SIAM J. OPTIM.
Vol. 32, No. 2, pp. 354–385

DISTRIBUTED OPTIMIZATION BASED ON GRADIENT TRACKING REVISITED: ENHANCING CONVERGENCE RATE VIA SURROGATION*

YING SUN†, GESUALDO SCUTARI‡, AND AMIR DANESHMAND‡

Abstract. We study distributed multiagent optimization over 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 (thus going beyond linearization, such as proximal-gradient) coupled with a perturbed consensus mechanism that aims to locally track 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 (nonaccelerated) schemes just minimizing F, whose rate depends on much larger quantities than κ_g. When the loss functions of the agents are similar, due to statistical data similarity or otherwise, 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 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 those achievable by first-order models; this is without exchanging any Hessian matrix over the network.

Key words. distributed optimization, gradient tracking, linear rate, machine learning, statistical similarity, surrogate functions

AMS subject classifications. 68Q25, 68R10, 68U05

DOI. 10.1137/19M1259973

1. Introduction.

We study distributed optimization over networks in the form

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

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 associated agent only. 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, multiagent control, and machine learning. An instance of particular interest to this work is

*Received by the editors May 10, 2019; accepted for publication (in revised form) January 9, 2022; published electronically April 5, 2022.
https://doi.org/10.1137/19M1259973
Funding: This work has been supported by the National Science Foundation under grants CIF 1719205 and CMMI 1832688, by the Army Research Office under grant W911NF1810238, and by the Office of Naval Research under grant N00014-21-1-2673.

†School of Electrical Engineering and Computer Science, The Pennsylvania State University, State College, PA 16801 USA (ysun@psu.edu).
‡School of Industrial Engineering, Purdue University, West Lafayette, IN 47907 USA (gscutari@purdue.edu, adaneshm@purdue.edu).

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. Letting $\mathcal{D}^{(i)} = \{\mathbf{z}_1^{(i)}, \dots, \mathbf{z}_n^{(i)}\}$ be 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 of parameter $\mathbf{x}$ to the data $\mathbf{z}_j^{(i)}$. Datasets are usually large and high-dimensional, which makes routing local data to other agents (let alone to a centralized node) infeasible or highly inefficient. Given the cost of communications (especially if compared with the speed of local processing), the challenge in such a network setting is designing distributed algorithms that are communication-efficient.

Our focus pertains to such a design in two settings (one being a special case of the other): (1) The scenario where no significant relationship can be assumed among the local functions f_i—this has been extensively studied in the literature and will be referred 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. This is the case, e.g., of ERM problems: when data samples are independent and identically distributed (i.i.d.) among machines, quantities such as the gradients and Hessian matrices of the local functions differ by $\beta = \tilde{\mathcal{O}}(1/\sqrt{n})$ (with high probability) [2, 10, 38, 58]; we will refer to this as the *β-related* setting. If properly exploited in the algorithmic design, such similarity can yield significant communication savings over general purpose optimization 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 (nonaccelerated) 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 agents' functions f_i are sufficiently similar—specifically, $1+\beta/\mu < \kappa_g$—a linear rate scaling with κ_g may be highly unsatisfactory. For instance, this is the case of some ill-conditioned functions. Another example is ERM losses with optimal regularization $\mu = \mathcal{O}(1/\sqrt{mn})$ and L-smooth constant $L = \mathcal{O}(1)$ (e.g., see [58, Table 1] for ridge regression); we have $\kappa_g = \mathcal{O}(\sqrt{m \cdot n})$ while $\beta/\mu = \mathcal{O}(\sqrt{m})$– the former grows with the local sample size n, while the latter is independent. Function similarity has been explicitly explored in DANE [38], a mirror-descent scheme for (P) with $G \equiv 0$ whereby workers perform a local data preconditioning via a suitably chosen Bregman divergence, and the master averages the solutions of the workers. For quadratic losses, DANE achieves communication complexity $\widetilde{\mathcal{O}}((\beta/\mu)^2 \log 1/\epsilon)$ (it is assumed that $\beta/\mu \geq 1$). More recently, [9] proposed CEASE, which achieves DANE's rate but for nonquadratic losses and $G \not\equiv 0$. Applying the convergence analysis of mirror descent in [19] to CEASE enhances its rate to $\widetilde{\mathcal{O}}((\beta/\mu) \log 1/\epsilon)$.

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. The literature of distributed algorithms over general network topologies—albeit vast—does not provide a satisfactory answer, leaving a gap between rate results over star-networks and what has been certified over general graphs; see section 1.2 for a review of the state of the art. In a nutshell, (i) there are no distributed schemes provably achieving a linear rate for (P) with $G \not\equiv 0$ and/or constraints (cf. Table 1). Furthermore, even considering the unconstrained minimization of F (i.e., $G \equiv 0$ 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 2); 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 is achieved over star-topologies. The goal of this paper is to close 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 graphs. SONATA was earlier proposed in the companion paper [36] for nonconvex problems and directed, time-varying graphs. It combines the use of surrogate functions in the agents' subproblems with a perturbed 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 on SONATA's rates below (cf. Table 3).
2. *The 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 Chebyshev 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 instead show a more pessimistic dependence on the optimization parameters and are proved under more restrictive assumptions; contrast Table 2 with Table 3.
3. *The 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 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 2). Notice that this is achieved without exchanging any Hessian matrix over the network but by leveraging function homogeneity via surrogation. When customized over star-topologies, SONATA's rates improve on DANE/CEASE's rates too.
4. *Time-varying directed graphs.* The above rate improvements are extended also to time-varying directed graphs. Because of space limitations, details can be found in the long version of the paper [42].

1.2. Related works. Early works on distributed optimization aimed at decentralizing the (sub)gradient algorithm. The Distributed Gradient Descent (DGD) was introduced in [26] for unconstrained instances of (P) and in [18] for least squares, both over undirected graphs. A refined convergence rate analysis of DGD [26] can be found in [55]. Subsequent variants of DGD include the projected (sub)gradient algorithm [27] and the push-sum gradient consensus algorithm [23], 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

TABLE 1

Existing linearly convergent distributed algorithms. SONATA is the only scheme achieving a linear rate in the presence of G in (P) *or constraints. The expression of the rates of the above nonaccelerated schemes (when available) is reported in Table* 2*. SONATA over time-varying graphs is discussed in the long version of the paper* [42].

