m^2} \sum_{i=1}^m \left(2m \sum_{j=1}^m L_{\text{mx}}^2 \|\mathbf{x}_i^\nu - \mathbf{x}_j^\nu\|^2 + 2m^2 \|\bar{\mathbf{y}}^\nu - \mathbf{y}_i^\nu\|^2 \right) \\
&= 4L_{\text{mx}}^2 \|\mathbf{x}_\perp^\nu\|^2 + 2\|\mathbf{y}_\perp^\nu\|^2.
\end{aligned}$$

□

We are ready to prove the linear convergence of the optimality gap up to consensus errors. The result is summarized in Proposition 3.4 below. The proof follows readily multiplying (33) and (35) by $\widetilde{\mu}_{\text{mn}} - \frac{L}{2}\alpha - \frac{1}{2}\epsilon_{opt}$ and $6(L^2 + \widetilde{L}_{\text{mx}}^2)/\mu$, respectively, adding them together to cancel out $\|\mathbf{d}^\nu\|$, and using (39) to bound $\|\boldsymbol{\delta}^\nu\|^2$.

Proposition 3.4. *The optimality gap p^ν [cf. (20)] satisfies*

$$p^{\nu+1} \leq \sigma(\alpha) \cdot p^\nu + \eta(\alpha) \cdot \left(4L_{\text{mx}}^2 \|\mathbf{x}_\perp^\nu\|^2 + 2\|\mathbf{y}_\perp^\nu\|^2\right), \tag{40}$$

where $\sigma(\alpha) \in (0, 1)$ and $\eta(\alpha) > 0$ are defined as

$$\sigma(\alpha) \triangleq 1 - \alpha \frac{\left(1 - \frac{\alpha}{2}\right) \widetilde{\mu}_{\text{mn}} + \frac{D_{\text{mn}}^\ell}{2}\alpha - \frac{1}{2}\epsilon_{opt}}{\frac{D_{\text{mx}}^2}{\mu} + \left(1 - \frac{\alpha}{2}\right) \widetilde{\mu}_{\text{mn}} + \frac{D_{\text{mn}}^\ell}{2}\alpha - \frac{1}{2}\epsilon_{opt}}, \tag{41}$$

$$\eta(\alpha) \triangleq \frac{\frac{1}{2}\epsilon_{opt}^{-1}\alpha \cdot \frac{D_{\text{mx}}^2}{\mu} + \frac{\alpha}{\mu} \cdot \left(\left(1 - \frac{\alpha}{2}\right) \widetilde{\mu}_{\text{mn}} + \frac{D_{\text{mn}}^\ell}{2}\alpha - \frac{1}{2}\epsilon_{opt} \right)}{\frac{D_{\text{mx}}^2}{\mu} + \left(1 - \frac{\alpha}{2}\right) \widetilde{\mu}_{\text{mn}} + \frac{D_{\text{mn}}^\ell}{2}\alpha - \frac{1}{2}\epsilon_{opt}}; \tag{42}$$

ϵ_{opt} satisfies (34); and L_{mx} and $\widetilde{\mu}_{\text{mn}}$, D_{mn}^ℓ, D_{mx} are defined in (4) and (24), respectively.

3.3.2. Step 2: $\|\mathbf{x}_\perp^\nu\|$ and $\|\mathbf{y}_\perp^\nu\|$ linearly converge up to $\mathcal{O}(\|\mathbf{d}^\nu\|)$

We upper bound $\|\mathbf{x}_\perp^\nu\|$ and $\|\mathbf{y}_\perp^\nu\|$ in terms of $\|\mathbf{d}^\nu\|$. We begin rewriting the SONATA algorithm (11a)-(11d) in vector-matrix form; using (21) and (25),we have

$$\mathbf{x}^{\nu+1} = \widehat{\mathbf{W}}(\mathbf{x}^\nu + \alpha\mathbf{d}^\nu) \tag{43a}$$

$$\mathbf{y}^{\nu+1} = \widehat{\mathbf{W}}(\mathbf{y}^\nu + \nabla\mathbf{f}^{\nu+1} - \nabla\mathbf{f}^\nu). \tag{43b}$$

Noting that $\mathbf{x}_\perp^\nu = (\mathbf{I} - \mathbf{J})\mathbf{x}^\nu$ [similarly, $\mathbf{y}_\perp^\nu = (\mathbf{I} - \mathbf{J})\mathbf{y}^\nu$] and $(\mathbf{I} - \mathbf{J})\widehat{\mathbf{W}} = \widehat{\mathbf{W}} - \mathbf{J}$ (due to the doubly stochasticity of $\mathbf{W}$), it follows from (43) that

$$\mathbf{x}_\perp^{\nu+1} = (\widehat{\mathbf{W}} - \mathbf{J})(\mathbf{x}_\perp^\nu + \alpha\mathbf{d}^\nu) \tag{44}$$

$$\mathbf{y}_\perp^{\nu+1} = (\widehat{\mathbf{W}} - \mathbf{J})(\mathbf{y}_\perp^\nu + \nabla\mathbf{f}^{\nu+1} - \nabla\mathbf{f}^\nu). \tag{45}$$

Using (44)-(45), Proposition 3.5 below establishes linear convergence of the consensus errors $\mathbf{x}_\perp^\nu$ and $\mathbf{y}_\perp^\nu$, up to a perturbation.

Proposition 3.5. *There holds:*

$$\|\mathbf{x}_\perp^{\nu+1}\| \leq \rho\|\mathbf{x}_\perp^\nu\| + \alpha\rho\|\mathbf{d}^\nu\|, \tag{46a}$$

$$\|\mathbf{y}_\perp^{\nu+1}\| \leq \rho\|\mathbf{y}_\perp^\nu\| + 2L_{\mathrm{mx}}\rho\|\mathbf{x}_\perp^\nu\| + \alpha L_{\mathrm{mx}}\rho\|\mathbf{d}^\nu\|, \tag{46b}$$

with ρ and L_{mx} defined in (26) *and* (4)*, respectively.*

Proof. We prove next (46b); (46a) follows readily from (44). Using (43a), (45), and the Lipschitz continuity of ∇f_i [cf. (1)], we can bound $\|\mathbf{y}_\perp^{\nu+1}\|$ as

$$\begin{aligned}\|\mathbf{y}_\perp^{\nu+1}\| &\leq \rho\|\mathbf{y}_\perp^\nu\| + \rho\|\nabla\mathbf{f}^{\nu+1} - \nabla\mathbf{f}^\nu\| \\ &\leq \rho\|\mathbf{y}_\perp^\nu\| + L_{\mathrm{mx}}\rho\|\underbrace{(\widehat{\mathbf{W}} - \mathbf{I})\mathbf{x}^\nu}_{=(\widehat{\mathbf{W}}-\mathbf{I})\mathbf{x}_\perp^\nu} + \alpha\widehat{\mathbf{W}}\mathbf{d}^\nu\| \\ &\leq \rho\|\mathbf{y}_\perp^\nu\| + 2L_{\mathrm{mx}}\rho\|\mathbf{x}_\perp^\nu\| + \alpha L_{\mathrm{mx}}\rho\|\mathbf{d}^\nu\|,\end{aligned}$$

where in the last inequality we used $\|\mathbf{W}\| \leq 1$. □

3.3.3. Step 3: $\|\mathbf{d}^\nu\| = \mathcal{O}(\sqrt{p^\nu} + \|\mathbf{y}_\perp^\nu\|)$ (closing the loop)

Given the inequalities in Propositions 3.4 and 3.5, to close the loop, one needs to link $\|\mathbf{d}^\nu\|$ to the quantities in the aforementioned inequalities, which is done next.

Proposition 3.6. *The following upper bound holds for $\|\mathbf{d}^\nu\|$:*

$$\|\mathbf{d}^\nu\|^2 \leq \frac{6}{\mu}\left(\left(\frac{D_{\mathrm{mx}}}{\widetilde{\mu}_{\mathrm{mn}}} + 1\right)^2 + \frac{4L_{\mathrm{mx}}^2}{\widetilde{\mu}_{\mathrm{mn}}^2}\right)p^\nu + \frac{3}{\widetilde{\mu}_{\mathrm{mn}}^2}\|\mathbf{y}_\perp^\nu\|^2. \tag{47}$$

where L_{mx} and $\widetilde{L}_{\mathrm{mx}}$, $\widetilde{\mu}_{\mathrm{mn}}$, D_{mx} are defined in (4) *and* (24)*, respectively.*

Proof. By optimality of $\widehat{\mathbf{x}}_i^\nu$ and $\mathbf{x}^\star$ we have

$$\begin{aligned}\left(\nabla\widetilde{f}_i(\widehat{\mathbf{x}}_i^\nu;\mathbf{x}_i^\nu) + \mathbf{y}_i^\nu - \nabla f_i(\mathbf{x}_i^\nu)\right)^\top(\mathbf{x}^\star - \widehat{\mathbf{x}}_i^\nu) + G(\mathbf{x}^\star) - G(\widehat{\mathbf{x}}_i^\nu) \geq 0, \\ \nabla F(\mathbf{x}^\star)^\top(\widehat{\mathbf{x}}_i^\nu - \mathbf{x}^\star) + G(\widehat{\mathbf{x}}_i^\nu) - G(\mathbf{x}^\star) \geq 0.\end{aligned}$$

Summing the two inequalities above yields

$$\begin{aligned}0 &\leq \left(\nabla F(\mathbf{x}^\star) - \mathbf{y}_i^\nu + \nabla f_i(\mathbf{x}_i^\nu) - \nabla\widetilde{f}_i(\widehat{\mathbf{x}}_i^\nu;\mathbf{x}_i^\nu) \pm \bar{\mathbf{y}}^\nu\right)^\top(\widehat{\mathbf{x}}_i^\nu - \mathbf{x}^\star) \\ &\leq \left(\nabla F(\mathbf{x}^\star) - \frac{1}{m}\sum_{j=1}^m \nabla f_j(\mathbf{x}_j^\nu) + \nabla f_i(\mathbf{x}_i^\nu) - \nabla\widetilde{f}_i(\widehat{\mathbf{x}}_i^\nu;\mathbf{x}_i^\nu)\right)^\top(\widehat{\mathbf{x}}_i^\nu - \mathbf{x}^\star) \\ &\quad + \|\bar{\mathbf{y}}^\nu - \mathbf{y}_i^\nu\|\|\widehat{\mathbf{x}}_i^\nu - \mathbf{x}^\star\| \\ &\leq \left(\nabla F(\mathbf{x}^\star) - \nabla F(\mathbf{x}_i^\nu) + \nabla f_i(\mathbf{x}_i^\nu) - \nabla\widetilde{f}_i(\widehat{\mathbf{x}}_i^\nu;\mathbf{x}_i^\nu)\right)^\top(\widehat{\mathbf{x}}_i^\nu - \mathbf{x}^\star) \\ &\quad + \|\bar{\mathbf{y}}^\nu - \mathbf{y}_i^\nu\|\|\widehat{\mathbf{x}}_i^\nu - \mathbf{x}^\star\| + \left\|\nabla F(\mathbf{x}_i^\nu) - \frac{1}{m}\sum_{j=1}^m \nabla f_j(\mathbf{x}_j^\nu)\right\|\|\widehat{\mathbf{x}}_i^\nu - \mathbf{x}^\star\|\end{aligned}$$

$$\begin{aligned}
&\leq \Big(\nabla F(\mathbf{x}^\star) - \nabla F(\mathbf{x}_i^\nu) + \nabla f_i(\mathbf{x}_i^\nu) \pm \nabla \widetilde{f}_i(\mathbf{x}^\star;\mathbf{x}_i^\nu) - \nabla \widetilde{f}_i(\widehat{\mathbf{x}}_i^\nu;\mathbf{x}_i^\nu)\Big)^\top (\widehat{\mathbf{x}}_i^\nu - \mathbf{x}^\star) \\
&\quad + \|\bar{\mathbf{y}}^\nu - \mathbf{y}_i^\nu\| \|\widehat{\mathbf{x}}_i^\nu - \mathbf{x}^\star\| + \left(\frac{1}{m}\sum_{j=1}^m L_j \left\|\mathbf{x}_i^\nu - \mathbf{x}_j^\nu\right\|\right) \|\widehat{\mathbf{x}}_i^\nu - \mathbf{x}^\star\| \\
&\leq \left(\int_0^1 \Big(\nabla^2 F(\theta\mathbf{x}^\star + (1-\theta)\mathbf{x}_i^\nu) - \nabla^2 \widetilde{f}_i\big(\theta\mathbf{x}^\star + (1-\theta)\mathbf{x}_i^\nu;\mathbf{x}_i^\nu\big)\Big)(\mathbf{x}^\star - \mathbf{x}_i^\nu)\,\mathrm{d}\theta\right)^\top (\widehat{\mathbf{x}}_i^\nu - \mathbf{x}^\star) \\
&\quad - \widetilde{\mu}_i \|\widehat{\mathbf{x}}_i^\nu - \mathbf{x}^\star\|^2 + \|\bar{\mathbf{y}}^\nu - \mathbf{y}_i^\nu\| \|\widehat{\mathbf{x}}_i^\nu - \mathbf{x}^\star\| + \left(\frac{1}{m}\sum_{j=1}^m L_j \left\|\mathbf{x}_i^\nu - \mathbf{x}_j^\nu\right\|\right) \|\widehat{\mathbf{x}}_i^\nu - \mathbf{x}^\star\|
\end{aligned}$$

$$\begin{aligned}
&\leq D_i \|\mathbf{x}^\star - \mathbf{x}_i^\nu\| \, \|\widehat{\mathbf{x}}_i^\nu - \mathbf{x}^\star\| - \widetilde{\mu}_i \|\widehat{\mathbf{x}}_i^\nu - \mathbf{x}^\star\|^2 + \|\bar{\mathbf{y}}^\nu - \mathbf{y}_i^\nu\| \|\widehat{\mathbf{x}}_i^\nu - \mathbf{x}^\star\| \\
&\quad + \left(\frac{1}{m}\sum_{j=1}^m L_j \left\|\mathbf{x}_i^\nu - \mathbf{x}_j^\nu\right\|\right) \|\widehat{\mathbf{x}}_i^\nu - \mathbf{x}^\star\|.
\end{aligned}$$

Rearranging terms and using the reverse triangle inequality we obtain the following bound for $\|\mathbf{d}_i^\nu\|$:

$$\begin{aligned}
&D_i \|\mathbf{x}^\star - \mathbf{x}_i^\nu\| + \|\bar{\mathbf{y}}^\nu - \mathbf{y}_i^\nu\| + \left(\frac{1}{m}\sum_{j=1}^m L_j \left\|\mathbf{x}_i^\nu - \mathbf{x}_j^\nu\right\|\right) \\
&\qquad\qquad \geq \widetilde{\mu}_i \|\widehat{\mathbf{x}}_i^\nu - \mathbf{x}^\star\| \geq \widetilde{\mu}_i \left(\|\mathbf{d}_i^\nu\| - \|\mathbf{x}^\star - \mathbf{x}_i^\nu\|\right). \qquad (48)
\end{aligned}$$

Therefore,

$$\begin{aligned}
\|\mathbf{d}_i^\nu\|^2 \leq\ & 3\left(\frac{D_i}{\widetilde{\mu}_i} + 1\right)^2 \|\mathbf{x}^\star - \mathbf{x}_i^\nu\|^2 + \frac{3}{\widetilde{\mu}_i^2}\|\bar{\mathbf{y}}^\nu - \mathbf{y}_i^\nu\|^2 + \frac{3}{\widetilde{\mu}_i^2}\left(\frac{1}{m}\sum_{j=1}^m L_j \left\|\mathbf{x}_i^\nu - \mathbf{x}_j^\nu\right\|\right)^2 \\
\leq\ & 3\left(\frac{D_i}{\widetilde{\mu}_i} + 1\right)^2 \|\mathbf{x}^\star - \mathbf{x}_i^\nu\|^2 + \frac{3}{\widetilde{\mu}_i^2}\|\bar{\mathbf{y}}^\nu - \mathbf{y}_i^\nu\|^2 + \frac{6L_{\mathrm{mx}}^2}{\widetilde{\mu}_i^2 m}\left(\sum_{j=1}^m \|\mathbf{x}_j^\nu - \mathbf{x}^\star\|^2 + m\|\mathbf{x}_i^\nu - \mathbf{x}^\star\|^2\right).
\end{aligned}$$

Summing over $i = 1, \ldots, m$, yields

$$\begin{aligned}
\|\mathbf{d}^\nu\|^2 \leq\ & \left(3\left(\frac{D_{\mathrm{mx}}}{\widetilde{\mu}_{\mathrm{mn}}} + 1\right)^2 + \frac{12L_{\mathrm{mx}}^2}{\widetilde{\mu}_{\mathrm{mn}}^2}\right)\sum_{j=1}^m \left\|\mathbf{x}_j^\nu - \mathbf{x}^\star\right\|^2 + \frac{3}{\widetilde{\mu}_{\mathrm{mn}}^2}\|\mathbf{y}_\perp^\nu\|^2 \\
\leq\ & \frac{6}{\mu}\left(\left(\frac{D_{\mathrm{mx}}}{\widetilde{\mu}_{\mathrm{mn}}} + 1\right)^2 + \frac{4L_{\mathrm{mx}}^2}{\widetilde{\mu}_{\mathrm{mn}}^2}\right)p^\nu + \frac{3}{\widetilde{\mu}_{\mathrm{mn}}^2}\|\mathbf{y}_\perp^\nu\|^2.
\end{aligned}$$

□

3.3.4. Step 4: Proof of the linear rate (chaining the inequalities)

We are now ready to prove linear rate of the SONATA algorithm. We build on the following intermediate result, introduced in [21].

Lemma 3.7. *Given the sequence $\{s^\nu\}$, define the transformations*

$$S^K(z) \triangleq \max_{\nu=0,\ldots,K} |s^\nu| z^{-\nu} \quad \text{and} \quad S(z) \triangleq \sup_{\nu\in\mathbb{N}} |s^\nu| z^{-\nu}, \tag{49}$$

for $z\in(0,1)$. If $S(z)$ is bounded, then $|s^\nu| = \mathcal{O}(z^\nu)$.

We show next how to chain the inequalities (40), (46) and (47) so that Lemma 3.7 can be applied to the sequences $\{p^\nu\}$, $\{\|\mathbf{x}_\perp^\nu\|^2\}$, $\{\|\mathbf{y}_\perp^\nu\|^2\}$ and $\{\|\mathbf{d}^\nu\|^2\}$, establishing thus their linear convergence.

Proposition 3.8. *Let $P^K(z)$, $X_\perp^K(z)$, $Y_\perp^K(z)$ and $D^K(z)$ denote the transformation (49) applied to the sequences $\{p^\nu\}$, $\{\|\mathbf{x}_\perp^\nu\|^2\}$, $\{\|\mathbf{y}_\perp^\nu\|^2\}$ and $\{\|\mathbf{d}^\nu\|^2\}$, respectively. Given the constants $\sigma(\alpha)$ and $\eta(\alpha)$ (defined in Proposition 3.4) and the free parameters $\epsilon_x, \epsilon_y > 0$ (to be determined), the following hold*

$$P^K(z) \leq G_P(\alpha, z) \cdot \left(4L_{\text{mx}}^2 X_\perp^K(z) + 2Y_\perp^K(z)\right) + \omega_p, \tag{50a}$$

$$X_\perp^K(z) \leq G_X(z) \cdot \rho^2\alpha^2 D^K(z) + \omega_x, \tag{50b}$$

$$Y_\perp^K(z) \leq G_Y(z) \cdot 8L_{\text{mx}}^2 \rho^2 X_\perp^K(z) + G_Y(z) \cdot 2L_{\text{mx}}^2 \rho^2\alpha^2 D^K(z) + \omega_y, \tag{50c}$$

$$D^K(z) \leq C_1 \cdot P^K(z) + C_2 \cdot Y_\perp^K(z), \tag{50d}$$

for all

$$z \in \left(\max\{\sigma(\alpha), \rho^2(1+\epsilon_x), \rho^2(1+\epsilon_y)\}, 1\right), \tag{51}$$

where

$$G_P(\alpha, z) \triangleq \frac{\eta(\alpha)}{z - \sigma(\alpha)}, \qquad \omega_p \triangleq \frac{z}{z-\sigma(\alpha)} \cdot p^0 \tag{52a}$$

$$G_X(z) \triangleq \frac{(1+\epsilon_x^{-1})}{z - \rho^2(1+\epsilon_x)}, \qquad \omega_x \triangleq \frac{z}{z-\rho^2(1+\epsilon_x)} \cdot \|\mathbf{x}_\perp^0\|^2, \tag{52b}$$

$$G_Y(z) \triangleq \frac{(1+\epsilon_y^{-1})}{z - \rho^2(1+\epsilon_y)}, \qquad \omega_y \triangleq \frac{z}{z-\rho^2(1+\epsilon_y)} \cdot \|\mathbf{y}_\perp^0\|^2, \tag{52c}$$

$$C_1 \triangleq \frac{6}{\mu}\left(\left(\frac{D_{\text{mx}}}{\widetilde{\mu}_{\text{mn}}} + 1\right)^2 + \frac{4L_{\text{mx}}^2}{\widetilde{\mu}_{\text{mn}}^2}\right), \qquad C_2 \triangleq \frac{4}{\widetilde{\mu}_{\text{mn}}^2}. \tag{52d}$$

Proof. Squaring (46) and using Young's inequality yield

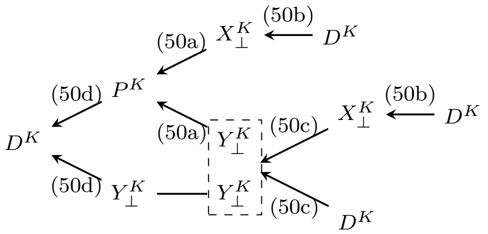
$$\begin{aligned} \|\mathbf{x}_\perp^{\nu+1}\|^2 &\leq \rho^2(1+\epsilon_x)\|\mathbf{x}_\perp^\nu\|^2 + \rho^2(1+\epsilon_x^{-1})\alpha^2\|\mathbf{d}^\nu\|^2 \\ \|\mathbf{y}_\perp^{\nu+1}\|^2 &\leq \rho^2(1+\epsilon_y)\|\mathbf{y}_\perp^\nu\|^2 + \rho^2(1+\epsilon_y^{-1})\Big(8L_{\text{mx}}^2\|\mathbf{x}_\perp^\nu\|^2 + 2\alpha^2 L_{\text{mx}}^2\|\mathbf{d}^\nu\|^2\Big), \end{aligned} \tag{53}$$

for arbitrary $\epsilon_x, \epsilon_y > 0$. The proof is completed by taking the maximum of both sides of (40), (47), and (53) over $\nu = 0, \ldots, K$ and using $\max_{\nu=0,\ldots,K} |s^{\nu+1}| z^{-\nu} \geq z \cdot \max_{\nu=0,\ldots,K} |s^\nu| z^{-\nu} - z \cdot |s^0|$, for any sequence $\{s^\nu\}$ and $z \in (0,1)$. □

Chaining the inequalities in Proposition 3.8 in the way shown in Fig. 3.3.4, we can bound $D^K(z)$ as (see Appendix A for the proof)

$$D^K(z) \leq \mathcal{P}(\alpha, z) \cdot D^K(z) + \mathcal{R}(\alpha, z), \tag{54}$$

Figure 1. Chain of the inequalities in Proposition 3.8 leading to (54).



where $\mathcal{P}(\alpha, z)$ is defined as

$$\begin{aligned}\mathcal{P}(\alpha, z) \triangleq\ & G_P(\alpha, z) \cdot G_X(z) \cdot C_1 \cdot 4L_{\text{mx}}^2 \cdot \rho^2 \cdot \alpha^2 \\ & + (G_P(\alpha, z) \cdot 2C_1 + C_2) \cdot G_Y(z) \cdot 2L_{\text{mx}}^2 \rho^2 \cdot \alpha^2 \\ & + (G_P(\alpha, z) \cdot 2C_1 + C_2) \cdot G_Y(z) \cdot 8L_{\text{mx}}^2 \rho^2 \cdot G_X(z) \cdot \rho^2 \cdot \alpha^2,\end{aligned} \tag{55}$$

and $\mathcal{R}(\alpha, z)$ is a remainder, which is bounded under (51).