Algorithms		[11, 12, 16, 17, 20, 22, 30, 39, 41]	[28, 48, 49, 57]	[21, 24, 25, 32]	**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 [20]		only [24, 32]	✓
	digraph		✓	only [24, 32]	✓

TABLE 2

Linear rate of existing nonaccelerated 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, respectively; $L_{\rm mx} \triangleq max_i L_i$, $\mu_{\rm mn} \triangleq \min_i \mu_i$; and $\rho \in [0,1)$ is the second largest eigenvalue modulus of the mixing matrix (cf. (3.14)*). The rates above include the quantities κ_l, $\hat{\kappa}$, and $\bar{\kappa}$ rather than the much more 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 of β, implying that faster rates are not certified when $1 + \beta/\mu < \kappa_g$ (β-related setting). Note that when the gossip matrix used in the algorithms above is symmetric and Chebyshev acceleration is employed, the dependence of the communication complexity on the network improves to $\sqrt{1-\rho}$.*

Algorithm	**Problem**	**Linear rate:** $\mathcal{O}\big(\delta \log(1/\epsilon)\big)$
EXTRA [39]	F	$\delta = \mathcal{O}\big(\frac{\kappa_\ell^2}{1-\rho}\big), \quad \kappa_\ell = \frac{L_{\rm mx}}{\mu_{\rm mn}}$
DIGing [24, 25]	F	$\delta = \frac{\hat{\kappa}^{1.5}}{(1-\rho)^2}, \quad \hat{\kappa} \triangleq \frac{L_{\rm mx}}{(1/m)\sum_i \mu_i}$
Harnessing [30]	F	$\delta = \frac{\kappa_\ell^2}{(1-\rho)^2}$
NIDS [16], ABC [14]	F	$\delta = \max\big\{\kappa_\ell, \frac{1}{1-\rho}\big\}$
Exact Diffusion [56]	F	$\delta = \frac{\bar{\kappa}^2}{1-\rho}, \quad \bar{\kappa} \triangleq \frac{L_{\rm mx}}{\mu_{\rm mx}}$
Augmented Lagrangian [12]	F	$\delta = \frac{\kappa_\ell}{1-\rho}$
ADMM [41]	F	$\frac{\kappa_\ell^4}{1-\rho}$

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 (cf. Definition 2.1*); the other quantities are defined as in Table* 2*. The extra averaging steps are performed using Chebyshev acceleration* [3, 33]*. The $\widetilde{O}$ notation hides log dependence on κ_g and β/μ (see section* 3.4.2 *for the exact expressions).*

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

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, a linear rate of the iterates is achievable, but it can only converge to an $\mathcal{O}(\alpha)$-neighborhood of the solution [26, 55].

Several subsequent attempts have been proposed to cope with this speed-accuracy dilemma, leading to algorithms converging to the *exact* solution by 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 [6, 7, 24, 28, 29, 30, 47, 48, 49, 51, 52, 53]; (ii) distributed schemes using ad hoc corrections of the local optimization direction [4, 39, 57]; and (iii) primal-dual-based methods [12, 17, 22, 33, 41]. We elaborate next on these works, focusing on schemes achieving a linear rate; Table 1 organizes these schemes based upon the setting in which their convergence is established, while Table 2 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 [6, 7] for problem (P) and in AUG-DGM [53] for strongly convex, smooth, unconstrained optimization. The work [43] introduced SONATA, extending NEXT over (time-varying) digraphs. A convergence rate analysis of [53] was later developed in [24, 30, 54], with [24] considering also (time-varying) digraphs. Other algorithms based on the idea of gradient tracking and implementable over digraphs are ADD-OPT [47] and [49]. Subsequent schemes [48], the Push-Pull [28] and the $\mathcal{AB}$ [32] algorithms, relaxed previous conditions on the mixing matrices used in the consensus and gradient tracking steps over digraphs, which need be neither row-stochastic nor column-stochastic. All the schemes above except 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 [39] and its variant over digraphs, EXTRA-PUSH [57], introduce two different weight matrices for any two consecutive iterations as well as leverage history of gradient information. They are applicable only to 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, [40] 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 directly correcting the gradient direction; this has been studied in [4] 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 [13, 41] and their inexact implementations [17, 21]; (ii) distributed augmented Lagrangian-based methods with randomized primal variable updates [12]; and (iii) a distributed dual ascent method employing tracking of the average of the primal variable [20]. All these schemes are applicable only to *smooth,*

unconstrained optimization over undirected graphs.

To summarize, the above literature review shows that currently there exists no distributed algorithm for the general formulation (P) that provably converges at a linear rate to the exact solution, in the presence of a nonsmooth function G or constraints (cf. Table 1), let alone mentions 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 [42], a few other related technical reports appeared online [1, 15, 31], 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 $\breve{\kappa} \triangleq L_{\rm mx}/\mu$. The authors in [31] proposed an inexact distributed projected gradient descent method for the unconstrained 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 $L_{\rm max}^2/\mu^2$), which depends on the global condition number κ_g. SONATA's rates compare favorably with those above. Furthermore, since both schemes [1] and [31] are gradient-type methods, unlike SONATA, their performance cannot benefit from function similarity, if any. On the other hand, [15] 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 [15] 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 those 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's rates exhibit a better dependence on the optimization parameters (κ_g versus 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. Section 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 section 3; in particular, linear convergence is proved in section 3.3, while a detailed discussion on the rate expression and its scalability properties is provided in section 3.4. The case of time-varying, possibly directed, graphs can be found in the long version of the paper, along with some numerical results supporting our theoretical findings [42].

2. Problem and network setting.

This section summarizes the assumptions on the optimization problem and network setting.

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. section 2.1.1), and (ii) the case where they are related (cf. section 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 implies

$$\mu_i \mathbf{I} \preceq \nabla^2 f_i(\mathbf{x}) \preceq L_i \mathbf{I} \quad \forall \mathbf{x} \in \mathcal{K},\ \forall i \in [m], \tag{2.1}$$

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 is assumed here to simplify our derivations.

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

$$\kappa_g \triangleq \frac{L}{\mu}. \tag{2.2}$$

Related quantities determining the (linear) convergence rate of existing distributed algorithms are (cf. Table 2):

$$\kappa_\ell \triangleq \frac{L_{\rm mx}}{\mu_{\rm mn}}, \quad \hat{\kappa} \triangleq \frac{L_{\rm mx}}{(1/m)\sum_i \mu_i}, \quad \breve{\kappa} \triangleq \frac{L_{\rm mx}}{\mu}, \quad \text{and} \quad \bar{\kappa} \triangleq \frac{L_{\rm mx}}{\mu_{\rm mx}}, \tag{2.3}$$

where

$$L_{\rm mx} \triangleq \max_{i=1,\ldots,m} L_i, \quad \mu_{\rm mn} \triangleq \min_{i=1,\ldots,m} \mu_i, \quad \text{and} \quad \mu_{\rm mx} \triangleq \max_{i=1,\ldots,m} \mu_i. \tag{2.4}$$

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; see, e.g., [42, Example 1].

In the setting above, our goal is to design linearly convergent algorithms whose iteration complexity is proportional to κ_g instead of the larger quantities in (2.3).

2.1.2. The β-related setting. This scenario 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). *Under Assumption* A*, let $\beta \geq 0$ be the smallest number such that $\left\|\nabla^2 F(\mathbf{x}) - \nabla^2 f_i(\mathbf{x})\right\|_2 \leq \beta$, for all $\mathbf{x} \in \mathcal{K}$.*

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

$$\beta \leq \max_{i=1,\ldots,m} \max\left\{|L - \mu_i|,\, |\mu - L_i|\right\}.$$

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

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

$$\mathbf{x}^\star \in \operatorname*{argmin}_{\mathbf{x} \in \mathcal{H}} F(\mathbf{x}), \quad \text{with} \quad F(\mathbf{x}) \triangleq \mathbb{E}_{\mathbf{z} \sim \mathcal{P}}\left[f(\mathbf{x}; \mathbf{z})\right], \tag{2.5}$$

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 $\mathbf{x}^\star \in \mathcal{B}_B \triangleq \{\mathbf{x} : \|\mathbf{x}\| \leq B\}$ 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 (2.5), 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 section 1, the ERM problem reads

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

which is clearly an instance of (P), satisfying Assumption A.

We derive the associated β/μ and contrast it 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 the use of $\widehat{\mathbf{x}}$ as proxy for $\mathbf{x}^\star$; we have [37, Th. 7]:

$$\lambda = \mathcal{O}\Big(\sqrt{G_f^2/(B^2\, N)}\Big), \tag{2.6}$$

where G_f is the Lipschitz constant of $f(\bullet; \mathbf{z})$ on $\mathcal{H} \bigcap \mathcal{B}_B$ for all $\mathbf{z} \in \mathcal{Z}$.

An estimate of β can be obtained by exploring the statistical similarity of f_i. Under the additional assumption that $\nabla^2 f(\bullet; \mathbf{z})$ is M-Lipschitz on $\mathcal{H}$, for all $\mathbf{z} \in \mathcal{Z}$, the following holds with high probability [59, Lemma 6]:

$$\sup_{\mathbf{x} \in \mathcal{B}_B} \left\| \nabla^2 f_i(\mathbf{x}; \mathbf{z}) - \nabla^2 \hat{F}(\mathbf{x}) \right\| \leq \beta = \widetilde{\mathcal{O}}\left(\sqrt{\frac{L^2\, d}{n}} \right), \tag{2.7}$$

for all $\mathbf{z} \in \mathcal{Z}$, $i \in [m]$, where $\widetilde{\mathcal{O}}$ hides the log-factor dependence.

Based on (2.6)–(2.7), an estimate of β/μ and κ_g reads

$$1 + \frac{\beta}{\mu} = 1 + \widetilde{\mathcal{O}}\left(L \sqrt{d\, m}\right) \quad \text{and} \quad \kappa_g = 1 + \widetilde{\mathcal{O}}\left(L \sqrt{d\, m\, n}\right). \tag{2.8}$$

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 (2.6) (larger n) are achieved at the cost of more iterations (larger κ_g). In this setting, it is desirable to design distributed algorithms whose rate depends on β/μ rather than κ_g.

2.2. Network setting. We model the network of agents 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.

The network setting covers, as a 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.

Algorithm 3.1. 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

$$\widehat{\mathbf{x}}_i^\nu \triangleq \operatorname*{argmin}_{\mathbf{x}_i \in \mathcal{K}} \underbrace{\widetilde{f}_i(\mathbf{x}_i; \mathbf{x}_i^\nu) + \big(\mathbf{y}_i^\nu - \nabla f_i(\mathbf{x}_i^\nu)\big)^\top (\mathbf{x}_i - \mathbf{x}_i^\nu)}_{\widetilde{F}_i(\mathbf{x}_i; \mathbf{x}_i^\nu)} + G(\mathbf{x}_i) \tag{3.1a}$$

and updates

$$\mathbf{x}_i^{\nu+\frac{1}{2}} = \mathbf{x}_i^\nu + \alpha \cdot \mathbf{d}_i^\nu, \quad \text{with} \quad \mathbf{d}_i^\nu \triangleq \widehat{\mathbf{x}}_i^\nu - \mathbf{x}_i^\nu; \tag{3.1b}$$

`[S.2] [Information Mixing]` Each agent i computes

(a) `Consensus`

$$\mathbf{x}_i^{\nu+1} = \sum_{j=1}^m w_{ij} \mathbf{x}_j^{\nu+\frac{1}{2}}, \tag{3.1c}$$

(b) `Gradient tracking`

$$\mathbf{y}_i^{\nu+1} = \sum_{j=1}^m w_{ij} \big(\mathbf{y}_j^\nu + \nabla f_j(\mathbf{x}_j^{\nu+1}) - \nabla f_j(\mathbf{x}_j^\nu)\big). \tag{3.1d}$$

end

3. The SONATA algorithm. We recall here the SONATA/NEXT algorithm [7, 36], 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 algorithm is described in Algorithm 3.1. 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 (3.1a) 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. (3.1b)], 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 (3.1c), while the y-variables are updated via the perturbed consensus (3.1d), aiming at tracking $\nabla F(\mathbf{x}_i^\nu)$.