Therefore, as long as $\mathcal{P}(\alpha, z) < 1$, (54) implies

$$D^K(z) \leq \frac{\mathcal{R}(\alpha, z)}{1 - \mathcal{P}(\alpha, z)} \leq B < +\infty \tag{56}$$

where B is a constant independent of K. Therefore, $D(z) \leq B$ and thus $\{\|\mathbf{d}^\nu\|^2\}$ converges R-linearly to zero at rate at least z (cf. Lemma 3.7). Applying the same argument to the other inequalities in Proposition 3.8, one can conclude that also the sequences $\{p^\nu\}$, $\{\|\mathbf{x}_\perp^\nu\|^2\}$ and $\{\|\mathbf{y}_\perp^\nu\|\}$ converge R-linearly to zero.

The last step consists to showing that there exist a sufficiently small step-size $\alpha \in (0, 1]$ and $z \in (0, 1)$ satisfying (51), such that $\mathcal{P}(\alpha, z) < 1$. This is proved in the Theorem 3.9 below.

Theorem 3.9. *Consider Problem* (P) *under Assumptions A-B; and the SONA- TA algorithm* (11a)-(11d)*, under Assumptions C and D, with $\widetilde{\mu}_{\text{mn}} \geq D_{\text{mn}}^\ell$. Then, there exists a sufficiently small step-size $\bar{\alpha} \in (0, 1]$ [see the proof for its expression] such that for all $\alpha < \bar{\alpha}$, $\{U(\mathbf{x}_i^\nu)\}$ converges to $U^\star$ at an R-linear rate, $i \in [m]$.*

Proof. The proof is organized in following two steps: **Step 1)** We first consider the "marginal" stable case by letting $z = 1$, and show that there exists $\bar{\alpha} > 0$ so that $\mathcal{P}(\alpha, 1) < 1$, for all $\alpha \in (0, \bar{\alpha})$; **Step 2)** Then, invoking the continuity of $\mathcal{P}(\alpha, z)$, we argue that, for any $\alpha \in (0, \bar{\alpha})$, one can find $\bar{z}(\alpha) < 1$ such that $\mathcal{P}\big(\alpha, \bar{z}(\alpha)\big) < 1$. This implies the boundedness of $D^K\big(\bar{z}(\alpha)\big)$, and thus $\|\mathbf{d}^\nu\|^2 = \mathcal{O}\big(\bar{z}(\alpha)^\nu\big)$ (cf. Lemma 3.7).

• **Step 1:** We begin optimizing the free parameters ϵ_x, ϵ_y, and ϵ_{opt}. Since the goal is to find the largest $\bar{\alpha}$ so that $\mathcal{P}(\alpha, 1) < 1$, for all $\alpha \in (0, \bar{\alpha})$, the optimal choice of ϵ_x, ϵ_y, and ϵ_{opt} is the one that minimizes $\mathcal{P}(\alpha, 1)$, that is,

$$\epsilon^\star = \underset{\epsilon > 0}{\text{argmin}} \frac{1 + \epsilon^{-1}}{1 - \rho^2(1 + \epsilon)} = \frac{1 - \rho}{\rho}. \tag{57}$$

We then set $\epsilon_x = \epsilon_y = \epsilon^\star$, and proceed to optimize ϵ_{opt}, which appears in $\eta(\alpha)$ and $\sigma(\alpha)$. Recalling the definition of $\eta(\alpha)$ and $\sigma(\alpha)$ (cf. Proposition 3.4) and the constraint (34), the problem boils down to minimize

$$G_P(\alpha,1) = \frac{\eta(\alpha)}{1-\sigma(\alpha)} = \frac{\frac{1}{2}\epsilon_{opt}^{-1}\cdot \frac{D_{\text{mx}}^2}{\mu} + \frac{1}{\mu}\cdot\left(\left(1-\frac{\alpha}{2}\right)\widetilde{\mu}_{\text{mn}} + \frac{D_{\text{mn}}^{\ell}}{2}\alpha - \frac{1}{2}\epsilon_{opt}\right)}{\left(1-\frac{\alpha}{2}\right)\widetilde{\mu}_{\text{mn}} + \frac{D_{\text{mn}}^{\ell}}{2}\alpha - \frac{1}{2}\epsilon_{opt}},$$

subject to $\epsilon_{opt} \in (0, 2\widetilde{\mu}_{\text{mn}} - \alpha(\widetilde{\mu}_{\text{mn}} - D_{\text{mn}}^{\ell}))$. To have a nonempty feasible set, we require $\alpha < 2\widetilde{\mu}_{\text{mn}}/(\widetilde{\mu}_{\text{mn}} - D_{\text{mn}}^{\ell})$ (recall that it is assumed $\widetilde{\mu}_{\text{mn}} \geq D_{\text{mn}}^{\ell}$). Setting the derivative of $G_P(\alpha,1)$ with respect to ϵ_{opt} to zero, yields $\epsilon_{opt}^{\star} = \left(1-\frac{\alpha}{2}\right)\widetilde{\mu}_{\text{mn}} + \alpha D_{\text{mn}}^{\ell}/2$, which is strictly feasible, and thus the solution.

Let $\mathcal{P}^{\star}(\alpha,z)$ denote the value of $\mathcal{P}(\alpha,z)$ corresponding to the optimal choice of the above parameters. The expression of $\mathcal{P}^{\star}(\alpha,1)$ reads

$$\begin{aligned}\mathcal{P}^{\star}(\alpha,1) \triangleq\, & G_P^{\star}(\alpha)\cdot C_1 \cdot 4L_{\text{mx}}^2 \cdot \frac{\rho^2}{(1-\rho)^2}\cdot\alpha^2 \\ & + \left(G_P^{\star}(\alpha)\cdot 2C_1 + C_2\right)\cdot 2L_{\text{mx}}^2\cdot\frac{\rho^2}{(1-\rho)^2}\cdot\alpha^2 \\ & + \left(G_P^{\star}(\alpha)\cdot 2C_1 + C_2\right)\cdot 8L_{\text{mx}}^2\cdot\frac{\rho^4}{(1-\rho)^4}\cdot\alpha^2,\end{aligned} \tag{58}$$

where

$$G_P^{\star}(\alpha) \triangleq \frac{\frac{D_{\text{mx}}^2}{\mu} + \frac{1}{\mu}\cdot\left(\left(1-\frac{\alpha}{2}\right)\widetilde{\mu}_{\text{mn}} + \frac{D_{\text{mn}}^{\ell}}{2}\alpha\right)^2}{\left(\left(1-\frac{\alpha}{2}\right)\widetilde{\mu}_{\text{mn}} + \frac{D_{\text{mn}}^{\ell}}{2}\alpha\right)^2}. \tag{59}$$

• **Step 2:** Since $\mathcal{P}^{\star}(\bullet,1)$ is continuous and monotonically increasing on $(0, 2\widetilde{\mu}_{\text{mn}}/(\widetilde{\mu}_{\text{mn}} - D_{\text{mn}}^{\ell})$, with $\mathcal{P}^{\star}(0,1) = 0$, there exists some $\bar{\alpha} < 2\widetilde{\mu}_{\text{mn}}/(\widetilde{\mu}_{\text{mn}} - D_{\text{mn}}^{\ell})$ such that $\mathcal{P}^{\star}(\alpha,1) < 1$, for all $\alpha \in (0,\bar{\alpha})$. One can verify that, for any $\alpha \in (0, 2\widetilde{\mu}_{\text{mn}}/(\widetilde{\mu}_{\text{mn}} - D_{\text{mn}}^{\ell}))$, $\mathcal{P}^{\star}(\alpha,z)$ is continuous at $z=1$. Therefore, for any fixed $\alpha \in (0,\bar{\alpha})$, $\mathcal{P}^{\star}(\alpha,1) < 1$ implies the existence of some $\bar{z}(\alpha) < 1$ such that $\mathcal{P}^{\star}(\alpha,\bar{z}(\alpha)) < 1$.

We conclude the proof providing the expression of a valid $\bar{\alpha}$. Restricting $\alpha \leq \widetilde{\mu}_{\text{mn}}/(\widetilde{\mu}_{\text{mn}} - D_{\text{mn}}^{\ell})$, we upper bound $G_P^{\star}(\alpha)$ by $G_P^{\star}(\widetilde{\mu}_{\text{mn}}/(\widetilde{\mu}_{\text{mn}} - D_{\text{mn}}^{\ell}))$. Using for $G_P^{\star}(\alpha)$ this upper bound in (58) and solving the resulting $\mathcal{P}^{\star}(\alpha,1) < 1$ for α, yield

$$\begin{aligned}\alpha < \alpha_1 \triangleq \Bigg(& G_P^{\star}\left(\frac{\widetilde{\mu}_{\text{mn}}}{\widetilde{\mu}_{\text{mn}} - D_{\text{mn}}^{\ell}}\right)\cdot C_1\cdot 4L_{\text{mx}}^2\cdot\frac{\rho^2}{(1-\rho)^2} \\ & + \left(G_P^{\star}\left(\frac{\widetilde{\mu}_{\text{mn}}}{\widetilde{\mu}_{\text{mn}} - D_{\text{mn}}^{\ell}}\right)\cdot 2C_1 + C_2\right)\cdot 2L_{\text{mx}}^2\cdot\frac{\rho^2}{(1-\rho)^2} \\ & + \left(G_P^{\star}\left(\frac{\widetilde{\mu}_{\text{mn}}}{\widetilde{\mu}_{\text{mn}} - D_{\text{mn}}^{\ell}}\right)\cdot 2C_1 + C_2\right)\cdot 8L_{\text{mx}}^2\cdot\frac{\rho^4}{(1-\rho)^4}\Bigg)^{-\frac{1}{2}}.\end{aligned} \tag{60}$$

Therefore, a valid $\bar{\alpha}$ is $\bar{\alpha} = \min\{\widetilde{\mu}_{\text{mn}}/(\widetilde{\mu}_{\text{mn}} - D_{\text{mn}}^{\ell}), \alpha_1\}$. □

The next theorem provides an explicit expression of the convergence rate in Theorem 3.9 in terms of the step-size α; the constants J, $A_{\frac{1}{2}}$, and α^* therein are defined in (B7), (B5) with $\theta = 1/2$, and (B9), respectively.

Theorem 3.10. *In the setting of Theorem 3.9, suppose that the step-size α satisfies $\alpha \in (0,\alpha_{\text{mx}})$, with $\alpha_{\text{mx}} \triangleq \min\{(1-\rho)^2/A_{\frac{1}{2}}, \widetilde{\mu}_{\text{mn}}/(\widetilde{\mu}_{\text{mn}} - D_{\text{mn}}), 1\}$. Then, $U(\mathbf{x}_i^{\nu}) - U^{\star} = \mathcal{O}(z^{\nu})$, for all $i \in [m]$, where*

$$z = \begin{cases} 1 - J \cdot \alpha & \textit{for } \alpha \in (0, \min\{\alpha^*, \alpha_{\text{mx}}\})\,, \\ \left(\rho + \sqrt{\alpha A_{\frac{1}{2}}}\right)^2 & \textit{for } \alpha \in [\min\{\alpha^*, \alpha_{\text{mx}}\}, \alpha_{\text{mx}})\,. \end{cases} \tag{61}$$

Proof. See Appendix B. □

3.4. *Discussion*

Theorem 3.10 provides a unified set of convergence conditions for different choices of surrogates and network topologies. To shed light on the expression of the rate and its dependence on the key optimization and network parameters, we customize here Theorem 3.10 to specific network topologies and surrogate functions. We begin considering star-networks (cf. Sec. 3.4.1) and then move to general graph topologies with no master node (cf. Sec. 3.4.2). We will customize the rate achieved by SONATA employing the following two surrogate functions $\widetilde{f}_i$, representing the two extreme choices in the spectrum of admissible surrogates:

- **Linearization:**

$$\widetilde{f}_i(\mathbf{x}_i; \mathbf{x}_i^\nu) \triangleq \nabla f_i(\mathbf{x}_i^\nu)^\top (\mathbf{x}_i - \mathbf{x}_i^\nu) + \frac{L}{2}\|\mathbf{x}_i - \mathbf{x}_i^\nu\|^2; \tag{62}$$

- **Local f_i:**

$$\widetilde{f}_i(\mathbf{x}_i; \mathbf{x}_i^\nu) \triangleq f_i(\mathbf{x}_i) + \frac{\beta}{2}\|\mathbf{x}_i - \mathbf{x}_i^\nu\|^2. \tag{63}$$

3.4.1. Star-networks: SONATA-Star

Convergence of SONATA-Star (Algorithm 2) is established in Corollary 3.11 below.

Corollary 3.11. *Consider Problem* (P) *under Assumption A over a star-network; let* $\{\mathbf{x}^\nu\}$ *be the sequence generated by SONATA-Star (Algorithm 2), based on the surrogate functions satisfying Assumption C and step-size* $\alpha \in (0, \min(2\widetilde{\mu}_{\text{mn}}/(\widetilde{\mu}_{\text{mn}} - D^\ell_{\text{mn}}), 1)]$*. Then, for all* $i = 1, \ldots, m$*,*

$$U(\mathbf{x}^\nu) - U^\star = \mathcal{O}(z^\nu), \quad \textit{with} \quad z = 1 - \alpha \cdot \frac{\left(1 - \frac{\alpha}{2}\right)\widetilde{\mu}_{\text{mn}} + \frac{\alpha D^\ell_{\text{mn}}}{2}}{\frac{D^2_{\text{mx}}}{2\mu} + \left(1 - \frac{\alpha}{2}\right)\widetilde{\mu}_{\text{mn}} + \frac{\alpha D^\ell_{\text{mn}}}{2}}. \tag{64}$$

In particular, when the surrogates (62) *and* (63) *are employed along with* $\alpha = 1$*, the rate above reduces to the following expressions:*

- ***Linearization*** (62)*:* $z \leq 1 - \kappa_g^{-1}$*. Therefore,* $U(\mathbf{x}^\nu) - U^\star \leq \epsilon$ *in at most* $\mathcal{O}\Big(\kappa_g \log(1/\epsilon)\Big)$ *iterations (communications);*
- ***Local*** f_i (63)*:*

$$z \leq 1 - \frac{1}{1 + 4 \cdot \frac{\beta}{\mu} \cdot \min\{1, \frac{\beta}{\mu}\}}. \tag{65}$$

Therefore, $U(\mathbf{x}^\nu) - U^\star \leq \epsilon$ *in at most*

$$\begin{cases} \mathcal{O}\left(1 \cdot \log\left(1/\epsilon\right)\right), & \textit{if } \beta \leq \mu, \\ \mathcal{O}\left(\frac{\beta}{\mu} \cdot \log\left(1/\epsilon\right)\right), & \textit{if } \beta > \mu, \end{cases} \tag{66}$$

iterations (communications).

Proof. See Appendix C. □

The following comments are in order. When linearization is employed, SONATA-Star matches the iteration complexity of the centralized proximal-gradient algorithm. When the f_i's are sufficiently similar, (65)-(66) proves that faster rates can be achieved if surrogates (63) are chosen over first-order approximations: when $\beta \ll L$, (66) is significantly faster than $\mathcal{O}\big(\kappa_g \log(1/\epsilon)\big)$. As case study, consider Example 2 (cf. Sec. 2.1.2): plugging (10) into Corollary 3.11 shows that using the surrogates (63) yields $\widetilde{\mathcal{O}}\big(L\sqrt{d\,m} \cdot \log(1/\epsilon)\big)$ iterations (communications); this contrasts with $\widetilde{\mathcal{O}}\big(L\sqrt{d\,m\,n} \cdot \log(1/\epsilon)\big)$, achieved by first-order methods (and SONATA-Star using linearization), which instead increases with the sample size n.

Comparison with DANE & CEASE Since SONATA-Star contains as special cases the DANE [37] and CEASE [7] algorithms, we contrast here Corollary 3.11 with their convergence rates. We recall that DANE is applicable to (P) when $G = 0$: For quadratic losses, it achieves an ϵ-optimal objective value in $\mathcal{O}\big((\beta/\mu)^2 \cdot \log(1/\epsilon)\big)$ iterations/communications (here $\beta/\mu \geq 1$). This rate is worse than (66). For nonquadratic losses, [37] did not show any rate improvement of DANE over plain gradient algorithms, i.e., $\mathcal{O}\big(\kappa_g \cdot \log(1/\epsilon)\big)$ while SONATA-star still retains $\mathcal{O}\big(\beta/\mu \cdot \log(1/\epsilon)\big)$. The CEASE algorithm is proved to achieve an ϵ-solution on the iterates in $\mathcal{O}\big((\beta/\mu)^2 \cdot \log(1/\epsilon)\big)$ iterations/communications (with $\beta/\mu \geq 1$); SONATA reaches the same error on the iterates in $\mathcal{O}\big(\beta/\mu \cdot \log(\kappa_g/\epsilon)\big)$ iterations/communications, which matches the order of the mirror-decent algorithm.

In the next section we extend the study to networks with no centralized nodes, sheding lights on the role of the network in achieving the same kind of results.

3.4.2. The general case

The convergence rate of SONATA over general graphs is summarized in Corollary 3.12 for the linearization surrogates (62) while Corollaries 3.13 and 3.14 consider the surrogates (63) based on local f_i, with Corollary 3.13 addressing the case $\beta \leq \mu$ and Corollary 3.14 the case $\beta > \mu$. The step-size α is tuned to obtain favorable rate expressions.

Corollary 3.12 (Linearization surrogates)**.** *In the setting of Theorem 3.10, let* $\{\mathbf{x}^\nu\}$ *be the sequence generated by SONATA, using the surrogates* (62) *and step-size* $\alpha = c \cdot \alpha_{\text{mx}}$, $c \in (0,1)$, *with* $\alpha_{\text{mx}} = \min\{1, (1-\rho)^2/(\rho \cdot 110\kappa_g(1+\beta/L)^2)\}$*. The number of iterations (communications) needed for* $U(\mathbf{x}_i^\nu) - U^\star \leq \epsilon$, $i \in [m]$, *is*

$$\textbf{Case I:} \qquad \mathcal{O}\left(\kappa_g \log(1/\epsilon)\right), \qquad \textit{if} \quad \frac{\rho}{(1-\rho)^2} \leq \frac{1}{110\,\kappa_g \left(1 + \frac{\beta}{L}\right)^2}, \tag{67}$$

$$\textbf{Case II:} \qquad \mathcal{O}\left(\frac{\left(\kappa_g + \beta/\mu\right)^2 \rho}{(1-\rho)^2}\log(1/\epsilon)\right), \qquad \textit{otherwise.} \tag{68}$$

Proof. See Appendix D. □

Corollary 3.13 (local f_i, $\beta \leq \mu$)**.** *Instate assumptions of Theorem 3.10 and suppose $\beta \leq \mu$. Consider SONATA using the surrogates* (63) *and step-size $\alpha = c \cdot \alpha_{\rm mx}$, $c \in (0,1)$, with $\alpha_{\rm mx} = \min\{1, (1-\rho)^2/(M\rho)\}$ and $M = 193\left(1+\frac{\beta}{\mu}\right)^2\left(\kappa_g + \frac{\beta}{\mu}\right)^2$. The number of iterations (communications) needed for $U(\mathbf{x}_i^\nu) - U^\star \leq \epsilon$, $i \in [m]$, is*

$$\textbf{Case I:} \qquad \mathcal{O}\left(1 \cdot \log(1/\epsilon)\right), \qquad \textit{if} \quad \frac{\rho}{(1-\rho)^2} \leq \frac{1}{193\left(1+\frac{\beta}{\mu}\right)^2\left(\kappa_g + \frac{\beta}{\mu}\right)^2}, \tag{69}$$

$$\textbf{Case II:} \qquad \mathcal{O}\left(\frac{\kappa_g^2\, \rho}{(1-\rho)^2}\log(1/\epsilon)\right), \qquad \textit{otherwise.} \tag{70}$$

Corollary 3.14 (local f_i, $\beta > \mu$)**.** *Instate assumptions of Theorem 3.10 and suppose $\beta > \mu$. Consider SONATA using the surrogates* (63) *and step-size $\alpha = c \cdot \alpha_{\rm mx}$, $c \in (0,1)$, with $\alpha_{\rm mx} = \min\{1, (1-\rho)^2/(M\rho)\}$ and $M = 253\left(1+\frac{L}{\beta}\right)\left(\kappa_g + \frac{\beta}{\mu}\right)$. The number of iterations (communications) needed for $U(\mathbf{x}_i^\nu) - U^\star \leq \epsilon$, $i \in [m]$, is*

$$\textbf{Case I:} \quad \mathcal{O}\left(\frac{\beta}{\mu} \cdot \log(1/\epsilon)\right) \qquad \textit{if} \quad \frac{\rho}{(1-\rho)^2} \leq \frac{1}{253\left(1+\frac{L}{\beta}\right)\left(\kappa_g + \frac{\beta}{\mu}\right)}, \tag{71}$$

$$\textbf{Case II:} \quad \mathcal{O}\left(\frac{\left(\kappa_g + (\beta/\mu)\right)^2 \rho}{(1-\rho)^2}\log(1/\epsilon)\right), \qquad \textit{otherwise .} \tag{72}$$

The proof of Corollaries 3.13 and 3.14 can be found in Appendix E.

Several comments are in order.

• **Order of the rate of centralized (nonaccelerated) methods (Case I):** For a fixed optimization problem, if the network is sufficiently connected (ρ “small”), its impact on the rate becomes negligible (the bottleneck is the optimization), and SONATA matches the *network-independent* rate order achieved on star-topologies (cf. Corollary 3.11) by the proximal gradient algorithm when linearization is employed [cf. (67)] and by the mirror-descent scheme when the local f_i's are used in the surrogates [cf. (69) and (71)].

• **Network-dependent rates (Case II):** As expected, the convergence rate deteriorates as ρ increases, i.e., the network connectivity gets worse. This translates in a less favorable dependence of the complexity on κ_g and β/μ (by a square factor) and network scalability of the order of $\rho/(1-\rho)^2$. When $\beta\sqrt{\rho} = \mathcal{O}(L)$ (e.g., the network is decently connected or $\beta = \mathcal{O}(L)$), the complexity becomes $\mathcal{O}\left(\kappa_g^2(1-\rho)^{-2}\log(1/\epsilon)\right)$, which compares favorably with that of existing distributed schemes, determined instead by the more pessimistic local quantities (3). The scalability of the rate with the network connectivity, $(1-\rho)^{-2}$, can be improved leveraging multiple rounds of communications or accelerated consensus protocols, as discussed below.

• **Linearization** (62) **vs. local** f_i (63) **surrogates:** As already observed in the setting of star-networks, the use of the local losses as surrogates employs a form of preconditioning in the local agents subproblems. When the f_i's are sufficiently similar to each other, so that $1+\beta/\mu < \kappa_g$, exploiting local Hessian information via (63) provably reduces the iteration/communication complexity over linear models (62)–contrast (67) with (69) and (71). Note that these faster rates are achieved without exchanging any matrices over the network, which is a key feature of SONATA. On the other hand, when the functions f_i are heterogeneous, the local surrogates (63) are no longer informative of the average-loss F and using linearization might yield better rates. Although these design recommendations are based on sufficient conditions, numerical results seem to confirm the above conclusions–see Sec. 5.

• **Multiple communications rounds and acceleration:** The discussion above shows that rates of the order of those of centralized methods can be achieved if the network is sufficiently connected (Case I). When this is not the case, one can still achieve the same iteration complexity at the cost of multiple, finite, rounds of communications per iteration. Specifically, let ρ_0 be the connectivity of the given network and suppose we run K steps of communications per iteration (computation) in (43a)-(43b); this yields an effective network with improved connectivity $\rho = \rho_0^K$. One can then choose K so that the ratio $\rho_0^K/(1-\rho_0^K)^2$ satisfies the condition triggering Case I in the Corollaries 3.12–3.14, as briefly summarized next.

1) Linearization: Invoking Corollary 3.12, one can check that the order of such a K is $K = \mathcal{O}(\log(\kappa_g(1+\beta/L)^2)/\log(1/\rho_0)) = \mathcal{O}(\log(\kappa_g(1+\beta/L)^2)/(1-\rho_0))$; therefore, SONATA using the surrogates (62) reaches an ϵ-solution in $\mathcal{O}\left(\kappa_g \log(1/\epsilon)\right)$ iterations and $\mathcal{O}\left(\kappa_g \cdot (1-\rho_0)^{-1} \log(\kappa_g(1+\beta/L)^2) \log(1/\epsilon)\right)$ communications. The dependence on the network connectivity ρ_0 can be further improved leveraging Chebyshev polynomials (see, e.g., [32, 45]): the final communication complexity of SONATA reads

$$\mathcal{O}\left(\frac{\kappa_g}{\sqrt{1-\rho_0}} \cdot \log\left(\kappa_g(1+\beta/L)^2\right)\, \log(1/\epsilon)\right).$$

2) Local f_i **surrogates:** Considering the case $\beta \geq \mu$ (Corollary 3.14), we can show that SONATA using the surrogates (63) and employing multiple rounds of communications per iteration, reaches an ϵ-solution in $\mathcal{O}\left(\beta/\mu \cdot \log(1/\epsilon)\right)$ iterations and $\mathcal{O}\left(\beta/\mu \cdot \log\left((\kappa_g + \beta/\mu)(1+L/\beta)\right)(1-\rho_0)^{-1} \log(1/\epsilon)\right)$ communications. If Chebyshev polynomials are used to accelerate the communications, the communication complexity further improves to

$$\mathcal{O}\left(\frac{\beta/\mu}{\sqrt{1-\rho_0}} \cdot \log\left((\kappa_g + \beta/\mu)(1+L/\beta)\right) \log(1/\epsilon)\right).$$

4. The SONATA algorithm over directed time-varying graphs

In this section we extend SONATA and its convergence analysis to solve Problem (P) over *directed, time-varying graphs* (Assumption B$'$). Note that (11a)-(11d) is not readily applicable to this setting, as constructing a doubly stochastic weight matrix compliant with a directed graph is generally infeasible or computationally costly–see e.g. [8]. Conditions on the weight matrices can be relaxed if the consensus/tracking schemes

(11c)-(11d) are properly changed to deal with the lack of doubly stochasticity.

Here, we consider the perturbed push-sum protocols as proposed in the companion paper [35] (but in the Adapt-Then-Combine (ATC) form). The resulting distributed algorithm, still termed SONATA, is formally described in Algorithm 3.

```
Algorithm 3 SONATA over time-varying directed graphs
Data: $\mathbf{x}_i^0 \in \mathcal{K}$, $\mathbf{y}_i^0 = \nabla f_i(\mathbf{x}_i^0)$, and $\phi_i^0 = 1$, $i \in [m]$.
Iterate: $\nu = 1, 2, ...$
[S.1] [Distributed Local Optimization] Each agent $i$ solves
        $\widehat{\mathbf{x}}_i^\nu \triangleq \underset{\mathbf{x}_i \in \mathcal{K}}{\operatorname{argmin}}\ \widetilde{f}_i(\mathbf{x}_i; \mathbf{x}_i^\nu) + \big(\mathbf{y}_i^\nu - \nabla f_i(\mathbf{x}_i^\nu)\big)^\top (\mathbf{x}_i - \mathbf{x}_i^\nu) + G(\mathbf{x}_i),$   (73a)
    and updates
        $\mathbf{x}_i^{\nu+\frac{1}{2}} = \mathbf{x}_i^\nu + \alpha \cdot \mathbf{d}_i^\nu, \quad \text{with} \quad \mathbf{d}_i^\nu \triangleq \widehat{\mathbf{x}}_i^\nu - \mathbf{x}_i^\nu;$   (73b)
[S.2] [Information Mixing] Each agent $i$ computes
    (a) Consensus
        $\phi_i^{\nu+1} = \sum_{j=1}^m c_{ij}^\nu \phi_j^\nu, \quad \mathbf{x}_i^{\nu+1} = \frac{1}{\phi_i^{\nu+1}} \sum_{j=1}^m c_{ij}^\nu \phi_j^\nu \mathbf{x}_j^{\nu+\frac{1}{2}},$   (73c)
    (b) Gradient tracking
        $\mathbf{y}_i^{\nu+1} = \frac{1}{\phi_i^{\nu+1}} \sum_{j=1}^m c_{ij}^\nu \left(\phi_j^\nu \mathbf{y}_j^\nu + \nabla f_j(\mathbf{x}_j^{\nu+1}) - \nabla f_j(\mathbf{x}_j^\nu)\right),$   (73d)
end
```

In the perturbed push-sum protocols (73c)-(73d), $\mathbf{C}^\nu \triangleq (c_{ij}^\nu)_{i,j=1}^m$ satisfies the assumption below.

Assumption E. *For each $\nu \geq 0$, the weight matrix $\mathbf{C}^\nu \triangleq (c_{ij}^\nu)_{i,j=1}^m$ has a sparsity pattern compliant with $\mathcal{G}^\nu$, i.e., there exists a constant c_ℓ such that, for all $\nu = 0, 1, \ldots,$*

E1 $c_{ii}^\nu \geq c_\ell > 0$, for all $i \in [m]$;
E2 $c_{ij}^\nu \geq c_\ell > 0$, if $(j, i) \in \mathcal{E}^\nu$; and $c_{ij}^\nu = 0$ otherwise.

Moreover, $\mathbf{C}^\nu$ is column stochastic, i.e., $\mathbf{1}^\top \mathbf{C}^\nu = \mathbf{1}^\top$, for all $\nu = 0, 1, \ldots$.

We conclude this section stating the counterparts of the definitions introduced in Sec. 2, adjusted here to the case of directed time-varying graphs. Using the column stochasticity of $\mathbf{C}^\nu$ and (73d), one can see that opposed to (18), the average gradient is now preserved on the weighted average of the $\mathbf{y}_i$'s:

$$\frac{1}{m} \sum_{i=1}^m \phi_i^{\nu+1} \mathbf{y}_i^{\nu+1} = \frac{1}{m} \sum_{i=1}^m \phi_i^\nu \mathbf{y}_i^\nu + \overline{\nabla \mathbf{f}}^{\nu+1} - \overline{\nabla \mathbf{f}}^\nu, \tag{74}$$

where $\overline{\nabla \mathbf{f}}^\nu$ is defined in (17). This suggests to decompose $\mathbf{y}^\nu$ into its weighted average and the consensus error, defined respectively as

$$\bar{\mathbf{y}}_{\boldsymbol{\phi}}^\nu \triangleq \frac{1}{m} \sum_{i=1}^m \phi_i^\nu \mathbf{y}_i^\nu \quad \text{and} \quad \mathbf{y}_{\boldsymbol{\phi},\perp}^\nu \triangleq \mathbf{y}^\nu - \mathbf{1}_m \otimes \bar{\mathbf{y}}_{\boldsymbol{\phi}}^\nu. \tag{75}$$

Accordingly, we define the weighted average of $\mathbf{x}^\nu$ and the consensus error as

$$\bar{\mathbf{x}}^\nu_{\boldsymbol{\phi}} \triangleq \frac{1}{m}\sum_{i=1}^m \phi_i^\nu \mathbf{x}_i^\nu \quad \text{and} \quad \mathbf{x}^\nu_{\boldsymbol{\phi},\perp} \triangleq \mathbf{x}^\nu - \mathbf{1}_m \otimes \bar{\mathbf{x}}^\nu_{\boldsymbol{\phi}}. \tag{76}$$

In addition, we also generalize the definition of the optimality gap as

$$p^\nu_{\boldsymbol{\phi}} \triangleq \sum_{i=1}^m \phi_i^\nu p_i^\nu, \quad \text{with} \quad p_i^\nu \triangleq \big(U(\mathbf{x}_i^\nu) - U^\star\big). \tag{77}$$

Finally, apart from the problem parameters L_i, L_{mx}, L, μ [cf. (4)] and algorithm parameters $\widetilde{\mu}_{\text{mn}}$, $\widetilde{L}_{\text{mx}}$, D^ℓ_{mn}, D_{mx} [cf. (24)], we introduce the following network parameters, borrowed from [35, Prop. 1]:

$$\phi_{lb} \triangleq c_\ell^{2(m-1)B}, \quad \phi_{ub} \triangleq m - c_\ell^{2(m-1)B}, \tag{78}$$

with c_ℓ and B given in Assumptions E and B$'$, respectively; and

$$c_0 \triangleq 2m \cdot \frac{1+\tilde{c}_\ell^{\,-(m-1)B}}{1-\tilde{c}_\ell^{\,-(m-1)B}}, \quad \rho_B \triangleq (1-\tilde{c}_\ell^{\,(m-1)B})^{\frac{1}{(m-1)B}}, \quad \tilde{c}_\ell \triangleq c_\ell^{2(m-1)B+1}/m. \tag{79}$$

Furthermore, we will use the following lower and upper bounds of ϕ_i^ν [35, Prop. 1]

$$\phi_{lb} \le \phi_i^\nu \le \phi_{ub}, \quad \text{for all } i \in [m], \quad \nu = 0, 1, \ldots .$$

4.1. *Linear convergence rate*

The proof of linear convergence of SONATA (Algorithm 3) follows the same path of the one developed in Sec. 3.3 for the case of undirected graphs. Hence, we omit similar derivations and highlight only the key differences. We will tacitly assume that Assumptions A, B$'$, C, and E are satisfied.

4.1.1. Step 1: $p^\nu_{\boldsymbol{\phi}}$ converges linearly up $\mathcal{O}(\|\mathbf{x}^\nu_{\boldsymbol{\phi},\perp}\|^2 + \|\mathbf{y}^\nu_{\boldsymbol{\phi},\perp}\|^2)$

This is counterpart of Proposition 3.4 (cf. Sec. 3.3), and stated as follows.

Proposition 4.1. *The optimality gap sequence $\{p^\nu_{\boldsymbol{\phi}}\}$ satisfies:*

$$p^{\nu+1}_{\boldsymbol{\phi}} \le \sigma(\alpha)\cdot p^\nu_{\boldsymbol{\phi}} + \eta(\alpha)\cdot \phi_{ub}\cdot \left(8L^2_{\text{mx}}\|\mathbf{x}^\nu_{\boldsymbol{\phi},\perp}\|^2 + 2\|\mathbf{y}^\nu_{\boldsymbol{\phi},\perp}\|^2\right), \tag{80}$$

where the constants L_{mx} and $\widetilde{\mu}_{\text{mn}}$ are defined in (4) *and* (24)*, respectively; and $\sigma(\alpha) \in (0,1)$ and $\eta(\alpha) > 0$ are defined in* (41)*.*

Proof. The proof follows closely that of Proposition 3.4 and thus is omitted. For completeness, we report it in the supporting materials. Here, we only notice that, instead of (27), we built on: $\sum_{i=1}^m \phi_i^{\nu+1} U(\mathbf{x}_i^{\nu+1}) \le \sum_{i=1}^m \phi_i^\nu U\big(\mathbf{x}_i^{\nu+\frac{1}{2}}\big)$, where we used $\sum_{j=1}^m c_{ij}^\nu \phi_j^\nu / \phi_i^{\nu+1} = 1$, for all $i \in [m]$. □

4.1.2. Step 2: Decay of $\|\mathbf{x}^\nu_{\boldsymbol{\phi},\perp}\|$ and $\|\mathbf{y}^\nu_{\boldsymbol{\phi},\perp}\|$

Lemma 4.2. *The following bounds hold for $\|\mathbf{x}^\nu_{\boldsymbol{\phi},\perp}\|$ and $\|\mathbf{y}^\nu_{\boldsymbol{\phi},\perp}\|$:*

$$\left\|\mathbf{x}_{\boldsymbol{\phi},\perp}^{\nu}\right\|^2 \leq 2c_0^2\rho_B^{2\nu}\left\|\mathbf{x}_{\boldsymbol{\phi},\perp}^{0}\right\|^2 + \frac{2c_0^2\rho_B^2}{1-\rho_B}\sum_{t=0}^{\nu-1}\rho_B^{\nu-1-t}\alpha^2\left\|\mathbf{d}^t\right\|^2 \tag{81a}$$

$$\left\|\mathbf{y}_{\boldsymbol{\phi},\perp}^{\nu}\right\|^2 \leq 2c_0^2\rho_B^{2\nu}\left\|\mathbf{y}_{\boldsymbol{\phi},\perp}^{0}\right\|^2 + \frac{2c_0^2\rho_B^2 m L_{\mathrm{mx}}^2\phi_{lb}^{-2}}{1-\rho_B}\sum_{t=0}^{\nu-1}\rho_B^{\nu-1-t}\left(8\left\|\mathbf{x}_{\boldsymbol{\phi},\perp}^{t}\right\|^2 + 2\alpha^2\left\|\mathbf{d}^t\right\|^2\right). \tag{81b}$$

where B and ρ_B are defined in (78)*, and ϵ_x and ϵ_y are arbitrary positive constants (to be determined).*

Proof. Using the result in [26, Lemma 5] and [35, Lemma 3, 11], we obtain

$$\left\|\mathbf{x}_{\boldsymbol{\phi},\perp}^{\nu}\right\| \leq c_0\left(\rho_B^{\nu}\left\|\mathbf{x}_{\boldsymbol{\phi},\perp}^{0}\right\| + \sum_{t=0}^{\nu-1}\rho_B^{(\nu-1)-t}(\rho_B\alpha\left\|\mathbf{d}^t\right\|)\right) \tag{82}$$

$$\left\|\mathbf{y}_{\boldsymbol{\phi},\perp}^{\nu}\right\| \leq c_0\left(\rho_B^{\nu}\left\|\mathbf{y}_{\boldsymbol{\phi},\perp}^{0}\right\| + \sqrt{m}L_{\mathrm{mx}}\phi_{lb}^{-1}\sum_{t=0}^{\nu-1}\rho_B^{(\nu-1)-t}\cdot\rho_B\left(2\left\|\mathbf{x}_{\boldsymbol{\phi},\perp}^{t}\right\| + \alpha\left\|\mathbf{d}^t\right\|\right)\right). \tag{83}$$

The rest of the proof follows similar steps as [52, Lemma 2], hence it is omitted. □

4.1.3. Step 3: $\|\mathbf{d}^{\nu}\| = \mathcal{O}(\sqrt{p_{\boldsymbol{\phi}}^{\nu}} + \|\mathbf{y}_{\boldsymbol{\phi},\perp}^{\nu}\|)$

Proposition 4.3. *The following upper bound holds for $\|\mathbf{d}^{\nu}\|$:*

$$\|\mathbf{d}^{\nu}\|^2 \leq \frac{6}{\mu\phi_{lb}}\left(\left(\frac{D_{\mathrm{mx}}}{\widetilde{\mu}_{\mathrm{mn}}}+1\right)^2 + \frac{4L_{\mathrm{mx}}^2}{\widetilde{\mu}_{\mathrm{mn}}^2}\right)p_{\boldsymbol{\phi}}^{\nu} + \frac{3}{\widetilde{\mu}_{\mathrm{mn}}^2}\|\mathbf{y}_{\boldsymbol{\phi},\perp}^{\nu}\|^2, \tag{84}$$

where L_{mx}, $\widetilde{L}_{\mathrm{mx}}$, $\widetilde{\mu}_{\mathrm{mn}}$, and D_{mx} are defined in (4) *and* (24)*, respectively.*

Proof. The proof follows similar path of that of Proposition 3.6 and thus is omitted. □

4.2. *Establishing linear rate*

We can now prove linear rate following the path introduced in Sec. 3.3; for sake of simplicity, we will use the same notation as in Sec. 3.3. We begin applying the transformation (49) to the sequences $\{p_{\boldsymbol{\phi}}^{\nu}\}_{\nu\in\mathbb{N}_+}$, $\{\|\mathbf{x}_{\boldsymbol{\phi},\perp}^{\nu}\|^2\}$, $\{\|\mathbf{y}_{\boldsymbol{\phi},\perp}^{\nu}\|^2\}$, and $\{\|\mathbf{d}^{\nu}\|^2\}$, satisfying the inequalities (80), (81a), (81b), and (84), respectively.

Proposition 4.4. *Let $P_{\boldsymbol{\phi}}^K(z)$, $D^K(z)$, $X_{\boldsymbol{\phi},\perp}^K(z)$, and $Y_{\boldsymbol{\phi},\perp}^K(z)$ denote the transformation* (49) *of the sequences $\{p_{\boldsymbol{\phi}}^{\nu}\}$, $\{\|\mathbf{d}^{\nu}\|^2\}$, $\{\|\mathbf{x}_{\boldsymbol{\phi},\perp}^{\nu}\|^2\}$ and $\{\|\mathbf{y}_{\boldsymbol{\phi},\perp}^{\nu}\|^2$ $\}$. Given the constants $\sigma(\alpha)$ and $\eta(\alpha)$, defined in Proposition 4.1, and the free parameters $\epsilon_x, \epsilon_y > 0$, the following holds:*

$$P_{\boldsymbol{\phi}}^K(z) \leq G_P(\alpha, z)\cdot\left(8\phi_{ub}L_{\mathrm{mx}}^2X_{\boldsymbol{\phi},\perp}^K(z) + 2\phi_{ub}Y_{\boldsymbol{\phi},\perp}^K(z)\right) + \omega_p \tag{85}$$

$$X^K_{\boldsymbol{\phi},\perp}(z) \leq G_X(z) \cdot \rho_B^2 \alpha^2 D^K(z) + \omega_x \tag{86}$$

$$Y^K_{\boldsymbol{\phi},\perp}(z) \leq G_Y(z) \cdot 2m\phi_{lb}^{-2} L_{\text{mx}}^2 \rho_B^2 \left(4X^K_{\boldsymbol{\phi},\perp}(z) + \alpha^2 D^K(z)\right) + \omega_y \tag{87}$$

$$D^K(z) \leq C_1 \cdot P^K_{\boldsymbol{\phi}}(z) + C_2 \cdot Y^K_{\boldsymbol{\phi},\perp}(z), \tag{88}$$

for all

$$z \in \left(\max\left\{\sigma(\alpha), \rho_B\right\}, 1\right), \tag{89}$$

where

$$G_P(\alpha, z) \triangleq \frac{\eta(\alpha)}{z - \sigma(\alpha)}, \qquad \omega_p \triangleq \frac{z}{z - \sigma(\alpha)} \cdot p^0_{\boldsymbol{\phi}} \tag{90}$$

$$G_X(z) \triangleq \frac{2c_0^2}{(1-\rho_B)(z-\rho_B)}, \qquad \omega_x \triangleq 2c_0^2 \left\|\mathbf{x}^0_{\boldsymbol{\phi},\perp}\right\|^2 \tag{91}$$

$$G_Y(z) \triangleq \frac{2c_0^2}{(1-\rho_B)(z-\rho_B)}, \qquad \omega_y \triangleq 2c_0^2 \left\|\mathbf{y}^0_{\boldsymbol{\phi},\perp}\right\|^2 \tag{92}$$

$$C_1 \triangleq \frac{6}{\mu\phi_{lb}} \left(\left(\frac{D_{\text{mx}}}{\widetilde{\mu}_{\text{mn}}} + 1\right)^2 + \frac{4L_{\text{mx}}^2}{\widetilde{\mu}_{\text{mn}}^2}\right), \qquad C_2 \triangleq \frac{4}{\widetilde{\mu}_{\text{mn}}^2}. \tag{93}$$

Proof. The proof of the first two inequalities (85) and (88) follows the same steps of those used to prove Proposition 3.8. Applying [43, Lemma 21] to (81a) and (81b) respectively gives (86) and (87).

□

Chaining the inequalities in Proposition 4.4 as done in for (50) (cf. Fig. 3.3.4), we can bound $D^K(z)$ as

$$D^K(z) \leq \mathcal{P}(\alpha, z) \cdot D^K(z) + \mathcal{R}(\alpha, z), \tag{94}$$

where $\mathcal{P}(\alpha, z)$ is defined as

$$\begin{aligned} \mathcal{P}(\alpha, z) \triangleq\ & G_P(\alpha, z) \cdot G_X(z) \cdot C_1 \cdot 8\phi_{ub} L_{\text{mx}}^2 \cdot \rho_B^2 \cdot \alpha^2 \\ & + (G_P(\alpha, z) \cdot 2\phi_{ub} \cdot C_1 + C_2) \cdot G_Y(z) \cdot 2m\phi_{lb}^{-2} L_{\text{mx}}^2 \cdot \rho_B^2 \cdot \alpha^2 \\ & + (G_P(\alpha, z) \cdot 2\phi_{ub} \cdot C_1 + C_2) \cdot G_Y(z) \cdot 8m\phi_{lb}^{-2} L_{\text{mx}}^2 \cdot G_X(z) \cdot \rho_B^4 \cdot \alpha^2 \end{aligned} \tag{95}$$

and $\mathcal{R}(\alpha, z)$ is a bounded remainder term.