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

• *On the subproblem* (3.1a) *and surrogate functions* $\widetilde{f}_i$. 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,

$$\widetilde{f}_i(\mathbf{x}_i;\mathbf{x}_i^\nu) = \nabla f_i(\mathbf{x}_i^\nu)^\top(\mathbf{x}_i-\mathbf{x}_i^\nu) + \frac{\tau_i}{2}\|\mathbf{x}_i-\mathbf{x}_i^\nu\|^2, \quad \text{with} \quad \tau_i>0, \tag{3.2}$$

which leads to the standard proximal-gradient update for $\widehat{\mathbf{x}}_i$. Note that if, in addition, $G=0$ and $\mathcal{K}=\mathbb{R}^d$, (3.1a)–(3.1c) reduces to the standard (ATC) consensus/gradient-tracking step (setting $\alpha=1$ and absorbing $1/\tau_i$ into the common stepsize γ) $\mathbf{x}_i^{\nu+1} = \sum_j w_{ij}(\mathbf{x}_i^\nu - \gamma\,\mathbf{y}_i^\nu)$ [24, 30, 53]. 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 (3.2), we have a surrogate retaining all of the structure of f_i, such as

$$\widetilde{f}_i(\mathbf{x}_i;\mathbf{x}_i^\nu) = f_i(\mathbf{x}_i) + \frac{\tau_i}{2}\|\mathbf{x}_i-\mathbf{x}_i^\nu\|^2, \quad \text{with} \quad \tau_i>0. \tag{3.3}$$

We refer the reader to [8, 34, 35] 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 (3.3) indeed yield faster rates of SONATA than those achievable using linear surrogates (3.2). 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, $D_i^\ell \le D_i^u$, such that

$$D_i^\ell\,\mathbf{I} \preceq \nabla^2\widetilde{f}_i(\mathbf{x},\mathbf{y}) - \nabla^2 F(\mathbf{x}) \preceq D_i^u\,\mathbf{I} \quad \forall \mathbf{x},\mathbf{y}\in\mathcal{K}; \qquad \text{let} \quad D_i \triangleq \max\{|D_i^\ell|,|D_i^u|\}. \tag{3.4}$$

For instance, (3.4) holds with $D_i=\max\{|\widetilde{\mu}_i-L|,|\widetilde{L}_i-\mu|\}$. Roughly speaking, the smaller D_i the better $\widetilde{F}_i$ (defined in (3.1a)) approximates F. 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 (3.3), may be more effective than mere linearization. Our theoretical findings confirm the above intuition; see section 3.4.

• *Consensus and gradient tracking steps* (3.1c)–(3.1d). In the consensus and tracking steps, the weights w_{ij} 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 $i=1,\dots,m$;

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–Hastings, and the maximum-degree rules [50].

Finally, we comment on 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

$$\bar{\mathbf{y}}^\nu \triangleq \frac{1}{m}\sum_{i=1}^m \mathbf{y}_i^\nu \quad \text{and} \quad \overline{\nabla\mathbf{f}}^\nu \triangleq \frac{1}{m}\sum_{i=1}^m \nabla f_i(\mathbf{x}_i^\nu). \tag{3.5}$$

Summing (3.1d) over $i\in[m]$ and invoking the doubly stochasticity of $\mathbf{W}$; we have

$$\bar{\mathbf{y}}^{\nu+1} = \bar{\mathbf{y}}^\nu + \overline{\nabla\mathbf{f}}^{\nu+1} - \overline{\nabla\mathbf{f}}^\nu. \tag{3.6}$$

Applying (3.6) inductively and using the initial condition $\mathbf{y}_i^0=\nabla f_i(\mathbf{x}_i^0)$, $i\in[m]$, yield

$$\bar{\mathbf{y}}^\nu = \overline{\nabla\mathbf{f}}^\nu, \quad \forall \nu=0,1,\dots. \tag{3.7}$$

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 $\|\mathbf{y}_i^\nu-\mathbf{y}_j^\nu\| \underset{\nu\to\infty}{\longrightarrow} 0$, $i\neq j$, (3.7) 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 $i\in[m]$.

Algorithm 3.2. SONATA on star-networks (SONATA-Star).

Data: $\mathbf{x}^0 \in \mathcal{K}$.

Iterate: $\nu = 1, 2, \ldots$

[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

$$\widehat{\mathbf{x}}_i^\nu \triangleq \operatorname*{argmin}_{\mathbf{x}_i \in \mathcal{K}} \widetilde{f}_i(\mathbf{x}_i; \mathbf{x}^\nu) + \big(\nabla F(\mathbf{x}^\nu) - \nabla f_i(\mathbf{x}_i^\nu)\big)^\top (\mathbf{x}_i - \mathbf{x}^\nu) + G(\mathbf{x}_i)$$

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

3.1. A special instance: SONATA on star-networks. Although the main focus of the paper is studying SONATA over mesh networks, it is worth discussing here a special instance over star-networks. Consider a star (undirected) 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. Problem (P) can be solved using Algorithm 3.2, 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. (3.2)) and $\alpha = 1$, reduces to the proximal gradient algorithm. When the surrogates (3.3) are used (and still $\alpha = 1$), SONATA-star coincides with the DANE algorithm [38] if $G = 0$ and to the CEASE (with averaging) algorithm [9] if $G \neq 0$. Nevertheless, our convergence rates improve on those of DANE and CEASE; see section 3.4.1.

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

$$p^\nu \triangleq \sum_{i=1}^m \big(U(\mathbf{x}_i^\nu) - U(\mathbf{x}^\star)\big), \tag{3.8}$$

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

We stack the local variables and gradients in the column vectors

$$\mathbf{x}^\nu \triangleq [\mathbf{x}_1^{\nu\top}, \ldots, \mathbf{x}_m^{\nu\top}]^\top,\ \mathbf{y}^\nu \triangleq [\mathbf{y}_1^{\nu\top}, \ldots, \mathbf{y}_m^{\nu\top}]^\top,\ \nabla \mathbf{f}^\nu \triangleq [\nabla f_1(\mathbf{x}_1^\nu)^\top, \ldots, \nabla f_m(\mathbf{x}_m^\nu)^\top]^\top. \tag{3.9}$$

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

$$\mathbf{x}_\perp^\nu \triangleq \mathbf{x}^\nu - \mathbf{1}_m \otimes \bar{\mathbf{x}}^\nu \quad \text{and} \quad \mathbf{y}_\perp^\nu \triangleq \mathbf{y}^\nu - \mathbf{1}_m \otimes \bar{\mathbf{y}}^\nu, \tag{3.10}$$

respectively, while the gradient tracking error is defined as

$$\boldsymbol{\delta}^\nu \triangleq [\boldsymbol{\delta}_1^{\nu\top}, \ldots, \boldsymbol{\delta}_m^{\nu\top}]^\top, \quad \text{with} \quad \boldsymbol{\delta}_i^\nu \triangleq \nabla F(\mathbf{x}_i^\nu) - \mathbf{y}_i^\nu, \quad i = 1, \ldots, m. \tag{3.11}$$

Recalling L_i, $\widetilde{L}_i$, $\widetilde{\mu}_i$, D_i^ℓ, and D_i as given in Assumptions A and C and (3.4), we introduce the following algorithm-dependent parameters:

$$\widetilde{\mu}_{\rm mn} \triangleq \min_{i\in[m]} \widetilde{\mu}_i, \quad \widetilde{L}_{\rm mx} \triangleq \max_{i\in[m]}, \widetilde{L}_i, \quad D^\ell_{\rm mn} \triangleq \min_{i\in[m]} D_i^\ell, \quad D_{\rm mx} \triangleq \max_{i\in[m]} D_i. \tag{3.12}$$

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

$$\widehat{\mathbf{W}} \triangleq \mathbf{W} \otimes \mathbf{I}_d \quad \text{and} \quad \mathbf{J} \triangleq \frac{1}{m}\mathbf{1}_m\mathbf{1}_m^\top \otimes \mathbf{I}_d. \tag{3.13}$$

Under Assumptions B and D, it is well known that (see, e.g., [46])

$$\rho \triangleq \sigma(\widehat{\mathbf{W}} - \mathbf{J}) < 1, \tag{3.14}$$

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

3.3. Linear convergence rate. Our proof of the 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 a 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

$$p^{\nu+1} \le \sum_{i=1}^m \sum_{j=1}^m w_{ij}\Big(U\big(\mathbf{x}_j^{\nu+\frac{1}{2}}\big) - U(\mathbf{x}^\star)\Big) = \sum_{i=1}^m \Big(U(\mathbf{x}_i^{\nu+\frac{1}{2}}) - U(\mathbf{x}^\star)\Big). \tag{3.15}$$

We can now bound $U(\mathbf{x}_j^{\nu+\frac{1}{2}})$, regarding the local optimization (3.1a)–(3.1b) 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 (3.1b) 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*

$$U(\mathbf{x}_i^{\nu+\frac{1}{2}}) \le U(\mathbf{x}_i^\nu) - \alpha\left(\left(1-\frac{\alpha}{2}\right)\widetilde{\mu}_i + \frac{\alpha}{2}\cdot D_i^\ell\right)\|\mathbf{d}_i^\nu\|^2 + \alpha\|\mathbf{d}_i^\nu\|\|\boldsymbol{\delta}_i^\nu\|, \tag{3.16}$$

with D_i^ℓ and $\boldsymbol{\delta}_i^\nu$ defined as in (3.4) *and* (3.11)*, respectively.*

Proof. Consider the Taylor expansion of F:

$$\begin{aligned} F(\mathbf{x}_i^{\nu+\frac{1}{2}}) &= F(\mathbf{x}_i^\nu) + \nabla F(\mathbf{x}_i^\nu)^\top(\alpha\mathbf{d}_i^\nu) + (\alpha\mathbf{d}_i^\nu)^\top\mathbf{H}(\alpha\mathbf{d}_i^\nu) \\ &\overset{(3.11)}{=} F(\mathbf{x}_i^\nu) + \big(\boldsymbol{\delta}_i^\nu\big)^\top(\alpha\mathbf{d}_i^\nu) + \big(\mathbf{y}_i^\nu\big)^\top(\alpha\mathbf{d}_i^\nu) + (\alpha\mathbf{d}_i^\nu)^\top\mathbf{H}(\alpha\mathbf{d}_i^\nu), \end{aligned} \tag{3.17}$$

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 $\widehat{\mathbf{x}}_i^\nu$ 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

$$G(\mathbf{x}_i^\nu) - G(\widehat{\mathbf{x}}_i^\nu) \ge (\mathbf{d}_i^\nu)^\top\big(\nabla\widetilde{f}_i(\widehat{\mathbf{x}}_i^\nu;\mathbf{x}_i^\nu) + \mathbf{y}_i^\nu - \nabla f_i(\mathbf{x}_i^\nu)\big) = (\mathbf{d}_i^\nu)^\top\big(\mathbf{y}_i^\nu + \widetilde{\mathbf{H}}_i\mathbf{d}_i^\nu\big), \tag{3.18}$$

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 (3.18) into (3.17) and using the convexity of G yields