Comparing (95) to (55) we can see that they share the same form and only differ in coefficients. Therefore, with the same argument as in the proof of Theorem 3.9 we can easily arrive at the following conclusion.

Theorem 4.5. *Consider Problem* (P) *under Assumptions A, and B′; and SONATA (Algorithm 3) under Assumptions C and E, with $\widetilde{\mu}_{\text{mn}} \geq D^{\ell}_{\text{mn}}$. Then, there exists a sufficiently small step-size $\bar{\alpha} \in (0, 1]$ such that, for all $\alpha < \bar{\alpha}$, $\{U(\mathbf{x}_i^\nu)\}$ converges to $U^\star$ at an R-linear rate, $i \in [m]$.*

Proof. We provide the proof in the supporting material. □

For sake of completeness, we provide an explicit expression of the linear rates in terms of the step-size α in the supporting material–see Theorem III.1. Table 4 summarizes the expression of the rates achieved by SONATA using the surrogate functions

(62) and (63)–a formal statement of these results along with the proofs can be found in the supporting material-see Corollaries IV.1, V.1 and V.2.

Surrogate	**Communication Rounds**	ρ_B **(network)**	β
linearization	$\mathcal{O}\left(\kappa_g \log\left(1/\epsilon\right)\right)$	$\rho_B = \mathcal{O}(\kappa_g^{-1}(1+\frac{\beta}{L})^{-2})$ or star-networks	arbitrary
	$\mathcal{O}\left(\frac{\left(\kappa_g+\beta/\mu\right)^2\rho_B}{(1-\rho_B)^2}\log(1/\epsilon)\right)$	arbitrary	
local f_i	$\mathcal{O}\left(1\cdot\log\left(1/\epsilon\right)\right)$	$\rho_B = \mathcal{O}\left(\left(1+\frac{\beta}{\mu}\right)^{-2}\left(\kappa_g+\frac{\beta}{\mu}\right)^{-2}\right)$ or star-networks	$\beta \leq \mu$
	$\mathcal{O}\left(\frac{\kappa_g^2\rho_B}{(1-\rho_B)^2}\log(1/\epsilon)\right)$	arbitrary	
	$\mathcal{O}\left(\frac{\beta}{\mu}\cdot\log\left(1/\epsilon\right)\right)$	$\rho_B = \mathcal{O}\left(\left(1+\frac{L}{\beta}\right)^{-1}\left(\kappa_g+\frac{\beta}{\mu}\right)^{-1}\right)$ or star-networks	$\beta > \mu$
	$\mathcal{O}\left(\frac{\left(\kappa_g+\beta/\mu\right)^2\rho_B}{(1-\rho_B)^2}\log(1/\epsilon)\right)$	arbitrary	

Table 4. Summary of convergence rates of SONATA over time-varying directed graphs: number of communication rounds to reach ϵ-accuracy.

The rate estimates in Table 4 are almost identical to those obtained in Sec. 3.4.2, with the difference that the network dependence now is expressed throughout ρ_B rather than ρ. Therefore, similar comments–as those stated in Sec. 3.4.2–apply to the rates in Table 4. For example, if the network is sufficiently connected (ρ_B "small"), its impact on the rate becomes negligible and SONATA matches the *network-independent* rate achieved on star-topology (cf. Corollary 3.11) or centralized settings. Specifically, when linearization surrogate (62) is used, this rate coincides with the rates of centralized proximal gradient algorithm.

5. Numerical Results

In this section, we corroborate numerically the complexity results proved in Corollaries 3.12–3.14. As a test problem, we consider the distributed ridge regression:

$$\min_{\mathbf{x}\in\mathbb{R}^d} \frac{1}{m}\left\{\frac{1}{2n}\|\mathbf{A}_i\mathbf{x}-\mathbf{b}_i\|^2+\lambda\|\mathbf{x}\|^2\right\}, \tag{96}$$

where the loss function of agent i is $f_i(\mathbf{x}) = \frac{1}{2n}\|\mathbf{A}_i\mathbf{x}-\mathbf{b}_i\|^2+\lambda\|\mathbf{x}\|^2$ [agent i owns data $(\mathbf{A}_i,\mathbf{b}_i)$]. Problem parameters are generated as follows. Each row of the measurement matrix $\mathbf{A}_i$ is independently and identically drawn from distribution $\mathcal{N}(\mathbf{0},\mathbf{\Sigma})$; and $\mathbf{b}_i$ is generated according to the linear model $\mathbf{b}_i = \mathbf{A}_i\mathbf{x}^* + \mathbf{n}_i$, where $\mathbf{x}^*$ is the ground truth, generated according to $\mathcal{N}(5\cdot\mathbf{1},\mathbf{I})$, and $\mathbf{n}_i \sim \mathcal{N}(\mathbf{0}, 0.1\cdot\mathbf{I})$ is the measurement noise. The covariance matrix $\mathbf{\Sigma}$ is constructed according to the eigenvalue decomposition $\mathbf{\Sigma} = \sum_{j=1}^d \lambda_j\mathbf{u}_j\mathbf{u}_j^\top$, where the eigenvalues $\{\lambda_j\}_{j=1}^d$ are uniformly distributed in $[\mu_0, L_0]$. The eigenvectors, forming $\mathbf{U} = [\mathbf{u}_1,\ldots,\mathbf{u}_d]$, are obtained via the QR decomposition of a random $d\times d$ matrix with standard Gaussian i.i.d. elements. The network is generated using an Erdős-Rényi model $G(m,p)$, with $m = 30$ nodes and each edge independently included in the graph with probability $p = 0.5$.

	Setting (S.I)	Setting (S.II)
Linearization	(s.1) $n = 10^3$ $\mu_0 = 1,\ L_0 = 10^3$ $K_\ell = 10,\ K_u = 100$	(s.4) $\lambda = 0$ $\mu_0 = 1,\ L_0 = 5,\ \kappa_g \approx 5$ $N_\ell = 10,\ N_u = 10^3$
Local f_i ($\beta \geq \mu$)	(s.2) same as above	(s.5) same as above
Local f_i ($\beta < \mu$)	(s.3) $n = 10^5$ $\mu_0 = 1,\ L_0 = 20$ $K_\ell = 1.1,\ K_u = 19$	(s.6) $\lambda = 0$ $\mu_0 = 1,\ L_0 = 2,\ \kappa_g \approx 2$ $N_\ell = 2 \times 10^3,\ N_u = 10^5$

Table 5. Simulation setup and parameter setting.

To investigate the impact of κ_g and β on the convergence rate, we specifically consider the following two scenarios:

(S.I) **Changing κ_g with fixed β**: We generate a sequence of instances of (96) with fixed β and increasing κ_g. To do so, we use the same data set $\{\mathbf{A}_i, \mathbf{b}_i\}$ across the different instances and change the regularization parameter λ, so that the condition number κ_g ranges in $[K_\ell, K_u]$.

(S.II) **Changing β with (almost) fixed κ_g**: We generate instances of (96) with decreasing β and (almost) fixed κ_g. To do so, we set $\lambda = 0$ and increased the local sample size n from N_ℓ to N_u; we set N_ℓ sufficiently large so that the empirical condition number κ_g is close to L_0/μ_0 for all instances.

We run SONATA using surrogates (62) (linearization) and (63) (local f_i)–we term it as SONATA-L and SONATA-F, respectively. The simulations parameters of the different experiments are summarized in Table 5; and the algorithmic parameters are set according to Corollaries 3.12–3.14.[1] We measure the algorithm's complexity using $T_\epsilon = \inf\left\{\nu \geq 0 \,|\, \frac{1}{m}\sum_{i=1}^m (F(\mathbf{x}_i^\nu) - F^\star)\ \leq 10^{-7}\right\}$.

In Table 5, we report the corresponding iteration complexity of SONATA for each simulation setup (s.1)-(s.6) in Table 5. Each figure is generated under one particular realization of the problem setting. Further, in order to compare the complexity of SONATA across different settings, all the simulations share the same network parameters, as well as the same data set whenever the problem parameters are the same. The results of our experiments are reported in Table 5; the curve are generated using only one random realization for visualization clarity. However, the behavior of the curves (e.g., scalability with respect to the parameters) is representative and consistent across all the random experiments we conducted.

The following comments are in order.

• **Scalability with respect to κ_g.** Consider setting (S.I) wherein β is fixed and λ is changing. Figures for (s.1)-(s.3) show that when $\alpha = 1$ (blue curve), the iteration complexity of SONATA-L scales linearly with respect to κ_g [as predicted by Corollary 3.12], while that of SONATA-F is invariant whenever $\beta < \mu$ [as stated in Corollary 3.13]. When $\beta \geq \mu$, the iteration complexity of SONATA-F grows as λ increases since β/μ decreases [cf. Corollary 3.14]. However, the increasing rate is much slower than SONATA-L, due to the fact that $(\beta/\mu)/\kappa_g = \beta/L \ll 1$ for large λ. When

[1] The expressions are not tight in terms of the absolute constants. To show convergence rate in both Cases I and II in Corollary 3.12-3.14, we enlarged the second term in the expression of α_{mx} by a constant factor.

$\alpha < 1$, the iteration complexity scales quadratically with respect to κ_g, in all settings, as predicted by our theory.

• **Scalability with respect to β.** Consider now setting (S.II), where we decrease the local sample size n to increase β. In contrast to setting (S.I), Figures for (s.4) and (s.5) show that, with $\alpha = 1$, the iteration complexity of SONATA-F scales linearly with β/μ when $\beta > \mu$, while that of SONATA-L is invariant–this is consistent with Corollaries 3.12 and 3.14. When $\alpha < 1$, the iteration complexity scales quadratically with respect to β/μ. Finally, the plot associated with (s.6) simply reveals that when $\beta < \mu$, iteration complexity of SONATA-F remains bounded, as stated in Corollary 3.13.

• **Linearization versus Local f_i.** We compare the performance of SONATA-L and SONATA-F in the setting (S.II), with parameters $\lambda = 0$, $\mu_0 = 1$, $L_0 = 100$, $N_\ell = 10$, $N_u = 10^5$. We consider a relatively connected network with edge activation probability $p = 0.9$ so that the step-size can be set to $\alpha = 1$, for all experiments. Note that such connectivity can also be achieved with a less connected network by running multiple but fixed rounds of consensus steps. Fig. 5 compares the iteration complexity as β increases, averaged over 100 Monte-Carlo realizations. We can see that for small β SONATA-F converges faster than SONATA-L; while for large β SONATA-L is faster. This can be explained using our results in Corollaries 3.12 and 3.14. As the complexity of SONATA-F and SONATA-L scales proportionally to β/μ and κ_g, respectively, when β/μ is comparatively smaller than κ_g, SONATA-F enjoys a better rate. But as β/μ increases, the rate deteriorates and eventually gets worse than that of SONATA-L.

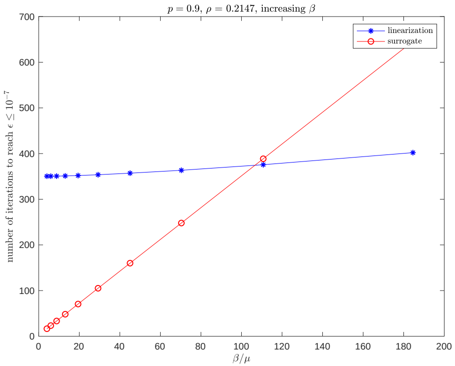


Figure 2. Complexity of SONATA-L versus SONATA-F.

Appendix A. Proof of (54)

Chaining the inequalities in (50) as shown in Fig. 3.3.4, we have

$$
\begin{aligned}
D^K(z) \leq & C_1 \cdot P^K(z) + C_2 \cdot Y_\perp^K(z) \\
& \leq C_1 \cdot \Big(G_P(\alpha, z) \cdot \big(4L_{\text{mx}}^2 X_\perp^K(z) + 2Y_\perp^K(z)\big) + \omega_p \Big) + C_2 \cdot Y_\perp^K(z) \\
& = C_1 \cdot G_P(\alpha, z) \cdot 4L_{\text{mx}}^2 X_\perp^K(z) + (C_1 \cdot G_P(\alpha, z) \cdot 2 + C_2) Y_\perp^K(z) + C_1 \cdot \omega_p \\
& \leq C_1 \cdot G_P(\alpha, z) \cdot 4L_{\text{mx}}^2 \cdot G_X(z) \cdot \rho^2 \alpha^2 D^K(z) \\
& \quad + (C_1 \cdot G_P(\alpha, z) \cdot 2 + C_2) \cdot G_Y(z) \cdot 8L_{\text{mx}}^2 \rho^2 X_\perp^K(z) \\
& \quad + (C_1 \cdot G_P(\alpha, z) \cdot 2 + C_2) \cdot G_Y(z) \cdot 2L_{\text{mx}}^2 \rho^2 \alpha^2 D^K(z) \\
& \quad + C_1 \cdot \omega_p + (C_1 \cdot G_P(\alpha, z) \cdot 2 + C_2) \cdot \omega_y + C_1 \cdot G_P(\alpha, z) \cdot 4L_{\text{mx}}^2 \cdot \omega_x
\end{aligned}
$$

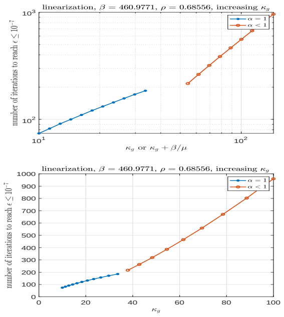
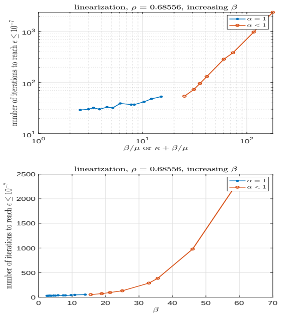
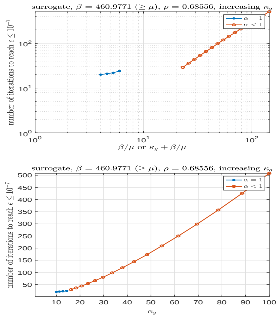
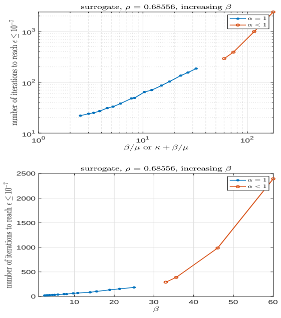
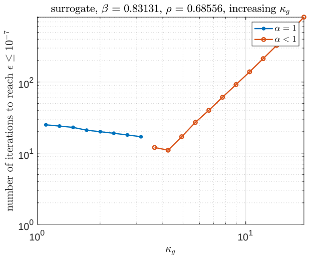
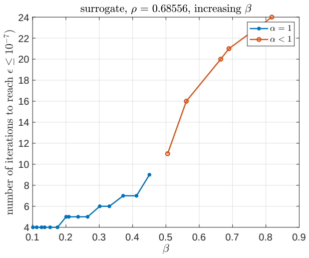
<table>
<tr><th></th><th>Setting (S.I)</th><th>Setting (S.II)</th></tr>
<tr><td>Linearization</td><td>(s.1)

</td><td>(s.4)

</td></tr>
<tr><td>Local f_i ($\beta \geq \mu$)</td><td>(s.2)

</td><td>(s.5)

</td></tr>
<tr><td>Local f_i ($\beta < \mu$)</td><td>(s.3)

</td><td>(s.6)

</td></tr>
</table>

Table 6. Iteration complexity of SONATA under the simulation settings in Table 5. Left (S.I): scalability of iteration complexity with respect to the condition number κ_g; Right (S.II): scalability of the iteration complexity with respect to the similarity parameter β.

$$\begin{aligned}
&\leq C_1 \cdot G_P(\alpha, z) \cdot 4L_{\text{mx}}^2 \cdot G_X(z) \cdot \rho^2 \alpha^2 D^K(z) \\
&\quad + (C_1 \cdot G_P(\alpha, z) \cdot 2 + C_2) \cdot G_Y(z) \cdot 8L_{\text{mx}}^2 \rho^2 \cdot G_X(z) \cdot \rho^2 \alpha^2 D^K(z) \\
&\quad + (C_1 \cdot G_P(\alpha, z) \cdot 2 + C_2) \cdot G_Y(z) \cdot 2L_{\text{mx}}^2 \rho^2 \alpha^2 D^K(z) \\
&\quad + C_1 \cdot \omega_p + (C_1 \cdot G_P(\alpha, z) \cdot 2 + C_2) \cdot \omega_y + C_1 \cdot G_P(\alpha, z) \cdot 4L_{\text{mx}}^2 \cdot \omega_x \\
&\quad + (C_1 \cdot G_P(\alpha, z) \cdot 2 + C_2) \cdot G_Y(z) \cdot 8L_{\text{mx}}^2 \rho^2 \cdot \omega_x.
\end{aligned}$$

Notice that, under (51), $G_P(\alpha, z)$, $G_X(z)$, $G_Y(z)$, and ω_p, ω_x, ω_y are all bounded, which implies that the reminder $\mathcal{R}(\alpha, z)$ in (50) is bounded as well. □

Appendix B. Proof of Theorem 3.10

We find the smallest z satisfying (51) such that $\mathcal{P}(\alpha,z)<1$, for $\alpha\in(0,\alpha_{\rm mx})$, with $\alpha_{\rm mx}\in(0,1)$ to be determined.

Let us begin considering the condition $z>\sigma(\alpha)$ in (51). To simplify the analysis, we impose instead the following stronger version

$$z\geq\sigma(\alpha)+\frac{(\theta\cdot\alpha)\cdot\left(\left(1-\frac{\alpha}{2}\right)\widetilde{\mu}_{\rm mn}+\frac{D^{\ell}_{\rm mn}}{2}\alpha-\frac{1}{2}\epsilon_{opt}\right)}{\frac{D^2_{\rm mx}}{\mu}+\left(1-\frac{\alpha}{2}\right)\widetilde{\mu}_{\rm mn}+\frac{D^{\ell}_{\rm mn}}{2}\alpha-\frac{1}{2}\epsilon_{opt}}\tag{B1}$$

for some $\theta\in(0,1)$, which will be chosen to tighten the bound. Notice that the RHS of (B1) is strictly larger than $\sigma(\alpha)$ but still strictly less than one, for any $\alpha\in(0,(2\widetilde{\mu}_{\rm mn}-\epsilon_{opt})/(\widetilde{\mu}_{\rm mn}-D^{\ell}_{\rm mn}))$, with given $\epsilon_{opt}\in(0,2\widetilde{\mu}_{\rm mn})$.

Observe that in the expression of $\mathcal{P}(\alpha,z)$, the only coefficient multiplying α^2 that depends on α is the optimization gain $G_P(\alpha,z)\triangleq\eta(\alpha)/(z-\sigma(\alpha))$. Using (B1), $G_P(\alpha,z)$ can be upper bounded as

$$\begin{aligned}G_P(\alpha,z)&\leq\inf_{\epsilon_{opt}\in(0,2\widetilde{\mu}_{\rm mn}-\alpha(\widetilde{\mu}_{\rm mn}-D^{\ell}_{\rm mn}))}\frac{\frac{1}{2}\epsilon_{opt}^{-1}\cdot\frac{D^2_{\rm mx}}{\mu}+\frac{1}{\mu}\cdot\left(\left(1-\frac{\alpha}{2}\right)\widetilde{\mu}_{\rm mn}+\frac{D^{\ell}_{\rm mn}}{2}\alpha-\frac{1}{2}\epsilon_{opt}\right)}{\left(1-\frac{\alpha}{2}\right)\widetilde{\mu}_{\rm mn}+\frac{D^{\ell}_{\rm mn}}{2}\alpha-\frac{1}{2}\epsilon_{opt}}\cdot\theta^{-1}\\&=G^{\star}_P(\alpha)\cdot\theta^{-1},\end{aligned}\tag{B2}$$

where the minimum is attained at $\epsilon^{\star}_{opt}\triangleq\widetilde{\mu}_{\rm mn}-\frac{\alpha}{2}(\widetilde{\mu}_{\rm mn}-D^{\ell}_{\rm mn})$; and $G^{\star}_P(\alpha)$ is defined in (59). Substituting the upper bound (B2) in $\mathcal{P}(\alpha,z)$ and setting therein $\epsilon_{opt}=\epsilon^{\star}_{opt}$, we get the following sufficient condition for $\mathcal{P}(\alpha,z)<1$:

$$\begin{aligned}&G^{\star}_P(\alpha)\cdot\theta^{-1}\cdot C_1\cdot 4L^2_{\rm mx}\cdot G_X(z)\cdot\rho^2\cdot\alpha^2\\&\qquad+\left(G^{\star}_P(\alpha)\cdot\theta^{-1}\cdot 2C_1+C_2\right)\cdot G_Y(z)\cdot 2L^2_{\rm mx}\rho^2\cdot\alpha^2\\&\quad+\left(G^{\star}_P(\alpha)\cdot\theta^{-1}\cdot 2C_1+C_2\right)\cdot G_Y(z)\cdot 8L^2_{\rm mx}\rho^2\cdot G_X(z)\cdot\rho^2\cdot\alpha^2<1.\end{aligned}\tag{B3}$$

To minimize the left hand side, we set $\epsilon_x=\epsilon_y=(\sqrt{z}-\rho)/\rho$. Furthermore, using the fact that $G^{\star}_P(\alpha)$ is monotonically increasing on $\alpha\in(0,2\widetilde{\mu}_{\rm mn}/(\widetilde{\mu}_{\rm mn}-D^{\ell}_{\rm mn}))$, and restricting $\alpha\in(0,\widetilde{\mu}_{\rm mn}/(\widetilde{\mu}_{\rm mn}-D^{\ell}_{\rm mn})]$, a sufficient condition for (B3) is

$$\alpha\leq\alpha(z)\triangleq\left(A_{1,\theta}\frac{1}{(\sqrt{z}-\rho)^2}+A_{2,\theta}\frac{1}{(\sqrt{z}-\rho)^2}+A_{3,\theta}\frac{1}{(\sqrt{z}-\rho)^4}\right)^{-1/2},\tag{B4}$$

where $A_{1,\theta}$, $A_{2,\theta}$ and $A_{3,\theta}$ are constants defined as

$$\begin{aligned}A_{1,\theta}&\triangleq G^{\star}_P(\widetilde{\mu}_{\rm mn}/(\widetilde{\mu}_{\rm mn}-D^{\ell}_{\rm mn}))\cdot\theta^{-1}\cdot C_1\cdot 4L^2_{\rm mx}\cdot\rho^2\\A_{2,\theta}&\triangleq\left(G^{\star}_P(\widetilde{\mu}_{\rm mn}/(\widetilde{\mu}_{\rm mn}-D^{\ell}_{\rm mn}))\cdot\theta^{-1}\cdot 2C_1+C_2\right)\cdot 2L^2_{\rm mx}\rho^2\\A_{3,\theta}&\triangleq\left(G^{\star}_P(\widetilde{\mu}_{\rm mn}/(\widetilde{\mu}_{\rm mn}-D^{\ell}_{\rm mn}))\cdot\theta^{-1}\cdot 2C_1+C_2\right)\cdot 8L^2_{\rm mx}\rho^4.\end{aligned}$$

Condition (B4) shows the rate z must satisfy

$$z\geq\left(\rho+\sqrt{A_\theta\alpha}\right)^2,\quad\text{with}\quad A_\theta\triangleq\sqrt{A_{1,\theta}+A_{2,\theta}+A_{3,\theta}}.\tag{B5}$$

Notice that, under $\epsilon_x=\epsilon_y=(\sqrt{z}-\rho)/\rho$, (B5) implies $z>\rho^2(1+\epsilon_x)=\rho^2(1+\epsilon_y)=\rho\sqrt{z}$, which are the other two conditions on z in (51). Therefore, overall, z must satisfy

(B1) and (B5). Letting $\epsilon_{opt} = \epsilon^\star_{opt}$ in (B1), the condition simplifies to

$$z \geq 1 - \frac{\widetilde{\mu}_{\rm mn} - \frac{\alpha}{2}(\widetilde{\mu}_{\rm mn} - D^\ell_{\rm mn})}{\frac{2D^2_{\rm mx}}{\mu} + \widetilde{\mu}_{\rm mn} - \frac{\alpha}{2}(\widetilde{\mu}_{\rm mn} - D^\ell_{\rm mn})} \cdot (1-\theta)\alpha.$$

Therefore, the overall convergence rate can be upper bounded by $\mathcal{O}(\bar{z}^\nu)$, where

$$\bar{z} = \inf_{\theta \in (0,1)} \max\left\{ \left(\rho + \sqrt{A_\theta \alpha}\right)^2, 1 - \frac{\widetilde{\mu}_{\rm mn} - \frac{\alpha}{2}(\widetilde{\mu}_{\rm mn} - D^\ell_{\rm mn})}{\frac{2D^2_{\rm mx}}{\mu} + \widetilde{\mu}_{\rm mn} - \frac{\alpha}{2}(\widetilde{\mu}_{\rm mn} - D^\ell_{\rm mn})} \cdot (1-\theta)\alpha \right\}. \quad \text{(B6)}$$

Finally, we further simplify (B6). Letting $\theta = 1/2$ and using $\alpha \in (0, \widetilde{\mu}_{\rm mn}/(\widetilde{\mu}_{\rm mn} - D^\ell_{\rm mn})]$, the second term in (B6) can be upper bounded by

$$1 - \underbrace{\frac{\widetilde{\mu}_{\rm mn}\mu}{4D^2_{\rm mx} + \widetilde{\mu}_{\rm mn}\mu} \cdot \frac{1}{2}}_{\triangleq J} \alpha. \quad \text{(B7)}$$

The condition $\bar{z} < 1$ imposes the following upper bound on α: $\alpha < \alpha_{\rm mx} = \min\{(1-\rho)^2/A_{\frac{1}{2}}, \widetilde{\mu}_{\rm mn}/(\widetilde{\mu}_{\rm mn} - D^\ell_{\rm mn}), 1\}$. Eq. (B6) then simplifies to

$$\bar{z} = \max\left\{ \left(\rho + \sqrt{\alpha A_{\frac{1}{2}}}\right)^2, 1 - J\alpha \right\}. \quad \text{(B8)}$$

Note that as α increases from 0, the first term in the max operator above is monotonically increasing from $\rho^2 < 1$ while the second term is monotonically decreasing from 1. Therefore, there must exist some α^* so that the two terms are equal, which is

$$\alpha^* = \left(\frac{-\rho\sqrt{A_{\frac{1}{2}}} + \sqrt{A_{\frac{1}{2}} + J(1-\rho^2)}}{A_{\frac{1}{2}} + J} \right)^2. \quad \text{(B9)}$$

To conclude, given the step-size satisfying $\alpha \in (0, \alpha_{\rm mx})$, the sequence $\{\|\mathbf{d}^\nu\|^2\}$ converges at rate $\mathcal{O}(z^\nu)$, with z given in (61). □

Appendix C. Proof of Corollary 3.11

Since $\mathbf{W} = \mathbf{J}$, we have $\boldsymbol{\delta}^\nu = \mathbf{0}$; then (33a) and (35) reduce to

$$p^{\nu+1} \leq p^\nu - \left(\left(1 - \frac{\alpha}{2}\right) \widetilde{\mu}_{\rm mn} + \frac{\alpha D^\ell_{\rm mn}}{2} \right) \alpha \|\mathbf{d}^\nu\|^2 \quad \text{(C1)}$$

and

$$\alpha \|\mathbf{d}^\nu\|^2 \geq \frac{2\mu}{D^2_{\rm mx}} \left(p^{\nu+1} - (1-\alpha)p^\nu\right), \quad \text{(C2)}$$

respectively. Combining (C1) and (C2) and using $\alpha < 2\widetilde{\mu}_{\rm mn}/(\widetilde{\mu}_{\rm mn} - D_{\rm mn})$, yield

$$p^{\nu+1} \leq \left(1 - \alpha \cdot \frac{\left(1 - \frac{\alpha}{2}\right)\widetilde{\mu}_{\rm mn} + \frac{\alpha D^\ell_{\rm mn}}{2}}{\frac{D^2_{\rm mx}}{2\mu} + \left(1 - \frac{\alpha}{2}\right)\widetilde{\mu}_{\rm mn} + \frac{\alpha D^\ell_{\rm mn}}{2}} \right) p^\nu, \quad \text{(C3)}$$

which proves (64).

We customize next (64) to the specific choices of the surrogate functions.

- **Linearization:** Consider the choice of $\widetilde{f}_i$ as in (62). We have $\widetilde{\mu}_{\text{mn}} = L$; and we can set $D^{\ell}_{\text{mn}} = 0$, $D_{\text{mx}} = L - \mu$, and $\alpha = 1$. Substituting these values in (64), we obtain $z \leq 1 - \kappa_g^{-1}$.
- **Local f_i:** Consider now $\widetilde{f}_i$ as in (63). By $\nabla^2 f_i(\mathbf{x}) \succeq \mathbf{0}$, for all $\mathbf{x} \in \mathcal{K}$, and Definition 2.1, we have $\mathbf{0} \preceq \nabla^2 \widetilde{f}_i(\mathbf{x}, \mathbf{y}) - \nabla^2 F(\mathbf{x}) \preceq 2\beta \mathbf{I}$, for all $\mathbf{x}, \mathbf{y} \in \mathcal{K}$. Therefore, we can set $D^{\ell}_{\text{mn}} = 0$, $D_{\text{mx}} = 2\beta$, and $\widetilde{\mu}_{\text{mn}} = \beta + (\mu - \beta)_+$. Using these values in (64), yields

$$z \begin{cases} = 1 - \alpha \cdot \frac{\beta\left(1 - \frac{\alpha}{2}\right)}{\frac{2\beta^2}{\mu} + \beta\left(1 - \frac{\alpha}{2}\right)}, & \text{if } \mu \leq \beta \\ \leq 1 - \alpha \cdot \frac{\mu\left(1 - \frac{\alpha}{2}\right)}{\frac{2\beta^2}{\mu} + \mu\left(1 - \frac{\alpha}{2}\right)}, & \text{if } \mu > \beta. \end{cases} \tag{C4}$$

Finally, setting $\alpha = \min\{1, 2\widetilde{\mu}_{\text{mn}}/((\mu - \beta)_+ + \beta)\} = 1$ in the expression above, yields (65). □

Appendix D. Proof of Corollary 3.12

According to Theorem 3.10, the rate z can be bounded as

$$z \leq \max\{z_1, z_2\}, \qquad \text{with} \qquad z_1 \triangleq 1 - \alpha \cdot J \quad \text{and} \quad z_2 \triangleq \left(\rho + \sqrt{\alpha A_{\frac{1}{2}}}\right)^2, \tag{D1}$$

where J and $A_{\frac{1}{2}}$ are defined in (B7) and (B5), respectively.

The proof consists in bounding properly z_1 and z_2 based upon the surrogate (62) postulated in the corollary. We begin particularizing the expressions of J and $A_{\frac{1}{2}}$. Since $\nabla^2 \widetilde{f}_i(\mathbf{x}_i; \mathbf{x}_i^{\nu}) = L$, one can set $\widetilde{\mu}_{\text{mn}} = L$, and (15) holds with $D^{\ell}_{\text{mn}} = 0$ and $D_{\text{mx}} = L - \mu$. Furthermore, by Assumption 2.1, it follows that $\beta \geq \lambda_{\max}(\nabla^2 f_i(\mathbf{x})) - L$, for all $\mathbf{x} \in \mathcal{K}$; hence, one can set $L_{\text{mx}} = L + \beta$. Next, we will substitute the above values into the expressions of J and $A_{\frac{1}{2}}$.

To do so, we need to particularize first the quantities $G_P^{\star}\left(\frac{\widetilde{\mu}_{\text{mn}}}{\widetilde{\mu}_{\text{mn}} - D^{\ell}_{\text{mn}}}\right)$ [cf. (59)], C_1 and C_2 [cf. (52d)]:

$$\begin{aligned} G_P^{\star}\left(\frac{\widetilde{\mu}_{\text{mn}}}{\widetilde{\mu}_{\text{mn}} - D^{\ell}_{\text{mn}}}\right) &= G_P^{\star}(1) = \frac{4(L - \mu)^2 + L^2}{\mu L^2}, \\ C_1 &= \frac{6}{\mu L^2}\left((2L - \mu)^2 + 4(L + \beta)^2\right), \quad \text{and} \quad C_2 = \frac{4}{L^2}. \end{aligned}$$

Accordingly, the expressions of J and $A_{\frac{1}{2}}$ read:

$$J = \frac{1}{2}\frac{\kappa_g}{4(\kappa_g - 1)^2 + \kappa_g} \in \left[\frac{1}{8\kappa_g}, \frac{1}{2}\right], \tag{D2}$$

and

$$
\begin{aligned}
&(A_{\frac{1}{2}})^2\\
=&\, G_P^\star(1)\cdot 2\cdot C_1\cdot 4L_{\mathrm{mx}}^2\cdot\rho^2+(G_P^\star(1)\cdot 4\cdot C_1+C_2)\cdot 2L_{\mathrm{mx}}^2\rho^2\\
&\;+(G_P^\star(1)\cdot 4\cdot C_1+C_2)\cdot 8L_{\mathrm{mx}}^2\rho^4\\
=&\,(24G_P^\star(1)\cdot C_1+5C_2)\cdot 2L_{\mathrm{mx}}^2\rho^2\\
=&\left(24\cdot\frac{4(L-\mu)^2+L^2}{\mu L^2}\cdot\frac{6}{\mu L^2}\left((2L-\mu)^2+4(L+\beta)^2\right)+20L^{-2}\right)\cdot 2(L+\beta)^2\rho^2\\
\leq&\left(24\cdot\frac{5}{\mu}\cdot\frac{24}{\mu L^2}\left(L^2+(L+\beta)^2\right)+20L^{-2}\right)\cdot 2(L+\beta)^2\rho^2\\
=&\left(24\cdot 24\cdot 5\left(1+\left(1+\frac{\beta}{L}\right)^2\right)\left(1+\frac{\beta}{L}\right)^2\kappa_g^2+20\left(1+\frac{\beta}{L}\right)^2\right)\cdot 2\rho^2\\
\leq&\,110^2\cdot\kappa_g^2\left(1+\frac{\beta}{L}\right)^4\rho^2,
\end{aligned}
\tag{D3}
$$

where in the last inequality we have used the fact that $\kappa_g \geq 1$.

Using the above expressions, in the sequel we upperbound z_1 and z_2.

By (D3), we have

$$
z_2 \leq \bar{z}_2 \triangleq \left(\rho+\sqrt{\alpha M\rho}\right)^2,\quad \text{with}\quad M\triangleq 110\cdot\kappa_g(1+\beta/L)^2. \tag{D4}
$$

Since $\alpha\in(0,1]$ must be chosen so that $z\in(0,1]$, we impose $\max\{z_1,\bar{z}_2\}<1$, implying $\alpha\leq\min\{J^{-1},(1-\rho)^2/(M\rho),1\}$. Since $J^{-1}>1$ [cf. (D2)], the condition on α reduces to $\alpha\leq\alpha_{\mathrm{mx}}\triangleq\min\{(1-\rho)^2/(M\rho),1\}$. Choose $\alpha=c\cdot\alpha_{\mathrm{mx}}$, for some given $c\in(0,1)$. Depending on the value of ρ, either $\alpha_{\mathrm{mx}}=1$ or $\alpha_{\mathrm{mx}}=(1-\rho)^2/(M\rho)$.

• **Case I:** $\alpha_{\mathrm{mx}}=1$**.** This corresponds to the case $M\rho\leq(1-\rho)^2$, which happens when the network is sufficiently connected (ρ is small). Note that, we also have $\rho\leq 1/110$, otherwise $M\rho\geq 110\,\kappa_g\,\rho>1>(1-\rho)^2$. In this setting, $\alpha=c\cdot\alpha_{\mathrm{mx}}=c$, and

$$
\begin{aligned}
z_1&=1-c\cdot J,\\
\bar{z}_2&=\left(\rho+\sqrt{cM\rho}\right)^2\overset{(a)}{\leq}\left(1-(1-\rho)+\sqrt{c(1-\rho)^2}\right)^2\\
&=\left(1-\left(1-\sqrt{c}\right)(1-\rho)\right)^2\leq 1-\left(1-\sqrt{c}\right)^2(1-\rho)^2\\
&\overset{(b)}{\leq}1-(1-\sqrt{c})^2(1-1/110)^2,
\end{aligned}
$$

where in (a) we used $M\rho\leq(1-\rho)^2$ and (b) follows from $\rho\leq 1/110$.

Therefore, z can be bounded as

$$
\begin{aligned}
z\leq\max\{z_1,\bar{z}_2\}&\leq 1-c\cdot\left(1-\sqrt{c}\right)^2\cdot(1-1/110)^2\cdot J\\
&\leq 1-c\cdot\left(1-\sqrt{c}\right)^2\cdot(1-1/110)^2\cdot\frac{1}{8\kappa_g}.
\end{aligned}
\tag{D5}
$$

• **Case II:** $\alpha_{\text{mx}} = (1-\rho)^2/(M\rho)$**.** This corresponds to the case $M\rho \geq (1-\rho)^2$. We have $\alpha = c \cdot \alpha_{\text{mx}} = c \cdot (1-\rho)^2/(M\rho)$,

$$z_1 = 1 - \frac{J\,c}{M\rho} \cdot (1-\rho)^2 \quad \text{and} \quad \bar{z}_2 = 1 - \left(1-\sqrt{c}\right)^2 (1-\rho)^2.$$

We claim that $(J\,c)/(M\rho) < 1$. Suppose this is not the case, that is, $M\rho \leq Jc$. Since $Jc < 1/2$ [cf. (D2)] and $M \geq 110\,\kappa$, $M\rho \leq Jc$ would imply $\rho < 1/(220\kappa_g)$. This however is in contradiction with the assumption $M\rho \geq (1-\rho)^2$, as it would lead to $1/2 > M\rho \geq (1-\rho)^2 > (1-1/(220\kappa_g))^2$.

Using $(J\,c)/(M\rho) < 1$, we can bound z

$$\begin{aligned} z &\leq \max\{z_1, \bar{z}_2\} \leq 1 - \frac{c\,J}{M\rho} \cdot \left(1-\sqrt{c}\right)^2 (1-\rho)^2 \\ &\leq 1 - c \cdot \left(1-\sqrt{c}\right)^2 \cdot \frac{1}{8\kappa_g} \cdot \frac{(1-\rho)^2}{110 \cdot \kappa_g \cdot (1+\beta/L)^2 \cdot \rho}. \end{aligned}$$

Appendix E. Proof of Corollaries 3.13 and 3.14

We follow similar steps as in Appendix D but customized to the surrogate (63). We begin particularizing the expressions of J and $A_{\frac{1}{2}}$.

In the setting of the corollary, we have: $\nabla^2 \widetilde{f}_i(\mathbf{x};\mathbf{y}) = \nabla^2 f_i(\mathbf{x}) + \beta\mathbf{I}$, for all $\mathbf{y} \in \mathcal{K}$; $\nabla^2 f_i(\mathbf{x}) \succeq \mathbf{0}$, for all $\mathbf{x} \in \mathcal{K}$; and, by Assumption 2.1, $\mathbf{0} \preceq \nabla^2 \widetilde{f}_i(\mathbf{x},\mathbf{y}) - \nabla^2 F(\mathbf{x}) \preceq 2\beta\mathbf{I}$, for all $\mathbf{x},\mathbf{y} \in \mathcal{K}$. Therefore, we can set $D^{\ell}_{\text{mn}} = 0$, $D_{\text{mx}} = 2\beta$, $\widetilde{\mu}_{\text{mn}} = \beta + (\mu-\beta)_+ = \max\{\beta,\mu\}$, and $L_{\text{mx}} = L + \beta$. Using these values, $G^{\star}_P\left(\frac{\widetilde{\mu}_{\text{mn}}}{\widetilde{\mu}_{\text{mn}} - D^{\ell}_{\text{mn}}}\right)$, C_1, and C_2 can be simplified as follows:

$$\begin{aligned} &G^{\star}_P\left(\frac{\widetilde{\mu}_{\text{mn}}}{\widetilde{\mu}_{\text{mn}} - D^{\ell}_{\text{mn}}}\right) = G^{\star}_P(1) = \frac{16\beta^2 + \max\{\beta,\mu\}^2}{\mu\max\{\beta,\mu\}^2}, \\ &C_1 = \frac{6}{\mu}\left(\left(\frac{2\beta}{\max\{\beta,\mu\}} + 1\right)^2 + \frac{4(L+\beta)^2}{\max\{\beta,\mu\}^2}\right), \quad \text{and} \quad C_2 = \frac{4}{\max\{\beta,\mu\}^2}. \end{aligned}$$

Accordingly, the expressions of J and $A_{\frac{1}{2}}$ read:

$$J = \frac{1}{2}\frac{1}{1 + 16\left(\frac{\beta}{\mu}\right) \cdot \min\left\{1, \frac{\beta}{\mu}\right\}}, \tag{E1}$$

and

$$\begin{aligned} &(A_{\frac{1}{2}})^2 \\ &\leq (24G^{\star}_P(1) \cdot C_1 + 5C_2) \cdot 2L^2_{\text{mx}}\rho^2 \\ &\leq \left(24 \cdot \frac{16\beta^2 + \max\{\beta,\mu\}^2}{\max\{\beta,\mu\}^2} \cdot \frac{6}{\mu^2}\left(\left(\frac{2\beta}{\max\{\beta,\mu\}} + 1\right)^2 + \frac{4(L+\beta)^2}{\max\{\beta,\mu\}^2}\right) + \frac{20}{\max\{\beta,\mu\}^2}\right) \cdot 2(L+\beta)^2\rho^2 \\ &= \begin{cases} \left(24 \cdot 17 \cdot 6 \cdot \left(9 + 4\left(1 + \frac{L}{\beta}\right)^2\right) \cdot \left(\kappa_g + \frac{\beta}{\mu}\right)^2 + 20\left(1 + \frac{L}{\beta}\right)^2\right) \cdot 2\rho^2, & \beta > \mu, \\ \left(24 \cdot \left(\frac{16\beta^2}{\mu^2} + 1\right) \cdot 6\left(\kappa_g + \frac{\beta}{\mu}\right)^2\left(\left(\frac{2\beta}{\mu} + 1\right)^2 + 4\left(\kappa_g + \frac{\beta}{\mu}\right)^2\right) + 20\left(\kappa_g + \frac{\beta}{\mu}\right)^2\right) \cdot 2\rho^2, & \beta \leq \mu; \end{cases} \\ &\leq M^2\rho^2, \end{aligned}$$

where

$$M = \begin{cases} 253\left(1+\frac{L}{\beta}\right)\left(\kappa_g+\frac{\beta}{\mu}\right), & \beta > \mu, \\ 193\left(1+\frac{\beta}{\mu}\right)^2\left(\kappa_g+\frac{\beta}{\mu}\right)^2, & \beta \leq \mu. \end{cases} \tag{E2}$$

Similarly to the proof of Corollary 3.12, we bound $z \leq \max\{z_1, z_2\}$ as

$$z \leq \max\{z_1, \bar{z}_2\}, \quad \text{with} \quad z_1 \triangleq 1 - \alpha \cdot J \;\text{ and }\; \bar{z}_2 \triangleq \left(\rho + \sqrt{\alpha\, M\, \rho}\right)^2, \tag{E3}$$

where J and M are now given by (E1) and (E2), respectively. For $\max\{z_1, z_2\} < 1$, we require $\alpha \leq \alpha_{\mathrm{mx}} \triangleq \min\{1, (1-\rho)^2/(M\rho)\}$, and choose $\alpha = c \cdot \alpha_{\mathrm{mx}}$, with arbitrary $c \in (0,1)$. We study separately the cases $\beta > \mu$ and $\beta \leq \mu$.
1) $\beta > \mu$. In this case we have

$$M = 253\left(1+\frac{L}{\beta}\right)\left(\kappa_g+\frac{\beta}{\mu}\right) \quad \text{and} \quad J = \frac{1}{2}\frac{1}{1+16\,(\beta/\mu)} \geq \frac{1}{34(\beta/\mu)}. \tag{E4}$$

Since $\alpha = c\alpha_{\mathrm{mx}} = c\min\{1, (1-\rho)^2/(M\rho)\}$, we study next the case $\alpha_{\mathrm{mx}} = 1$ and $\alpha_{\mathrm{mx}} = (1-\rho)^2/(M\rho)$ separately.