(3.19)
$$\begin{aligned} &F(\mathbf{x}_i^{\nu+\frac{1}{2}}) \\ &\le F(\mathbf{x}_i^\nu) + (\boldsymbol{\delta}_i^\nu)^\top(\alpha \mathbf{d}_i^\nu) + \alpha\left(-(\mathbf{d}_i^\nu)^\top \widetilde{\mathbf{H}}_i \mathbf{d}_i^\nu + (\alpha \mathbf{d}_i^\nu)^\top \mathbf{H}(\mathbf{d}_i^\nu)\right) + G(\mathbf{x}_i^\nu) - G(\mathbf{x}_i^{\nu+\frac{1}{2}}). \end{aligned}$$

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

(3.20)
$$\begin{aligned} &\alpha\mathbf{H} - \widetilde{\mathbf{H}}_i \\ &\overset{(3.1b)}{=} \int_0^\alpha (1-\theta/\alpha)\nabla^2 F(\theta\widehat{\mathbf{x}}_i^\nu + (1-\theta)\mathbf{x}_i^\nu)d\theta - \int_0^1 \nabla^2 \widetilde{f}_i(\theta\widehat{\mathbf{x}}_i^\nu + (1-\theta)\mathbf{x}_i^\nu;\mathbf{x}_i^\nu)d\theta \\ &\overset{(3.4)}{\preceq} -\int_0^\alpha (1-\theta/\alpha)\cdot(D_i^\ell)\,\mathbf{I}\,d\theta - \int_0^\alpha (\theta/\alpha)\nabla^2 \widetilde{f}_i(\theta\widehat{\mathbf{x}}_i + (1-\theta)\mathbf{x}_i^\nu;\mathbf{x}_i^\nu)d\theta \\ &\quad - \int_\alpha^1 \nabla^2 \widetilde{f}_i(\theta\,\widehat{\mathbf{x}}_i^\nu + (1-\theta)\,\mathbf{x}_i^\nu;\mathbf{x}_i^\nu)d\theta \overset{(a)}{\preceq} -\frac{1}{2}\alpha\,(D_i^\ell)\,\mathbf{I} - \left(1-\frac{\alpha}{2}\right)\widetilde{\mu}_i\,\mathbf{I}, \end{aligned}$$

where (a) follows from Assumption C(iii). Substituting (3.20) into (3.19) completes the proof. ∎

We can now substitute (3.16) into (3.15) and get

$$p^{\nu+1} \le p^\nu + \sum_{i=1}^m \left\{\alpha\|\mathbf{d}_i^\nu\|\|\boldsymbol{\delta}_i^\nu\| - \alpha\left(1-\frac{\alpha}{2}\right)\widetilde{\mu}_i\|\mathbf{d}_i^\nu\|^2 - \frac{D_i^\ell}{2}\alpha^2\|\mathbf{d}_i^\nu\|^2\right\} \tag{3.21a}$$

$$\overset{(a)}{\le} p^\nu - \left(\left(1-\frac{\alpha}{2}\right)\widetilde{\mu}_{\rm mn} + \frac{\alpha D_{\rm mn}^\ell}{2} - \frac{1}{2}\epsilon_{opt}\right)\alpha\|\mathbf{d}^\nu\|^2 + \frac{1}{2}\epsilon_{opt}^{-1}\,\alpha\cdot\|\boldsymbol{\delta}^\nu\|^2, \tag{3.21b}$$

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

$$\left(1-\frac{\alpha}{2}\right)\widetilde{\mu}_{\rm mn} + \frac{\alpha D_{\rm mn}^\ell}{2} - \frac{1}{2}\epsilon_{opt} > 0, \tag{3.22}$$

and $D_{\rm mn}^\ell$ is defined in (3.12).

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

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

$$\alpha\,\|\mathbf{d}^\nu\|^2 \ge \frac{\mu}{D_{\rm mx}^2}\left(p^{\nu+1} - (1-\alpha)p^\nu - \frac{\alpha}{\mu}\|\boldsymbol{\delta}^\nu\|^2\right), \tag{3.23}$$

where $D_{\rm mx}$ *is defined as in* (3.12).

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

$$G(\mathbf{x}^\star) - G(\widehat{\mathbf{x}}_i^\nu) \ge -(\mathbf{x}^\star - \widehat{\mathbf{x}}_i^\nu)^\top\left(\nabla\widetilde{f}_i(\widehat{\mathbf{x}}_i^\nu;\mathbf{x}_i^\nu) + \mathbf{y}_i^\nu - \nabla f_i(\mathbf{x}_i^\nu)\right). \tag{3.24}$$

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{(3.24)}{\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 \\
&\geq U(\widehat{\mathbf{x}}_i^\nu) - \frac{1}{2\mu}\Big\|\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)\Big\|^2 \\
&\geq U(\widehat{\mathbf{x}}_i^\nu) - \frac{1}{\mu}\Big\|\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)\Big\|^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

$$
\|\mathbf{d}^\nu\|^2 \geq \frac{\mu}{D_{\mathrm{mx}}^2}\left(\sum_{i=1}^m \big(U(\widehat{\mathbf{x}}_i^\nu) - U(\mathbf{x}^\star)\big) - \frac{1}{\mu}\|\boldsymbol{\delta}^\nu\|^2\right). \tag{3.25}
$$

Using (3.15) 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

$$
\alpha\sum_{i=1}^m \big(U(\widehat{\mathbf{x}}_i^\nu) - U(\mathbf{x}^\star)\big) \geq p^{\nu+1} - (1-\alpha)p^\nu. \tag{3.26}
$$

Combining (3.25) with (3.26) provides the desired result (3.23). ∎

As a last step, we upper bound $\|\boldsymbol{\delta}^\nu\|^2$ in (3.21) 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 $\|\boldsymbol{\delta}^\nu\|^2$:*

$$
\|\boldsymbol{\delta}^\nu\|^2 \leq 4L_{\mathrm{mx}}^2\|\mathbf{x}_\perp^\nu\|^2 + 2\|\mathbf{y}_\perp^\nu\|^2, \tag{3.27}
$$

where L_{mx} is defined as in (2.4).

Proof.

$$
\begin{aligned}
\|\boldsymbol{\delta}^\nu\|^2 &\overset{(3.11)}{=} \sum_{i=1}^m \|\nabla F(\mathbf{x}_i^\nu) \pm \bar{\mathbf{y}}^\nu - \mathbf{y}_i^\nu\|^2 \\
&\overset{(3.5)}{=} \frac{1}{m^2}\sum_{i=1}^m \Big\|\sum_{j=1}^m \nabla f_j(\mathbf{x}_i^\nu) - \sum_{j=1}^m \nabla f_j(\mathbf{x}_j^\nu) + m\cdot\bar{\mathbf{y}}^\nu - m\cdot\mathbf{y}_i^\nu\Big\|^2 \\
&\overset{(2.1),\,(2.4)}{\leq} \frac{1}{m^2}\sum_{i=1}^m\left(2m\sum_{j=1}^m L_{\mathrm{mx}}^2\|\mathbf{x}_i^\nu - \mathbf{x}_j^\nu\|^2 + 2m^2\|\bar{\mathbf{y}}^\nu - \mathbf{y}_i^\nu\|^2\right) \\
&= 4L_{\mathrm{mx}}^2\|\mathbf{x}_\perp^\nu\|^2 + 2\|\mathbf{y}_\perp^\nu\|^2.
\end{aligned}
$$

∎

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 by multiplying (3.21) and (3.23) by $\widetilde{\mu}_{\mathrm{mn}} - \frac{L}{2}\alpha - \frac{1}{2}\epsilon_{opt}$ and $6(L^2 + \widetilde{L}_{\mathrm{mx}}^2)/\mu$, respectively, adding them together to cancel out $\|\mathbf{d}^\nu\|$, and using (3.27) to bound $\|\boldsymbol{\delta}^\nu\|^2$.

PROPOSITION 3.4. *The optimality gap p^ν (cf. (3.8)) satisfies*

$$p^{\nu+1} \le \sigma(\alpha)\cdot p^\nu + \eta(\alpha)\cdot\left(4L_{\rm mx}^2\|\mathbf{x}_\perp^\nu\|^2 + 2\|\mathbf{y}_\perp^\nu\|^2\right), \tag{3.28}$$

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

$$\sigma(\alpha) \triangleq 1-\alpha\,\frac{\left(1-\frac{\alpha}{2}\right)\widetilde{\mu}_{\rm mn}+\frac{D_{\rm mn}^\ell}{2}\alpha-\frac{1}{2}\epsilon_{opt}}{\frac{D_{\rm mx}^2}{\mu}+\left(1-\frac{\alpha}{2}\right)\widetilde{\mu}_{\rm mn}+\frac{D_{\rm mn}^\ell}{2}\alpha-\frac{1}{2}\epsilon_{opt}}, \tag{3.29}$$

$$\eta(\alpha) \triangleq \frac{\frac{1}{2}\epsilon_{opt}^{-1}\alpha\cdot\frac{D_{\rm mx}^2}{\mu}+\frac{\alpha}{\mu}\cdot\left(\left(1-\frac{\alpha}{2}\right)\widetilde{\mu}_{\rm mn}+\frac{D_{\rm mn}^\ell}{2}\alpha-\frac{1}{2}\epsilon_{opt}\right)}{\frac{D_{\rm mx}^2}{\mu}+\left(1-\frac{\alpha}{2}\right)\widetilde{\mu}_{\rm mn}+\frac{D_{\rm mn}^\ell}{2}\alpha-\frac{1}{2}\epsilon_{opt}}; \tag{3.30}$$

ϵ_{opt} satisfies (3.22); and $L_{\rm mx}$ and $\widetilde{\mu}_{\rm mn}$, $D_{\rm mn}^\ell$, $D_{\rm mx}$ are defined in (2.4) and (3.12), 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 (3.1a)–(3.1d) in vector-matrix form; using (3.9) and (3.13), we have

$$\mathbf{x}^{\nu+1} = \widehat{\mathbf{W}}(\mathbf{x}^\nu+\alpha\mathbf{d}^\nu), \tag{3.31a}$$

$$\mathbf{y}^{\nu+1} = \widehat{\mathbf{W}}(\mathbf{y}^\nu+\nabla\mathbf{f}^{\nu+1}-\nabla\mathbf{f}^\nu). \tag{3.31b}$$

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 (3.31) that

$$\mathbf{x}_\perp^{\nu+1} = (\widehat{\mathbf{W}}-\mathbf{J})(\mathbf{x}_\perp^\nu+\alpha\mathbf{d}^\nu), \tag{3.32}$$

$$\mathbf{y}_\perp^{\nu+1} = (\widehat{\mathbf{W}}-\mathbf{J})(\mathbf{y}_\perp^\nu+\nabla\mathbf{f}^{\nu+1}-\nabla\mathbf{f}^\nu). \tag{3.33}$$

Using (3.32)–(3.33), 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. *The following hold:*

$$\|\mathbf{x}_\perp^{\nu+1}\| \le \rho\|\mathbf{x}_\perp^\nu\| + \alpha\rho\|\mathbf{d}^\nu\|, \tag{3.34a}$$

$$\|\mathbf{y}_\perp^{\nu+1}\| \le \rho\|\mathbf{y}_\perp^\nu\| + 2L_{\rm mx}\rho\|\mathbf{x}_\perp^\nu\| + \alpha L_{\rm mx}\rho\|\mathbf{d}^\nu\|, \tag{3.34b}$$

with ρ and $L_{\rm mx}$ defined as in (3.14) and (2.4), respectively.