- **Case I:** $\alpha_{\mathrm{mx}} = 1$. We have $M\rho \leq (1-\rho)^2$, $\alpha = c$, and thus

  $$z_1 = 1 - c \cdot J \quad \text{and} \quad \bar{z}_2 \leq 1 - \left(1-\sqrt{c}\right)^2 (1-\rho)^2.$$

  Since $M \geq 253$ and $(1-\rho)^2 \leq 1$, it must be $\rho \leq 1/253$. Therefore, the rate z can be bounded as

  $$\begin{aligned} z \leq \max\{z_1, \bar{z}_2\} &\leq 1 - c \cdot \left(1-\sqrt{c}\right)^2 \cdot J \cdot (1-\rho)^2 \\ &\leq 1 - c \cdot \left(1-\sqrt{c}\right)^2 \cdot \left(1 - \frac{1}{253}\right)^2 \cdot \frac{1}{34} \cdot \frac{\mu}{\beta}. \end{aligned}$$

- **Case II:** $\alpha_{\mathrm{mx}} = (1-\rho)^2/(M\rho)$. This corresponds to $M\rho \geq (1-\rho)^2$, $\alpha = c \cdot (1-\rho)^2/(M\rho)$, and

  $$z_1 = 1 - \frac{J\,c}{M\rho} \cdot (1-\rho)^2 \quad \text{and} \quad \bar{z}_2 \leq 1 - \left(1-\sqrt{c}\right)^2 (1-\rho)^2.$$

  Using the same argument as in the proof of Corollary 3.12–Case II, one can show that $(c\,J)/(M\rho) < 1$. Therefore,

  $$\begin{aligned} z \leq \max\{z_1, \bar{z}_2\} &\leq 1 - \left(1-\sqrt{c}\right)^2 \cdot c\,J \cdot \frac{(1-\rho)^2}{M\rho} \\ &\overset{\text{(E4)}}{\leq} 1 - c \cdot \left(1-\sqrt{c}\right)^2 \cdot \frac{1}{34} \cdot \frac{(1-\rho)^2}{253\left(\kappa_g + \frac{\beta}{\mu}\right)^2 \rho}. \end{aligned}$$

2) $\boldsymbol{\beta \le \mu}$**.** In this case we have

$$M = 193\left(1+\frac{\beta}{\mu}\right)^2\left(\kappa_g+\frac{\beta}{\mu}\right)^2 \quad \text{and} \quad J = \frac{1}{2}\frac{1}{1+16\,(\beta/\mu)^2}. \tag{E5}$$

- **Case I:** $\alpha_{\mathrm{mx}} = 1$**.** Following the same reasoning as $\mu \le \beta$, we can prove

$$z \le \max\{z_1,\bar{z}_2\} \le 1 - c\cdot\left(1-\sqrt{c}\right)^2\cdot\left(1-\frac{1}{193}\right)^2\cdot\frac{1}{2+32\left(\frac{\beta}{\mu}\right)^2}. \tag{E6}$$

- **Case II:** $\alpha_{\mathrm{mx}} = (1-\rho)^2/(M\rho)$**.** We claim that $(c\,J)/(M\rho) \le 1$, otherwise $\rho \le c/386$, which would lead to the following contradiction $c/2 \ge (c\,J) > M\rho \ge (1-\rho)^2 \ge (1-c/386)^2$. Therefore,

$$\begin{aligned} z \le \max\{z_1,\bar{z}_2\} &\le 1 - c\cdot\left(1-\sqrt{c}\right)^2\cdot\frac{1}{2+32\left(\frac{\beta}{\mu}\right)^2}\frac{(1-\rho)^2}{193\left(1+\frac{\beta}{\mu}\right)^2\left(\kappa_g+\frac{\beta}{\mu}\right)^2\rho} \\ &\le 1 - c'\cdot\frac{(1-\rho)^2}{\kappa_g^2\,\rho}, \end{aligned}$$

where $c' \in (0,1)$ is a suitable constant, independent on β/μ, κ_g, and ρ. □

References

[1] S.A. Alghunaim, K. Yuan, and A.H. Sayed, *A linearly convergent proximal gradient algorithm for decentralized optimization*, arXiv:1905.07996 (2019).

[2] Y. Arjevani and O. Shamir, *Communication complexity of distributed convex learning and optimization*, in *Proc. of the 28th International Conference on Neural Information Processing Systems (NIPS)*, Vol. 1. 2005, pp. 1756–1764.

[3] A. Berahas, R. Bollapragada, N.S. Keskar, and E. Wei, *Balancing Communication and Computation in Distributed Optimization*, IEEE Trans. Autom. Control (to appear, 2019).

[4] P. Di Lorenzo and G. Scutari, *Distributed nonconvex optimization over networks*, in *Proc. of 2015 IEEE 6th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, Dec., Cancun. 2015, pp. 229–232.

[5] P. Di Lorenzo and G. Scutari, *NEXT: In-network nonconvex optimization*, IEEE Trans. Signal Inf. Process. Netw. 2 (2016), pp. 120–136.

[6] F. Facchinei, G. Scutari, and S. Sagratella, *Parallel selective algorithms for nonconvex big data optimization*, IEEE Trans. Signal Process. 63 (2015), pp. 1874–1889.

[7] J. Fan, Y. Guo, and K. Wang, *Communication-efficient accurate statistical estimation*, arXiv:1906.04870 (2019).

[8] B. Gharesifard and J. Cortés, *When does a digraph admit a doubly stochastic adjacency matrix?*, in *Proc. of the 2010 American Control Conference*, June. 2010, pp. 2440–2445.

[9] D. Jakovetic, J.M.F. Moura, and J. Xavier, *Linear convergence rate of a class of distributed augmented lagrangian algorithms*, IEEE Trans. Autom. Control 60 (2015), pp. 922–936.

[10] D. Jakovetic, *A Unification and Generalization of Exact Distributed First-Order Methods*, IEEE Trans. Signal Inf. Process. Netw. 5 (2019), pp. 31–46.

[11] D. Jakovetic, J. Xavier, and J.M. Moura, *Cooperative convex optimization in networked systems: Augmented Lagrangian algorithms with directed gossip communication*, IEEE Trans. Signal Process. 59 (2011), pp. 3889–3902.

[12] X. Jinming, Y. Tian, Y. Sun, and G. Scutari, *Distributed algorithms for composite optimization: Unified and tight convergence analysis*, arXiv:2002.11534 (2020).

[13] B. Li, S. Cen, Y. Chen, and Y. Chi, *Communication-efficient distributed optimization in networks with gradient tracking and variance reduction*, arXiv:1909.05844v3 (2019).

[14] Z. Li, W. Shi, and M. Yan, *A decentralized proximal-gradient method with network independent step-sizes and separated convergence rates*, IEEE Transactions on Signal Processing 67 (2019), pp. 4494–4506.

[15] Q. Ling, W. Shi, G. Wu, and A. Ribeiro, *DLM: Decentralized linearized alternating direction method of multipliers*, IEEE Trans. Signal Process. 63 (2015), pp. 4051–4064.

[16] C.G. Lopes and A.H. Sayed, *Diffusion Least-Mean Squares Over Adaptive Networks: Formulation and Performance Analysis*, IEEE Trans. Signal Process. 56 (2008), pp. 3122–3136.

[17] H. Lu, R.M. Freund, and Y. Nesterov, *Relatively smooth convex optimization by first-order methods, and applications*, SIAM J. on Optimization 28 (2020), pp. 333–354.

[18] M. Maros and J. Jalden, *PANDA: A Dual Linearly Converging Method for Distributed Optimization Over Time-Varying Undirected Graphs*, 2018 IEEE Conference on Decision and Control (CDC) (2018), pp. 6520–6525.

[19] M. Maros and J. Jalden, *On the Q-linear convergence of Distributed Generalized ADMM under non-strongly convex function components*, IEEE Trans. Signal Inf. Process. Netw. PP (2019), pp. 1–1.

[20] A. Mokhtari, W. Shi, Q. Ling, and A. Ribeiro, *Dqm: Decentralized quadratically approximated alternating direction method of multipliers*, IEEE Transactions on Signal Processing 64 (2016), pp. 5158–5173.

[21] A. Nedić, A. Olshevsky, and W. Shi, *Achieving geometric convergence for distributed optimization over time-varying graphs*, SIAM Journal on Optimization 27 (2017), pp. 2597–2633.

[22] A. Nedic and A. Olshevsky, *Distributed optimization over time-varying directed graphs*, IEEE Trans. Autom. Control 60 (2015), pp. 601–615.

[23] A. Nedić and A. Ozdaglar, *Distributed subgradient methods for multi-agent optimization*, IEEE Trans. Autom. Control 54 (2009), pp. 48–61.

[24] A. Nedić, A. Ozdaglar, and P.A. Parrilo, *Constrained consensus and optimization in multi-agent networks*, IEEE Trans. Autom. Control 55 (2010), pp. 922–938.

[25] A. Nedić, A. Olshevsky, W. Shi, and C.A. Uribe, *Geometrically convergent distributed optimization with uncoordinated step-sizes*, in *2017 American Control Conference*. 2017, pp. 3950–3955.

[26] A. Nedić and A. Ozdaglar, *Convergence rate for consensus with delays*, Journal of Global Optimization 47 (2010), pp. 437–456.

[27] S. Pu, W. Shi, J. Xu, and A. Nedic, *A Push-Pull Gradient Method for Distributed Optimization in Networks*, in *2018 IEEE Conference on Decision and Control (CDC)*. 2018, pp. 3385–3390.

[28] G. Qu and N. Li, *Accelerated Distributed Nesterov Gradient Descent for smooth and strongly convex functions*, in *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Sept. 2016, pp. 209–216.

[29] G. Qu and N. Li, *Harnessing Smoothness to Accelerate Distributed Optimization*, IEEE Control Netw. Syst. 5 (2018), pp. 1245–1260.

[30] A. Rogozin and A. Gasnikov, *Projected gradient method for decentralized optimization over time-varying networks*, arXiv:1911.08527 (2019).

[31] F. Saadatniaki, R. Xin, and U.A. Khan, *Decentralized optimization over time-varying directed graphs with row and column-stochastic matrices*, IEEE Transactions on Automatic Control (2020), pp. 1–1.

[32] K. Scaman, F. Bach, S. Bubeck, Y.T. Lee, and L. Massoulié, *Optimal Algorithms for Smooth and Strongly Convex Distributed Optimization in Networks*, in *Proc. of the 34th International Conference on Machine Learning*, Vol. 70. 2017, pp. 3027–3036.

[33] G. Scutari, F. Facchinei, and L. Lampariello, *Parallel and distributed methods for con-*

strained nonconvex optimization–Part I: Theory, IEEE Trans. Signal Process. 65 (2017), pp. 1929–1944.

[34] G. Scutari and Y. Sun, *Parallel and Distributed Successive Convex Approximation Methods for Big-Data Optimization*, Springer Verlag Series, 2018.

[35] G. Scutari and Y. Sun, *Distributed nonconvex constrained optimization over time-varying digraphs*, Math. Prog. 176 (2019), pp. 497–544.

[36] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan, *Stochastic Convex Optimization*, in *Proc. of the 22nd Annual Conference on Learning Theory (COLT)*, June 18-21, Montreal, Canada. 2009.

[37] O. Shamir, N. Srebro, and T. Zhang, *Communication-Efficient Distributed Optimization using an Approximate Newton-type Method*, in *Proc. of the 31st International Conference on Machine Learning (PMLR)*, Vol. 32. 2014, pp. 1000–1008.

[38] W. Shi, Q. Ling, G. Wu, and W. Yin, *EXTRA: An exact first-order algorithm for decentralized consensus optimization*, SIAM J. Optim. 25 (2015), pp. 944–966.

[39] W. Shi, Q. Ling, G. Wu, and W. Yin, *A proximal gradient algorithm for decentralized composite optimization*, IEEE Trans. Signal Process. 63 (2015), pp. 6013–6023.

[40] W. Shi, Q. Ling, K. Yuan, G. Wu, and W. Yin, *On the linear convergence of the ADMM in decentralized consensus optimization*, IEEE Trans. Signal Process. 62 (2014), pp. 1750–1761.

[41] Y. Sun, A. Daneshmand, and G. Scutari, *Convergence rate of distributed optimization algorithms based on gradient tracking*, arXiv:1905.02637v1 (2019).

[42] Y. Sun, G. Scutari, and D. Palomar, *Distributed nonconvex multiagent optimization over time-varying networks*, in Proc. of the Asilomar Conference on Signals, Systems, and Computers (2016).

[43] Y. Tian, Y. Sun, and G. Scutari, *Achieving linear convergence in distributed asynchronous multi-agent optimization*, IEEE Trans. on Automatic Control (2020).

[44] J. Tsitsiklis, *Problems in decentralized decision making and computation*, Ph.D. dissertation, Dept. of Electrical Engineering and Computer Science, MIT (1984).

[45] A. Wien, *Iterative solution of large linear systems*, Lecture Notes, TU Wien, 2011.

[46] C. Xi, V.S. Mai, R. Xin, E.H. Abed, and U.A. Khan, *Linear convergence in optimization over directed graphs with row-stochastic matrices*, IEEE Trans. Autom. Control 63 (2018), pp. 3558–3565.

[47] C. Xi and U.A. Khan, *ADD-OPT: Accelerated distributed directed optimization*, IEEE Trans. Autom. Control 63 (2018), pp. 1329–1339.

[48] C. Xi and U.A. Khan, *A linear algorithm for optimization over directed graphs with geometric convergence*, IEEE Contr. Syst. Lett. 2 (2018), pp. 315–320.

[49] L. Xiao, S. Boyd, and S. Lall, *A scheme for robust distributed sensor fusion based on average consensus*, in *Proc. of the 4th international symposium on Information processing in sensor networks*, April, Los Angeles, CA. 2005, pp. 63–70.

[50] R. Xin and U. Khan, *Distributed heavy-ball: A generalization and acceleration of first-order methods with gradient tracking*, arXiv:1808.02942 (2018).

[51] R. Xin, D. Jakovetic, and U.A. Khan, *Distributed nesterov gradient methods over arbitrary graphs*, arXiv:1901.06995 (2018).

[52] J. Xu, S. Zhu, Y.C. Soh, and L. Xie, *Augmented distributed gradient methods for multi-agent optimization under uncoordinated constant stepsizes*, in *Proc. of the 54th IEEE Conference on Decision and Control (CDC 2015)*, Dec., Osaka, Japan. 2015, pp. 2055–2060.

[53] J. Xu, S. Zhu, Y.C. Soh, and L. Xie, *Convergence of Asynchronous Distributed Gradient Methods Over Stochastic Networks*, IEEE Trans. Autom. Control 63 (2018), pp. 434–448.

[54] K. Yuan, Q. Ling, and W. Yin, *On the Convergence of Decentralized Gradient Descent*, SIAM J. Optim. 26 (2016), pp. 1835–1854.

[55] K. Yuan, B. Ying, X. Zhao, and A.H. Sayed, *Exact diffusion for distributed optimization and learning—part ii: Convergence analysis*, IEEE Transactions on Signal Processing 67 (2018), pp. 724–739.

[56] J. Zeng and W. Yin, *ExtraPush for convex smooth decentralized optimization over directed networks*, J. Comput. Math. 35 (2017), pp. 383–396.

[57] Y. Zhang and X. Lin, *DiSCO: Distributed Optimization for Self-Concordant Empirical Loss*, in *Proc. of the 32nd International Conference on Machine Learning (PMLR)*, Vol. 37. 2015, pp. 362–370.

[58] Y. Zhang and L. Xiao, *Communication-efficient distributed optimization of self-concordant empirical loss*, in *Large-Scale and Distributed Optimization, number 2227 in Lecture Notes in Mathematics*, chap. 11, Springer, 2018, pp. 289–341.

Supporting Material

I. Proof of Proposition 4.1

We begin introducing some intermediate results.

Lemma I.1. *Consider Problem* (P) *under Assumption A; and SONATA (Algorithm 3) under Assumptions C and E. Then, there holds*

$$U(\mathbf{x}_i^{\nu+\frac{1}{2}}) \leq U(\mathbf{x}_i^\nu) - \alpha\left(\left(1-\frac{\alpha}{2}\right)\widetilde{\mu}_i + \frac{\alpha}{2}\cdot D_i^\ell\right)\|\mathbf{d}_i^\nu\|^2 + \alpha\|\mathbf{d}_i^\nu\|\|\boldsymbol{\delta}_i^\nu\|, \tag{1}$$

with $\boldsymbol{\delta}_i^\nu$ *defined in* (23).

Proof. Consider the Taylor expansion of F:

$$\begin{aligned} F(\mathbf{x}_i^{\nu+\frac{1}{2}}) &= F(\mathbf{x}_i^\nu) + \nabla F(\mathbf{x}_i^\nu)^\top(\alpha\mathbf{d}_i^\nu) + (\alpha\mathbf{d}_i^\nu)^\top\mathbf{H}(\alpha\mathbf{d}_i^\nu), \\ &\overset{(23)}{=} F(\mathbf{x}_i^\nu) + \left(\boldsymbol{\delta}_i^\nu\right)^\top(\alpha\mathbf{d}_i^\nu) + \left(\mathbf{y}_i^\nu\right)^\top(\alpha\mathbf{d}_i^\nu) + (\alpha\mathbf{d}_i^\nu)^\top\mathbf{H}(\alpha\mathbf{d}_i^\nu), \end{aligned} \tag{2}$$

where $\mathbf{H} \triangleq \int_0^1 (1-\theta)\nabla^2 F(\theta\mathbf{x}_i^{\nu+\frac{1}{2}} + (1-\theta)\mathbf{x}_i^\nu)d\theta$.

Invoking the optimality of $\widehat{\mathbf{x}}_i^\nu$, we have

$$G(\mathbf{x}_i^\nu) - G(\widehat{\mathbf{x}}_i^\nu) \geq (\mathbf{d}_i^\nu)^\top\left(\nabla\widetilde{f}_i(\widehat{\mathbf{x}}_i^\nu;\mathbf{x}_i^\nu) + \mathbf{y}_i^\nu - \nabla f_i(\mathbf{x}_i^\nu)\right) = (\mathbf{d}_i^\nu)^\top\left(\mathbf{y}_i^\nu + \widetilde{\mathbf{H}}_i\mathbf{d}_i^\nu\right) \tag{3}$$

where the equality follows from $\nabla\widetilde{f}_i(\mathbf{x}_i^\nu;\mathbf{x}_i^\nu) = \nabla f_i(\mathbf{x}_i^\nu)$ and the integral form of the mean value theorem; and $\widetilde{\mathbf{H}}_i \triangleq \int_0^1 \nabla^2\widetilde{f}_i(\theta\,\widehat{\mathbf{x}}_i^\nu + (1-\theta)\,\mathbf{x}_i^\nu;\mathbf{x}_i^\nu)d\theta$.

Substituting (3) in (2) and using the convexity of G yield

$$\begin{aligned} &F(\mathbf{x}_i^{\nu+\frac{1}{2}}) \\ &\leq F(\mathbf{x}_i^\nu) + (\boldsymbol{\delta}_i^\nu)^\top(\alpha\mathbf{d}_i^\nu) + (\alpha\mathbf{d}_i^\nu)^\top\mathbf{H}(\alpha\mathbf{d}_i^\nu) + \alpha\left(G(\mathbf{x}_i^\nu) - G(\widehat{\mathbf{x}}_i^\nu) - (\mathbf{d}_i^\nu)^\top\widetilde{\mathbf{H}}_i\mathbf{d}_i^\nu\right) \\ &\leq F(\mathbf{x}_i^\nu) + (\boldsymbol{\delta}_i^\nu)^\top(\alpha\mathbf{d}_i^\nu) + \alpha\left(-(\mathbf{d}_i^\nu)^\top\widetilde{\mathbf{H}}_i\mathbf{d}_i^\nu + (\alpha\mathbf{d}_i^\nu)^\top\mathbf{H}(\mathbf{d}_i^\nu)\right) + G(\mathbf{x}_i^\nu) - G(\mathbf{x}_i^{\nu+\frac{1}{2}}). \end{aligned} \tag{4}$$

It remains to bound $\alpha\mathbf{H} - \widetilde{\mathbf{H}}_i$. We proceed as follows:

$$\begin{aligned} &\alpha\mathbf{H} - \widetilde{\mathbf{H}}_i \\ &= \alpha\int_0^1 (1-\theta)\nabla^2 F(\theta\mathbf{x}_i^{\nu+\frac{1}{2}} + (1-\theta)\mathbf{x}_i^\nu)d\theta - \int_0^1 \nabla^2\widetilde{f}_i(\theta\widehat{\mathbf{x}}_i^\nu + (1-\theta)\mathbf{x}_i^\nu;\mathbf{x}_i^\nu)d\theta \\ &\overset{(11b)}{=} \int_0^\alpha (1-\theta/\alpha)\nabla^2 F(\theta\widehat{\mathbf{x}}_i^\nu + (1-\theta)\mathbf{x}_i^\nu)d\theta - \int_0^1 \nabla^2\widetilde{f}_i(\theta\widehat{\mathbf{x}}_i^\nu + (1-\theta)\mathbf{x}_i^\nu;\mathbf{x}_i^\nu)d\theta \\ &\overset{(a)}{\preceq} -\int_0^\alpha (1-\theta/\alpha)\cdot(D_i^\ell)\,\mathbf{I}\,d\theta - \int_0^\alpha (\theta/\alpha)\nabla^2\widetilde{f}_i(\theta\widehat{\mathbf{x}}_i + (1-\theta)\mathbf{x}_i^\nu;\mathbf{x}_i^\nu)d\theta \\ &\quad - \int_\alpha^1 \nabla^2\widetilde{f}_i(\theta\,\widehat{\mathbf{x}}_i^\nu + (1-\theta)\,\mathbf{x}_i^\nu;\mathbf{x}_i^\nu)d\theta \\ &\overset{(b)}{\preceq} -\frac{1}{2}\alpha\,(D_i^\ell)\,\mathbf{I} - \left(1-\frac{\alpha}{2}\right)\widetilde{\mu}_i\,\mathbf{I}, \end{aligned} \tag{5}$$

where in (a) we used $\nabla^2 F(\theta\widehat{\mathbf{x}}_i^\nu + (1-\theta)\mathbf{x}_i^\nu) \preceq -(D_i^\ell)\mathbf{I} + \nabla^2 \widetilde{f}_i(\theta\widehat{\mathbf{x}}_i^\nu + (1-\theta)\mathbf{x}_i^\nu; \mathbf{x}_i^\nu)$ [cf. (15)] while (b) follows from the fact that $\widetilde{f}_i$ is $\widetilde{\mu}_i$-strongly convex (cf. Assumption C). Substituting (5) into (4) completes the proof □

We connect now the individual decreases in (1) with that of the optimality gap $p_{\boldsymbol{\phi}}^\nu$, defined in (77). Notice that

$$\sum_{i=1}^m \phi_i^{\nu+1} U(\mathbf{x}_i^{\nu+1}) \leq \sum_{i=1}^m \sum_{j=1}^m c_{ij}\phi_j^\nu U\left(\mathbf{x}_j^{\nu+\frac{1}{2}}\right) = \sum_{i=1}^m \phi_i^\nu U(\mathbf{x}_i^{\nu+\frac{1}{2}}), \tag{6}$$

due to the convexity of U, column-stochasticity of $\{c_{ij}^\nu\}_{i,j}$ and $\sum_{j=1}^m c_{ij}^\nu \phi_j^\nu / \phi_i^{\nu+1} = 1$, for all $i = 1, \ldots, m$. Summing (1) over $i = 1, \ldots m$, and using (6), we obtain

$$\begin{aligned} p_{\boldsymbol{\phi}}^{\nu+1} &\leq p_{\boldsymbol{\phi}}^\nu + \sum_{i=1}^m \phi_i^\nu \left\{ \alpha \|\mathbf{d}_i^\nu\| \|\boldsymbol{\delta}_i^\nu\| - \alpha\left(1 - \frac{\alpha}{2}\right)\widetilde{\mu}_i \|\mathbf{d}_i^\nu\|^2 - \frac{D_i^\ell}{2}\alpha^2 \|\mathbf{d}_i^\nu\|^2 \right\} \\ &\overset{(a)}{\leq} p_{\boldsymbol{\phi}}^\nu - \left(\left(1 - \frac{\alpha}{2}\right)\widetilde{\mu}_{\mathrm{mn}} + \frac{\alpha D_{\mathrm{mn}}}{2} - \frac{1}{2}\epsilon_{opt} \right) \alpha \sum_{i=1}^m \phi_i^\nu \|\mathbf{d}_i^\nu\|^2 + \frac{1}{2}\epsilon_{opt}^{-1}\, \alpha \cdot \phi_{ub} \cdot \|\boldsymbol{\delta}^\nu\|^2, \end{aligned} \tag{7}$$

where in (a) we used Young's inequality, with $\epsilon_{opt} > 0$ satisfying

$$\left(1 - \frac{\alpha}{2}\right)\widetilde{\mu}_{\mathrm{mn}} + \frac{\alpha D_{\mathrm{mn}}^L}{2} - \frac{1}{2}\epsilon_{opt} > 0. \tag{8}$$

Next we lower bound $\|\mathbf{d}^\nu\|^2$ in terms of the optimality gap.

Lemma I.2. *In the setting of Lemma 3.1, there holds:*

$$\alpha \sum_{i=1}^m \phi_i^\nu \|\mathbf{d}_i^\nu\|^2 \geq \frac{\mu}{D_{\mathrm{mx}}^2} \left(p_{\boldsymbol{\phi}}^{\nu+1} - (1-\alpha) p_{\boldsymbol{\phi}}^\nu - \frac{\alpha}{\mu} \sum_{i=1}^m \phi_i^\nu \|\boldsymbol{\delta}_i^\nu\|^2 \right) \tag{9}$$

with D_{mx} defined in (24).

Proof. Invoking the optimality condition of $\widehat{\mathbf{x}}_i^\nu$, yields

$$G(\mathbf{x}^\star) - G(\widehat{\mathbf{x}}_i^\nu) \geq -(\mathbf{x}^\star - \widehat{\mathbf{x}}_i^\nu)^\top \Big(\nabla \widetilde{f}_i(\widehat{\mathbf{x}}_i^\nu; \mathbf{x}_i^\nu) + \mathbf{y}_i^\nu - \nabla f_i(\mathbf{x}_i^\nu) \Big). \tag{10}$$

Using the μ-strong convexity of F, we can write

$$\begin{aligned} U(\mathbf{x}^\star) &\geq U(\widehat{\mathbf{x}}_i^\nu) + G(\mathbf{x}^\star) - G(\widehat{\mathbf{x}}_i^\nu) + \nabla F(\widehat{\mathbf{x}}_i^\nu)^\top (\mathbf{x}^\star - \widehat{\mathbf{x}}_i^\nu) + \frac{\mu}{2}\|\mathbf{x}^\star - \widehat{\mathbf{x}}_i^\nu\|^2 \\ &\overset{(36)}{\geq} U(\widehat{\mathbf{x}}_i^\nu) + \Big(\nabla F(\widehat{\mathbf{x}}_i^\nu) - \nabla \widetilde{f}_i(\widehat{\mathbf{x}}_i^\nu; \mathbf{x}_i^\nu) - \big(\mathbf{y}_i^\nu - \nabla f_i(\mathbf{x}_i^\nu)\big) \Big)^\top (\mathbf{x}^\star - \widehat{\mathbf{x}}_i^\nu) + \frac{\mu}{2}\|\mathbf{x}^\star - \widehat{\mathbf{x}}_i^\nu\|^2 \\ &= U(\widehat{\mathbf{x}}_i^\nu) + \frac{\mu}{2} \Big\| \mathbf{x}^\star - \widehat{\mathbf{x}}_i^\nu + \frac{1}{\mu}\Big(\nabla F(\widehat{\mathbf{x}}_i^\nu) - \nabla \widetilde{f}_i(\widehat{\mathbf{x}}_i^\nu; \mathbf{x}_i^\nu) - \big(\mathbf{y}_i^\nu - \nabla f_i(\mathbf{x}_i^\nu)\big) \Big) \Big\|^2 \\ &\quad - \frac{1}{2\mu} \Big\| \nabla F(\widehat{\mathbf{x}}_i^\nu) - \nabla \widetilde{f}_i(\widehat{\mathbf{x}}_i^\nu; \mathbf{x}_i^\nu) - \big(\mathbf{y}_i^\nu - \nabla f_i(\mathbf{x}_i^\nu)\big) \Big\|^2 \end{aligned}$$

$$
\begin{aligned}
&\geq U(\widehat{\mathbf{x}}_i^\nu) - \frac{1}{2\mu}\left\|\nabla F(\widehat{\mathbf{x}}_i^\nu) \pm \nabla F(\mathbf{x}_i^\nu) - \nabla \widetilde{f}_i(\widehat{\mathbf{x}}_i^\nu;\mathbf{x}_i^\nu) - \big(\mathbf{y}_i^\nu - \nabla f_i(\mathbf{x}_i^\nu)\big)\right\|^2 \\
&\geq U(\widehat{\mathbf{x}}_i^\nu) - \frac{1}{\mu}\left\|\nabla F(\widehat{\mathbf{x}}_i^\nu) - \nabla F(\mathbf{x}_i^\nu) + \nabla f_i(\mathbf{x}_i^\nu) - \nabla \widetilde{f}_i(\widehat{\mathbf{x}}_i^\nu;\mathbf{x}_i^\nu)\right\|^2 - \frac{1}{\mu}\|\boldsymbol{\delta}_i^\nu\|^2 \\
&= U(\widehat{\mathbf{x}}_i^\nu) - \frac{1}{\mu}\left\|\int_0^1 \Big(\nabla^2 F(\theta\widehat{\mathbf{x}}_i^\nu + (1-\theta)\mathbf{x}_i^\nu) - \nabla^2 \widetilde{f}_i(\theta\widehat{\mathbf{x}}_i^\nu + (1-\theta)\mathbf{x}_i^\nu;\mathbf{x}_i^\nu)\Big)\,(\mathbf{d}_i^\nu)\,\mathrm{d}\theta\right\|^2 - \frac{1}{\mu}\|\boldsymbol{\delta}_i^\nu\|^2 \\
&\geq U(\widehat{\mathbf{x}}_i^\nu) - \frac{D_i^2}{\mu}\,\|\mathbf{d}_i^\nu\|^2 - \frac{1}{\mu}\|\boldsymbol{\delta}_i^\nu\|^2,
\end{aligned}
$$

where $D_i = \max\{|D_i^\ell|, |D_i^u|\}$.

Rearranging the terms and summing over $i = 1, \ldots, m$, yields

$$
\sum_{i=1}^m \phi_i^\nu \|\mathbf{d}_i^\nu\|^2 \geq \frac{\mu}{D_{\mathrm{mx}}^2}\left(\sum_{i=1}^m \phi_i^\nu\big(U(\widehat{\mathbf{x}}_i^\nu) - U(\mathbf{x}^\star)\big) - \frac{1}{\mu}\sum_{i=1}^m \phi_i^\nu \|\boldsymbol{\delta}_i^\nu\|^2\right). \tag{11}
$$

Using (27) in conjunction with $U(\mathbf{x}_i^{\nu+\frac{1}{2}}) \leq \alpha U(\widehat{\mathbf{x}}_i^\nu) + (1-\alpha)U(\mathbf{x}_i^\nu)$ leads to

$$
\alpha \sum_{i=1}^m \phi_i^\nu \left(U(\widehat{\mathbf{x}}_i^\nu) - U(\mathbf{x}^\star)\right) \geq p_{\boldsymbol{\phi}}^{\nu+1} - (1-\alpha)p_{\boldsymbol{\phi}}^\nu. \tag{12}
$$

Combining (11) with (12) yields the desired result (9). □

As last step, we upper bound $\|\boldsymbol{\delta}^\nu\|^2$ in (33) in terms of the consensus errors $\|\mathbf{x}_\perp^\nu\|^2$ and $\|\mathbf{y}_\perp^\nu\|^2$.

Lemma I.3. *The tracking error* $\|\boldsymbol{\delta}^\nu\|^2$ *can be bounded as*

$$
\|\boldsymbol{\delta}^\nu\|^2 \leq 8L_{\mathrm{mx}}^2\|\mathbf{x}_{\boldsymbol{\phi},\perp}^\nu\|^2 + 2\|\mathbf{y}_{\boldsymbol{\phi},\perp}^\nu\|^2., \tag{13}
$$

where L_{mx} *is defined in* (4).

Proof.

$$
\begin{aligned}
\|\boldsymbol{\delta}^\nu\|^2 &\overset{(23)}{=} \sum_{i=1}^m \|\nabla F(\mathbf{x}_i^\nu) \pm \bar{\mathbf{y}}_{\boldsymbol{\phi}}^\nu - \mathbf{y}_i^\nu\|^2 \\
&\overset{(74)}{=} \frac{1}{m^2}\sum_{i=1}^m \left\|\sum_{j=1}^m \nabla f_j(\mathbf{x}_i^\nu) - \sum_{j=1}^m \nabla f_j(\mathbf{x}_j^\nu) + m\cdot\bar{\mathbf{y}}_{\boldsymbol{\phi}}^\nu - m\cdot\mathbf{y}_i^\nu\right\|^2 \\
&\overset{A2,(4)}{\leq} \frac{1}{m^2}\sum_{i=1}^m\left(2m\sum_{j=1}^m L_{\mathrm{mx}}^2\|\mathbf{x}_i^\nu - \mathbf{x}_j^\nu\|^2 + 2m^2\|\bar{\mathbf{y}}_{\boldsymbol{\phi}}^\nu - \mathbf{y}_i^\nu\|^2\right) \\
&\leq 8L_{\mathrm{mx}}^2\|\mathbf{x}_{\boldsymbol{\phi},\perp}^\nu\|^2 + 2\|\mathbf{y}_{\boldsymbol{\phi},\perp}^\nu\|^2.
\end{aligned}
$$

□

The linear convergence of the optimality gap up to consensus errors as stated in Proposition follows readily multiplying (9) by $\left(1-\frac{\alpha}{2}\right)\widetilde{\mu}_{\mathrm{mn}} + \frac{\alpha D_{\mathrm{mn}}}{2} - \frac{1}{2}\epsilon_{opt}$ and adding with (7) to cancel out $\|\mathbf{d}^\nu\|$, and using (39) to bound $\|\boldsymbol{\delta}^\nu\|^2$.

II. Proof of Theorem 4.5

Following the same steps as in the proof of Theorem 3.9, we derive the optimal ϵ_{opt} appearing in $\eta(\alpha)$ and $\sigma(\alpha)$:

$$\epsilon^{\star}_{opt} = \left(1 - \frac{\alpha}{2}\right)\widetilde{\mu}_{\mathrm{mn}} + \alpha D^{\ell}_{\mathrm{mn}}/2, \tag{1}$$

where α must satisfy

$$\alpha < 2\widetilde{\mu}_{\mathrm{mn}}/(\widetilde{\mu}_{\mathrm{mn}} - D^{\ell}_{\mathrm{mn}}). \tag{2}$$

Setting $\epsilon_{opt} = \epsilon^{\star}_{opt}$ and denoting the corresponding $\mathcal{P}(\alpha, z)$ as $\mathcal{P}^{\star}(\alpha, z)$, the expression of $\mathcal{P}^{\star}(\alpha, 1)$ reads

$$\begin{aligned}\mathcal{P}^{\star}(\alpha, 1) \triangleq{} & G^{\star}_{P}(\alpha) \cdot C_1 \cdot 8\phi_{ub} L^2_{\mathrm{mx}} \cdot \frac{2c_0^2\rho_B^2}{(1-\rho_B)^2}\alpha^2 \\ & + \left(G^{\star}_{P}(\alpha) \cdot 2\phi_{ub} \cdot C_1 + C_2\right) \cdot 2m\phi_{lb}^{-2} L^2_{\mathrm{mx}} \cdot \frac{2c_0^2\rho_B^2}{(1-\rho_{\bar{B}})^2}\alpha^2 \\ & + \left(G^{\star}_{P}(\alpha) \cdot 2\phi_{ub} \cdot C_1 + C_2\right) \cdot 8m\phi_{lb}^{-2} L^2_{\mathrm{mx}} \cdot \frac{4c_0^4\rho_B^4}{(1-\rho_{\bar{B}})^4}\alpha^2,\end{aligned} \tag{3}$$

where

$$G^{\star}_{P}(\alpha) \triangleq \frac{\frac{D^2_{\mathrm{mx}}}{\mu} + \frac{1}{\mu} \cdot \left(\left(1 - \frac{\alpha}{2}\right)\widetilde{\mu}_{\mathrm{mn}} + \frac{D^{\ell}_{\mathrm{mn}}}{2}\alpha\right)^2}{\left(\left(1 - \frac{\alpha}{2}\right)\widetilde{\mu}_{\mathrm{mn}} + \frac{D^{\ell}_{\mathrm{mn}}}{2}\alpha\right)^2}. \tag{4}$$

Since $\mathcal{P}^{\star}(\bullet, 1)$ is continuous and monotonically increasing on $(0, 2\widetilde{\mu}_{\mathrm{mn}}/(\widetilde{\mu}_{\mathrm{mn}} - D^{\ell}_{\mathrm{mn}})$, with $\mathcal{P}^{\star}(0, 1) = 0$. A upperbound of α can be found by setting

$$\begin{align}\alpha < \alpha_2 \triangleq \Bigg(& G^{\star}_{P}\left(\frac{\widetilde{\mu}_{\mathrm{mn}}}{\widetilde{\mu}_{\mathrm{mn}} - D^{\ell}_{\mathrm{mn}}}\right) \cdot C_1 \cdot 8\phi_{ub} L^2_{\mathrm{mx}} \cdot \frac{2c_0^2\rho_B^2}{(1-\rho_B)^2}\alpha^2 \tag{5} \\ & + \left(G^{\star}_{P}\left(\frac{\widetilde{\mu}_{\mathrm{mn}}}{\widetilde{\mu}_{\mathrm{mn}} - D^{\ell}_{\mathrm{mn}}}\right) \cdot 2\phi_{ub} \cdot C_1 + C_2\right) \cdot 2m\phi_{lb}^{-2} L^2_{\mathrm{mx}} \cdot \frac{2c_0^2\rho_B^2}{(1-\rho_{\bar{B}})^2}\alpha^2 \tag{6} \\ & + \left(G^{\star}_{P}\left(\frac{\widetilde{\mu}_{\mathrm{mn}}}{\widetilde{\mu}_{\mathrm{mn}} - D^{\ell}_{\mathrm{mn}}}\right) \cdot 2\phi_{ub} \cdot C_1 + C_2\right) \cdot 8m\phi_{lb}^{-2} L^2_{\mathrm{mx}} \cdot \frac{4c_0^4\rho_B^4}{(1-\rho_{\bar{B}})^4}\Bigg)^{-1/2}. \tag{7}\end{align}$$

Therefore, a valid $\bar{\alpha}$ is $\bar{\alpha} = \min\{\widetilde{\mu}_{\mathrm{mn}}/(\widetilde{\mu}_{\mathrm{mn}} - D^{\ell}_{\mathrm{mn}}), \alpha_2\}$.

III. Explicit expression of the linear rate in the time-varying directed network setting

The following theorem provides an explicit expression of the convergence rate in Theorem 4.5, in terms of the step-size α; the constants J and $A_{\frac{1}{2}}$ therein are defined in (8) and (5) with $\theta = 1/2$, respectively.

Theorem III.1. *In the setting of Theorem 4.5, suppose that the step-size α satisfies $\alpha \in (0, \alpha_{\mathrm{mx}})$, with $\alpha_{\mathrm{mx}} \triangleq \min\{(1-\rho_B)/A_{\frac{1}{2}}, , \widetilde{\mu}_{\mathrm{mn}}/(\widetilde{\mu}_{\mathrm{mn}} - D^{\ell}_{\mathrm{mn}}), 1\}$. Then $\{U(\mathbf{x}_i^{\nu})\}$ converges to $U^\star$ at the R-linear rate $\mathcal{O}(z^\nu)$, for all $i = 1, \ldots, m$, where*

$$z = \begin{cases} 1 - J \cdot \alpha, & \text{if } \alpha \in (0, \min\{\alpha^*, \alpha_{\mathrm{mx}}\}), \\ \rho_B + A_{\frac{1}{2}}\alpha, & \text{if } \alpha \in [\min\{\alpha^*, \alpha_{\mathrm{mx}}\}, \alpha_{\mathrm{mx}}). \end{cases} \tag{1}$$

Proof. The proof follows similar steps as the proof of Theorem 3.10. For sake of simplicity, we used the same notation as therein. We find the smallest z satisfying (89) such that $\mathcal{P}(\alpha, z) < 1$, for $\alpha \in (0, \alpha_{\mathrm{mx}})$, and $\alpha_{\mathrm{mx}} \in (0,1)$ to be determined[recall that $\mathcal{P}(\alpha, z)$ is defined in (95)].

Using exactly the same argument as Theorem 3.10 we have the following two conditions on z:

$$z \geq \sigma(\alpha) + \frac{(\theta \cdot \alpha) \cdot \left(\left(1 - \frac{\alpha}{2}\right)\widetilde{\mu}_{\mathrm{mn}} + \frac{D^{\ell}_{\mathrm{mn}}}{2}\alpha - \frac{1}{2}\epsilon_{opt}\right)}{\frac{D^2_{\mathrm{mx}}}{\mu} + \left(1 - \frac{\alpha}{2}\right)\widetilde{\mu}_{\mathrm{mn}} + \frac{D^{\ell}_{\mathrm{mn}}}{2}\alpha - \frac{1}{2}\epsilon_{opt}} \tag{2}$$

for some $\theta \in (0,1)$; and

$$\begin{aligned} & G_P^\star(\alpha) \cdot \theta^{-1} \cdot G_X(z) \cdot C_1 \cdot 8\phi_{ub} L^2_{\mathrm{mx}} \cdot \rho_B^2 \cdot \alpha^2 \\ & + \left(G_P^\star(\alpha) \cdot \theta^{-1} \cdot 2\phi_{ub} \cdot C_1 + C_2\right) \cdot G_Y(z) \cdot 2m\phi_{lb}^{-2} L^2_{\mathrm{mx}} \cdot \rho_B^2 \cdot \alpha^2 \\ & + \left(G_P^\star(\alpha) \cdot \theta^{-1} \cdot 2\phi_{ub} \cdot C_1 + C_2\right) \cdot G_Y(z) \cdot 8m\phi_{lb}^{-2} L^2_{\mathrm{mx}} \cdot G_X(z) \cdot \rho_B^4 \cdot \alpha^2 < 1. \end{aligned} \tag{3}$$

Using the fact that $G_P^\star(\alpha)$ is monotonically increasing on $\alpha \in (0, 2\widetilde{\mu}_{\mathrm{mn}}/(\widetilde{\mu}_{\mathrm{mn}} - D^{\ell}_{\mathrm{mn}}))$, and restricting $\alpha \in (0, \widetilde{\mu}_{\mathrm{mn}}/(\widetilde{\mu}_{\mathrm{mn}} - D^{\ell}_{\mathrm{mn}})]$, a sufficient condition for (3) is

$$\alpha \leq \alpha(z) \triangleq \left(A_{1,\theta}\frac{1}{z - \rho_B} + A_{2,\theta}\frac{1}{z - \rho_B} + A_{3,\theta}\frac{1}{(z-\rho_B)^2}\right)^{-1/2}, \tag{4}$$

where $A_{1,\theta}$, $A_{2,\theta}$ and $A_{3,\theta}$ are constants defined as

$$\begin{aligned} A_{1,\theta} &\triangleq G_P^\star\left(\frac{\widetilde{\mu}_{\mathrm{mn}}}{\widetilde{\mu}_{\mathrm{mn}} - D^{\ell}_{\mathrm{mn}}}\right) \cdot \theta^{-1} \cdot C_1 \cdot 8\phi_{ub} L^2_{\mathrm{mx}} \cdot \frac{2c_0^2\rho_B^2}{1 - \rho_B} \\ A_{2,\theta} &\triangleq \left(G_P^\star\left(\frac{\widetilde{\mu}_{\mathrm{mn}}}{\widetilde{\mu}_{\mathrm{mn}} - D^{\ell}_{\mathrm{mn}}}\right) \cdot \theta^{-1} \cdot 2\phi_{ub} \cdot C_1 + C_2\right) \cdot 2m\phi_{lb}^{-2} L^2_{\mathrm{mx}} \cdot \frac{2c_0^2\rho_B^2}{1 - \rho_B} \\ A_{3,\theta} &\triangleq \left(G_P^\star\left(\frac{\widetilde{\mu}_{\mathrm{mn}}}{\widetilde{\mu}_{\mathrm{mn}} - D^{\ell}_{\mathrm{mn}}}\right) \cdot \theta^{-1} \cdot 2\phi_{ub} \cdot C_1 + C_2\right) \cdot 8m\phi_{lb}^{-2} L^2_{\mathrm{mx}} \cdot \frac{4c_0^4\rho_B^4}{(1 - \rho_B)^2}. \end{aligned}$$

Lower bounding $z - \rho_B$ by $(z - \rho_B)^2$ we obtain

$$z \geq \rho_B + A_\theta \alpha, \quad \text{with} \quad A_\theta \triangleq \sqrt{A_{1,\theta} + A_{2,\theta} + A_{3,\theta}}. \tag{5}$$

Letting $\epsilon_{opt} = \epsilon^\star_{opt}$ in (2), the condition reduces to

$$z \geq 1 - \frac{\widetilde{\mu}_{\mathrm{mn}} - \frac{\alpha}{2}(\widetilde{\mu}_{\mathrm{mn}} - D^{\ell}_{\mathrm{mn}})}{\frac{2D^2_{\mathrm{mx}}}{\mu} + \widetilde{\mu}_{\mathrm{mn}} - \frac{\alpha}{2}(\widetilde{\mu}_{\mathrm{mn}} - D^{\ell}_{\mathrm{mn}})} \cdot (1 - \theta)\alpha. \tag{6}$$

Therefore, the overall convergence rate can be upper bounded by $\mathcal{O}(\bar{z}^{\nu})$, where

$$\bar{z} = \inf_{\theta\in(0,1)} \max\left\{\rho_B + A_\theta\alpha, 1 - \frac{\widetilde{\mu}_{\text{mn}} - \frac{\alpha}{2}(\widetilde{\mu}_{\text{mn}} - D^{\ell}_{\text{mn}})}{\frac{2D^2_{\text{mx}}}{\mu} + \widetilde{\mu}_{\text{mn}} - \frac{\alpha}{2}(\widetilde{\mu}_{\text{mn}} - D^{\ell}_{\text{mn}})} \cdot (1-\theta)\alpha\right\}, \tag{7}$$

with A_θ defined in (5).

Finally, we further simplify (7). Letting $\theta = 1/2$ and using $\alpha \in (0, \widetilde{\mu}_{\text{mn}}/(\widetilde{\mu}_{\text{mn}} - D^{\ell}_{\text{mn}})]$, the second term in the max of (7) can be upper bounded by

$$1 - \underbrace{\frac{\widetilde{\mu}_{\text{mn}}\mu}{4D^2_{\text{mx}} + \widetilde{\mu}_{\text{mn}}\mu} \cdot \frac{1}{2}}_{\triangleq J}\alpha. \tag{8}$$

The condition $\bar{z} < 1$ imposes the following upper bound on α: $\alpha < \alpha_{\text{mx}} = \min\{(1-\rho_B)/A_{\frac{1}{2}}, \widetilde{\mu}_{\text{mn}}/(\widetilde{\mu}_{\text{mn}} - D^{\ell}_{\text{mn}}), 1\}$. Eq. (7) then simplifies to (1) with $\alpha^* = (1-\rho_B)/(A_{\frac{1}{2}} + J)$ that equates $1 - J\alpha$ and $\rho_B + A_{\frac{1}{2}}\alpha$. □

IV. Rate estimate using linearization surrogate (62) (time-varying directed network case)

Corollary IV.1 (Linearization surrogates)**.** *In the setting of Theorem III.1, let* $\{\mathbf{x}^{\nu}\}$ *be the sequence generated by SONATA (Algorithm 3), using the surrogates* (62) *and step-size* $\alpha = c\cdot\alpha_{\text{mx}}$, $c \in (0,1)$, *where* $\alpha_{\text{mx}} = \min\{1, (1-\rho_B)^2/(C_M\cdot\kappa_g(1+\beta/L)^2)\}$ *and* C_M *is a constant defined in* (6)*. The number of iterations (communications) needed for* $U(\mathbf{x}_i^{\nu}) - U^\star \leq \epsilon$, $i \in [m]$, *is*

$$\mathcal{O}\left(\kappa_g \log(1/\epsilon)\right), \qquad \textit{if} \quad \frac{\rho_B}{(1-\rho_B)^2} \leq \frac{1}{C_M \cdot \kappa_g\left(1+\frac{\beta}{L}\right)^2}, \tag{1}$$

$$\mathcal{O}\left(\frac{\left(\kappa_g + \beta/\mu\right)^2\rho_B}{(1-\rho_B)^2}\log(1/\epsilon)\right), \qquad \textit{otherwise.} \tag{2}$$

Proof. According to Theorem III.1, the rate z can be bounded as

$$z \leq \max\{z_1, z_2\}, \qquad \text{with} \qquad z_1 \triangleq 1 - \alpha\cdot J \;\; \text{and} \;\; z_2 \triangleq \rho_B + A_{\frac{1}{2}}\alpha, \tag{3}$$

where J and $A_{\frac{1}{2}}$ are defined in (8) and (5), respectively.

The proof consists in bounding properly z_1 and z_2 based upon the surrogate (62) postulated in the corollary. We begin particularizing the expressions of J and $A_{\frac{1}{2}}$. Since $\nabla^2\widetilde{f}_i(\mathbf{x}_i;\mathbf{x}_i^{\nu}) = L$, one can set $\widetilde{\mu}_{\text{mn}} = L$, and (15) holds with $D^{\ell}_{\text{mn}} = 0$ and $D_{\text{mx}} = L-\mu$. Furthermore, by Assumption 2.1, it follows that $\beta \geq \lambda_{\max}(\nabla^2 f_i(\mathbf{x})) - L$, for all $\mathbf{x} \in \mathcal{K}$; hence, one can set $L_{\text{mx}} = L + \beta$. Next, we will substitute the above values into the expressions of J and $A_{\frac{1}{2}}$.

To do so, we need to particularize first the quantities $G_P^\star\left(\frac{\widetilde{\mu}_{\text{mn}}}{\widetilde{\mu}_{\text{mn}} - D^{\ell}_{\text{mn}}}\right)$ [cf. (4)], C_1 and C_2 [cf. (93)]:

$$
\begin{aligned}
&G_P^\star\left(\frac{\widetilde{\mu}_{\text{mn}}}{\widetilde{\mu}_{\text{mn}} - D_{\text{mn}}^\ell}\right) = G_P^\star\left(1\right) = \frac{4(L-\mu)^2 + L^2}{\mu L^2},\\
&C_1 = \frac{6}{\mu \phi_{lb} L^2}\left((2L-\mu)^2 + 4(L+\beta)^2\right), \quad \text{and} \quad C_2 = \frac{4}{L^2}.
\end{aligned}
$$

Accordingly, the expressions of J and $A_{\frac{1}{2}}$ read:

$$
J = \frac{1}{2}\frac{\kappa_g}{4(\kappa_g - 1)^2 + \kappa_g} \in \left[\frac{1}{8\kappa_g}, \frac{1}{2}\right], \tag{4}
$$

and

$$
\begin{aligned}
&(A_{\frac{1}{2}})^2\\
=&G_P^\star(1)\cdot 2\cdot C_1 \cdot 8\phi_{ub}\cdot L_{\text{mx}}^2 \cdot \frac{2c_0^2\rho_B^2}{1-\rho_B}\\
&+ (G_P^\star(1)\cdot 2\cdot C_1\cdot 2\phi_{ub} + C_2)\cdot 2m\phi_{lb}^{-2} L_{\text{mx}}^2 \cdot \frac{2c_0^2\rho_B^2}{1-\rho_B}\\
&+ (G_P^\star(1)\cdot 2\cdot C_1\cdot 2\phi_{ub} + C_2)\cdot 8m\phi_{lb}^{-2} L_{\text{mx}}^2 \cdot \frac{4c_0^4\rho_B^4}{(1-\rho_B)^2}\\
\leq& (G_P^\star(1)\cdot 2\cdot C_1\cdot 12\phi_{ub} + C_2)\cdot 8m\phi_{lb}^{-2} L_{\text{mx}}^2 \cdot \frac{4c_0^4\rho_B^2}{(1-\rho_B)^2}\\
\leq& \left[\frac{4(L-\mu)^2 + L^2}{\mu L^2}\cdot \frac{12}{\mu\phi_{lb}L^2}\left((2L-\mu)^2 + 4(L+\beta)^2\right)\cdot 12\phi_{ub} + \frac{4}{L^2}\right]\\
&\qquad\qquad\qquad\qquad\cdot 8m\phi_{lb}^{-2}\left(L+\beta\right)^2 \cdot \frac{4c_0^4\rho_B^2}{(1-\rho_B)^2}\\
\leq& C_M^2\cdot\kappa_g^2\left(1+\frac{\beta}{L}\right)^4\cdot\frac{\rho_B^2}{(1-\rho_B)^2},
\end{aligned} \tag{5}
$$

where

$$
C_M \triangleq 608\cdot\phi_{lb}^{-1}\cdot c_0\,\sqrt{\frac{\phi_{ub}}{\phi_{lb}}\cdot m}, \tag{6}
$$

and in the first inequality we have used the fact that $\phi_{lb} < 1$ and $c_0 > 1$, and the last inequality holds since $\kappa_g \geq 1$ and $\frac{\phi_{ub}}{\phi_{lb}} \geq 1$. Using the above expressions, in the sequel we upperbound z_1 and z_2.

By (5), we have

$$
z_2 \leq \bar{z}_2 \triangleq \rho_B + \alpha M\cdot\frac{\rho_B}{1-\rho_B}, \quad \text{with} \quad M \triangleq C_M\cdot\kappa_g(1+\beta/L)^2. \tag{7}
$$

Since $\alpha\in(0,1]$ must be chosen so that $z\in(0,1]$, we impose $\max\{z_1,\bar{z}_2\} < 1$, implying $\alpha \leq \min\{J^{-1}, (1-\rho_B)^2/(M\rho_B), 1\}$. Since $J^{-1} > 1$ [cf. (4)], the condition on α reduces to $\alpha \leq \alpha_{\text{mx}} \triangleq \min\{1, (1-\rho_B)^2/(M\rho_B)\} < 1$. Choose $\alpha = c\cdot\alpha_{\text{mx}}$, for some given $c\in(0,1)$. Depending on the value of ρ_B, either $\alpha_{\text{mx}} = 1$ or $\alpha_{\text{mx}} = (1-\rho_B)^2/(M\rho_B)$.

- **Case I:** $\alpha_{\text{mx}} = 1$. This corresponds to the case $M\rho_B \leq (1-\rho_B)^2$. Note that, we also have $\rho_B \leq 1/C_M$, otherwise $M\rho_B \geq C_M\,\kappa_g\,\rho_B > 1 > (1-\rho_B)^2$. In this setting,

$\alpha = c \cdot \alpha_{\text{mx}} = c$, and

$$\begin{aligned} z_1 &= 1 - c \cdot J, \\ \bar{z}_2 &= \rho_B + cM \cdot \frac{\rho_B}{1-\rho_B} \overset{(a)}{\leq} 1 - (1-c)(1-\rho_B) \\ &\overset{(b)}{\leq} 1 - (1-c)\left(1 - \frac{1}{C_M}\right), \end{aligned}$$

where in (a) we used $M\rho_B \leq (1-\rho_B)^2$ and (b) follows from $\rho_B \leq 1/C_M$.

Therefore, z can be bounded as

$$\begin{aligned} z \leq \max\{z_1, \bar{z}_2\} &\leq 1 - c \cdot (1-c)\left(1 - \frac{1}{C_M}\right) \cdot J \\ &\leq 1 - c \cdot (1-c)\left(1 - \frac{1}{C_M}\right) \cdot \frac{1}{8\kappa_g}. \end{aligned} \tag{8}$$

- **Case II:** $\alpha_{\text{mx}} = (1-\rho_B)^2/(M\rho_B)$. This corresponds to $M\rho_B > (1-\rho_B)^2$. We have $\alpha = c \cdot \alpha_{\text{mx}}$,

$$\begin{aligned} z_1 =& 1 - \frac{J\,c}{M\rho_B} \cdot (1-\rho_B)^2, \\ \bar{z}_2 =& 1 - (1-c)\,(1-\rho_B)\,. \end{aligned}$$

Now we can bound z. Since $Jc/(M\rho_B) < 1$ (by the same reasoning as in proof of Proposition 3.12),

$$\begin{aligned} z \leq \max\{z_1, \bar{z}_2\} &\leq 1 - \frac{Jc}{M\rho_B} \cdot (1-c)\,(1-\rho_B)^2 \\ &\overset{(4)}{\leq} 1 - \frac{c\,(1-c)}{8C_M} \cdot \frac{(1-\rho_B)^2}{\kappa_g^2(1+\beta/L)^2\rho_B} \\ &= 1 - \frac{c\,(1-c)}{8C_M} \cdot \frac{(1-\rho_B)^2}{(\kappa_g+\beta/\mu)^2\rho_B}. \end{aligned} \tag{9}$$

□

V. Rate estimate using local f_i (63) (time-varying directed network case)

Corollary V.1 (Local f_i, $\beta \leq \mu$). *Instate assumptions of Theorem III.1 and suppose $\beta \leq \mu$. Consider SONATA (Algorithm 3) using the surrogates* (63) *and stepsize $\alpha = c \cdot \alpha_{\text{mx}}$, $c \in (0,1)$, with $\alpha_{\text{mx}} = \min\{1, (1-\rho_B)^2/(\tilde{M}_2\rho_B)\}$ where $\tilde{M}_2 = 1087\tilde{C}_M\left(1+\frac{\beta}{\mu}\right)^2\left(\kappa_g+\frac{\beta}{\mu}\right)^2$ and the constant $\tilde{C}_M$ is defined in* (7)*. The number of iterations (communications) needed for $U(\mathbf{x}_i^\nu) - U^\star \leq \epsilon$, $i \in [m]$, is*

$$\mathcal{O}\left(1 \cdot \log(1/\epsilon)\right), \qquad \textit{if} \quad \frac{\rho_B}{(1-\rho_B)^2} \leq \frac{1}{1087 \cdot \tilde{C}_M \cdot \left(1+\frac{\beta}{\mu}\right)^2\left(\kappa_g+\frac{\beta}{\mu}\right)^2}, \tag{1}$$

$$\mathcal{O}\left(\frac{\kappa_g^2\rho_B}{(1-\rho_B)^2}\log(1/\epsilon)\right), \qquad \textit{otherwise.} \qquad (2)$$

Corollary V.2 (Local f_i, $\beta > \mu$)**.** *Instate assumptions of Theorem III.1 and suppose $\beta > \mu$. Consider SONATA (Algorithm 3) using the surrogates* (63) *and stepsize $\alpha = c \cdot \alpha_{\mathrm{mx}}$, $c \in (0,1)$, where $\alpha_{\mathrm{mx}} = \min\{1, (1-\rho_B)^2/(\widetilde{M}_1\rho_B)\}$ with $\widetilde{M}_1 = 1428\tilde{C}_M\left(1+\frac{L}{\beta}\right)\left(\kappa_g+\frac{\beta}{\mu}\right)$ and the constant $\tilde{C}_M$ is defined in* (7)*. The number of iterations (communications) needed for $U(\mathbf{x}_i^\nu) - U^\star \leq \epsilon$, $i \in [m]$, is*

$$\mathcal{O}\left(\frac{\beta}{\mu}\log(1/\epsilon)\right), \qquad \textit{if} \quad \frac{\rho_B}{(1-\rho_B)^2} \leq \frac{1}{1428 \cdot \tilde{C}_M \cdot \left(1+\frac{L}{\beta}\right)\left(\kappa_g+\frac{\beta}{\mu}\right)}, \qquad (3)$$

$$\mathcal{O}\left(\frac{\left(\kappa_g+(\beta/\mu)\right)^2\rho_B}{(1-\rho_B)^2}\log(1/\epsilon)\right), \qquad \textit{otherwise.} \qquad (4)$$

Proof. In the setting of the corollary, we have: $\nabla^2\widetilde{f}_i(\mathbf{x};\mathbf{y}) = \nabla^2 f_i(\mathbf{x}) + \beta\mathbf{I}$, for all $\mathbf{y} \in \mathcal{K}$; $\nabla^2 f_i(\mathbf{x}) \succeq \mathbf{0}$, for all $\mathbf{x} \in \mathcal{K}$; and, by Assumption 2.1, $\mathbf{0} \preceq \nabla^2\widetilde{f}_i(\mathbf{x},\mathbf{y}) - \nabla^2 F(\mathbf{x}) \preceq 2\beta\mathbf{I}$, for all $\mathbf{x},\mathbf{y} \in \mathcal{K}$. Therefore, we can set $D_{\mathrm{mn}}^{\ell} = 0$, $D_{\mathrm{mx}} = 2\beta$, $\widetilde{\mu}_{\mathrm{mn}} = \beta + (\mu-\beta)_+ = \max\{\beta,\mu\}$, and $L_{\mathrm{mx}} = L + \beta$.

Using these values, $G_P^\star\left(\frac{\widetilde{\mu}_{\mathrm{mn}}}{\widetilde{\mu}_{\mathrm{mn}}-D_{\mathrm{mn}}^{\ell}}\right)$, C_1, and C_2 can be simplified as follows:

$$G_P^\star\left(\frac{\widetilde{\mu}_{\mathrm{mn}}}{\widetilde{\mu}_{\mathrm{mn}}-D_{\mathrm{mn}}^{\ell}}\right) = G_P^\star(1) = \frac{16\beta^2+\max\{\beta,\mu\}^2}{\mu\max\{\beta,\mu\}^2},$$
$$C_1 = \frac{6}{\mu\phi_{lb}}\left(\left(\frac{2\beta}{\max\{\beta,\mu\}}+1\right)^2+\frac{4(L+\beta)^2}{\max\{\beta,\mu\}^2}\right), \quad \text{and} \quad C_2 = \frac{4}{\max\{\beta,\mu\}^2}.$$

Accordingly, the expressions of J and $A_{\frac{1}{2}}$ read:

$$J = \frac{1}{2}\frac{1}{1+16\left(\frac{\beta}{\mu}\right)\cdot\min\left\{1,\frac{\beta}{\mu}\right\}}, \qquad (5)$$

and

$$
\begin{aligned}
&(A_{\frac{1}{2}})^2 \\
&\leq (G_P^\star(1) \cdot 2 \cdot C_1 \cdot 12\phi_{ub} + C_2) \cdot 8m\phi_{lb}^{-2} L_{\mathrm{mx}}^2 \cdot \frac{4c_0^4\rho_B^2}{(1-\rho_B)^2} \\
&\leq \left(\frac{16\beta^2 + \max\{\beta,\mu\}^2}{\mu \max\{\beta,\mu\}^2} \cdot \frac{12}{\mu\phi_{lb}} \left(\left(\frac{2\beta}{\max\{\beta,\mu\}} + 1 \right)^2 + \frac{4(L+\beta)^2}{\max\{\beta,\mu\}^2} \right) \cdot 12\phi_{ub} + \frac{4}{\max\{\beta,\mu\}^2} \right) \\
&\qquad\qquad\qquad\qquad\qquad \cdot 8m\phi_{lb}^{-2}(L+\beta)^2 \cdot \frac{4c_0^4\rho_B^2}{(1-\rho_B)^2} \\
&\leq \begin{cases} \left(2448 \cdot \frac{\phi_{ub}}{\phi_{lb}} \cdot \left(9 + 4\left(1 + \frac{L}{\beta}\right)^2 \right) \cdot \left(\kappa_g + \frac{\beta}{\mu}\right)^2 + 4\left(1+\frac{L}{\beta}\right)^2 \right) \cdot 8m\phi_{lb}^{-2} \cdot \frac{4c_0^4\rho_B^2}{(1-\rho_B)^2}, & \beta > \mu, \\ \left(144 \cdot \frac{\phi_{ub}}{\phi_{lb}} \cdot \left(\frac{16\beta^2}{\mu^2} + 1 \right) \left(\kappa_g + \frac{\beta}{\mu}\right)^2 \left(\left(\frac{2\beta}{\mu} + 1 \right)^2 + 4\left(\kappa_g + \frac{\beta}{\mu}\right)^2 \right) + 4\left(\kappa_g + \frac{\beta}{\mu}\right)^2 \right) & \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \cdot 8m\phi_{lb}^{-2} \cdot \frac{4c_0^4\rho_B^2}{(1-\rho_B)^2}, & \beta \leq \mu; \end{cases} \\
&\leq \tilde{M}^2 \frac{\rho_B^2}{1-\rho_B^2},
\end{aligned}
$$

where

$$
\tilde{M} \triangleq \begin{cases} 1428 \cdot \tilde{C}_M \left(1 + \frac{L}{\beta}\right)\left(\kappa_g + \frac{\beta}{\mu}\right), & \beta > \mu, \\ 1087 \cdot \tilde{C}_M \left(1 + \frac{\beta}{\mu}\right)^2 \left(\kappa_g + \frac{\beta}{\mu}\right)^2, & \beta \leq \mu, \end{cases} \tag{6}
$$

and

$$
\tilde{C}_M \triangleq c_0^2 \phi_{lb}^{-1} \sqrt{\frac{\phi_{ub}}{\phi_{lb}} \cdot m}, \tag{7}
$$

and the last inequality holds since $\kappa_g \geq 1$ and $\frac{\phi_{ub}}{\phi_{lb}} \geq 1$.

Similarly, we bound $z \leq \max\{z_1, z_2\}$ as

$$
z \leq \max\{z_1, \bar{z}_2\}, \quad \text{with} \quad z_1 \triangleq 1 - \alpha \cdot J \;\text{ and }\; \bar{z}_2 \triangleq \rho_B + \alpha\, \tilde{M} \cdot \frac{\rho_B}{1-\rho_B}, \tag{8}
$$

where J and $\tilde{M}$ are now given by (5) and (6), respectively. For $\max\{z_1, z_2\} < 1$, we require $\alpha \leq \alpha_{\mathrm{mx}} \triangleq \min\{1, (1-\rho_B)^2/(\tilde{M}\rho_B)\}$, and choose $\alpha = c \cdot \alpha_{\mathrm{mx}}$, with arbitrary $c \in (0,1)$.

- **Case I:** $\alpha_{\mathrm{mx}} = 1$. This correspond to $\tilde{M}\rho_B \leq (1-\rho_B)^2$, $\alpha = c$, hence,

$$
z_1 = 1 - c \cdot J \quad \text{and} \quad \bar{z}_2 \leq 1 - (1-c)\,(1-\rho_B).
$$

Since $\tilde{M} \geq 1087 \cdot \tilde{C}_M$ and $(1-\rho_B)^2 \leq 1$, it must be $\rho_B \leq 1/(1087 \cdot \tilde{C}_M)$. Therefore, the rate z can be bounded as

$$
\begin{aligned}
z \leq \max\{z_1, \bar{z}_2\} &\leq 1 - c \cdot (1-c) \cdot J \cdot (1-\rho_B) \\
&\leq 1 - c \cdot (1-c) \cdot \left(1 - \frac{1}{1087 \cdot \tilde{C}_M}\right) \cdot \frac{1}{34} \cdot \frac{\mu}{\beta},
\end{aligned}
$$

when $\beta > \mu$, and

$$z \leq 1 - c \cdot (1-c) \cdot \left(1 - \frac{1}{1087 \cdot \tilde{C}_M}\right) \cdot \frac{1}{2 + 32\left(\frac{\beta}{\mu}\right)^2},$$

when $\beta \leq \mu$.

- **Case II:** $\alpha_{\mathrm{mx}} = (1-\rho_B)^2/(\tilde{M}\rho_B)$. This corresponds to $\tilde{M}\rho_B > (1-\rho_B)^2$. We have $\alpha = c \cdot \alpha_{\mathrm{mx}}$. Similarly to inequality (9) in proof of Corollary IV.1, we have

$$z \leq \max\{z_1, \bar{z}_2\} \leq 1 - \frac{c\,J}{\tilde{M}\rho_B} \cdot (1-c)\,(1-\rho_B)^2,$$

which yields

$$z \leq 1 - \frac{c\,(1-c)}{1428 \cdot 34 \cdot \tilde{C}_M} \cdot \frac{(1-\rho_B)^2}{(\kappa_g + \beta/\mu)^2 \rho_B},$$

when $\beta > \mu$, and

$$\begin{aligned} z &\leq 1 - \frac{c\,(1-c)}{34 \cdot 1087 \cdot \tilde{C}_M} \cdot \frac{(1-\rho_B)^2}{(1+\beta/\mu)^2(\kappa_g + \beta/\mu)^2\rho_B} \\ &\leq 1 - \frac{c\,(1-c)}{34 \cdot 16 \cdot 1087 \cdot \tilde{C}_M} \cdot \frac{(1-\rho_B)^2}{\kappa_g^2 \rho_B}, \end{aligned}$$

when $\beta \leq \mu$; the last inequality holds due to $\kappa_g \geq 1$.

□