Proof. We prove next (3.34b); (3.34a) follows readily from (3.32). Using (3.31a), (3.33), and the Lipschitz continuity of ∇f_i (cf. (2.1)), we can bound $\|\mathbf{y}_\perp^{\nu+1}\|$ as

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

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

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* $\|\mathbf{d}^\nu\|$*:*

$$\|\mathbf{d}^\nu\|^2 \leq \frac{6}{\mu}\left(\left(\frac{D_{\mathrm{mx}}}{\widetilde{\mu}_{\mathrm{mn}}}+1\right)^2+\frac{4L_{\mathrm{mx}}^2}{\widetilde{\mu}_{\mathrm{mn}}^2}\right)p^\nu+\frac{3}{\widetilde{\mu}_{\mathrm{mn}}^2}\|\mathbf{y}_\perp^\nu\|^2, \tag{3.35}$$

where L_{mx} *and* $\widetilde{L}_{\mathrm{mx}}$, $\widetilde{\mu}_{\mathrm{mn}}$, D_{mx} *are defined as in* (2.4) *and* (3.12)*, 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 while adding and subtracting $\bar{\mathbf{y}}^\nu$ yields

$$\begin{aligned}0&\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\|\\&\leq\left(\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)\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(\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\left(\nabla^2F(\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)\right)(\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\|\\&\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 $\|\mathbf{d}_i^\nu\|$:

$$\begin{aligned}D_i\|\mathbf{x}^\star-\mathbf{x}_i^\nu\|+\|\bar{\mathbf{y}}^\nu-\mathbf{y}_i^\nu\|+\left(\frac{1}{m}\sum_{j=1}^m L_j\left\|\mathbf{x}_i^\nu-\mathbf{x}_j^\nu\right\|\right)\\ \geq\widetilde{\mu}_i\|\widehat{\mathbf{x}}_i^\nu-\mathbf{x}^\star\|\geq\widetilde{\mu}_i\left(\|\mathbf{d}_i^\nu\|-\|\mathbf{x}^\star-\mathbf{x}_i^\nu\|\right).\end{aligned} \tag{3.36}$$

Therefore,

$$\|\mathbf{d}_i^\nu\|^2\leq\ 3\left(\frac{D_i}{\widetilde{\mu}_i}+1\right)^2\|\mathbf{x}^\star-\mathbf{x}_i^\nu\|^2+\frac{3}{\widetilde{\mu}_i^2}\|\bar{\mathbf{y}}^\nu-\mathbf{y}_i^\nu\|^2+\frac{3}{\widetilde{\mu}_i^2}\left(\frac{1}{m}\sum_{j=1}^m L_j\left\|\mathbf{x}_i^\nu-\mathbf{x}_j^\nu\right\|\right)^2$$

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

Summing over $i=1,\dots,m$ and using the μ-strong convexity of U completes the proof. ∎

3.3.4. Step 4: Proof of the linear rate (chaining the inequalities). We are now ready to prove the linear rate of the SONATA algorithm. We build on the following intermediate result, introduced in [24].

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

$$S^K(z)\triangleq\max_{\nu=0,\dots,K}|s^\nu|z^{-\nu}\quad\text{and}\quad S(z)\triangleq\sup_{\nu\in\mathbb{N}}|s^\nu|z^{-\nu}\tag{3.37}$$

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 (3.28), (3.34), and (3.35) 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* (3.37) *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:*

$$P^K(z)\leq G_P(\alpha,z)\cdot\left(4L_{\rm mx}^2X_\perp^K(z)+2Y_\perp^K(z)\right)+\omega_p,\tag{3.38a}$$

$$X_\perp^K(z)\leq G_X(z)\cdot\rho^2\alpha^2D^K(z)+\omega_x,\tag{3.38b}$$

$$Y_\perp^K(z)\leq G_Y(z)\cdot 8L_{\rm mx}^2\rho^2X_\perp^K(z)+G_Y(z)\cdot 2L_{\rm mx}^2\rho^2\alpha^2D^K(z)+\omega_y,\tag{3.38c}$$

$$D^K(z)\leq C_1\cdot P^K(z)+C_2\cdot Y_\perp^K(z),\tag{3.38d}$$

for all

$$z\in\left(\max\{\sigma(\alpha),\rho^2(1+\epsilon_x),\rho^2(1+\epsilon_y)\},1\right),\tag{3.39}$$

where

$$G_P(\alpha,z)\triangleq\frac{\eta(\alpha)}{z-\sigma(\alpha)},\qquad\omega_p\triangleq\frac{z}{z-\sigma(\alpha)}\cdot p^0,\tag{3.40a}$$

$$G_X(z)\triangleq\frac{(1+\epsilon_x^{-1})}{z-\rho^2(1+\epsilon_x)},\qquad\omega_x\triangleq\frac{z}{z-\rho^2(1+\epsilon_x)}\cdot\|\mathbf{x}_\perp^0\|^2,\tag{3.40b}$$

$$G_Y(z)\triangleq\frac{(1+\epsilon_y^{-1})}{z-\rho^2(1+\epsilon_y)},\qquad\omega_y\triangleq\frac{z}{z-\rho^2(1+\epsilon_y)}\cdot\|\mathbf{y}_\perp^0\|^2,\tag{3.40c}$$

$$C_1\triangleq\frac{6}{\mu}\left(\left(\frac{D_{\rm mx}}{\widetilde{\mu}_{\rm mn}}+1\right)^2+\frac{4L_{\rm mx}^2}{\widetilde{\mu}_{\rm mn}^2}\right),\qquad C_2\triangleq\frac{4}{\widetilde{\mu}_{\rm mn}^2}.\tag{3.40d}$$

Proof. Squaring (3.34) and using Young's inequality yields

$$\begin{aligned}&\|\mathbf{x}_\perp^{\nu+1}\|^2\leq\rho^2(1+\epsilon_x)\|\mathbf{x}_\perp^\nu\|^2+\rho^2(1+\epsilon_x^{-1})\alpha^2\|\mathbf{d}^\nu\|^2\\&\|\mathbf{y}_\perp^{\nu+1}\|^2\leq\rho^2(1+\epsilon_y)\|\mathbf{y}_\perp^\nu\|^2+\rho^2(1+\epsilon_y^{-1})\Big(8L_{\rm mx}^2\|\mathbf{x}_\perp^\nu\|^2+2\alpha^2L_{\rm mx}^2\|\mathbf{d}^\nu\|^2\Big)\end{aligned}\tag{3.41}$$

for arbitrary $\epsilon_x,\epsilon_y>0$. The proof is completed by taking the maximum of both sides of (3.28), (3.35), and (3.41) over $\nu=0,\dots,K$ and using $\max_{\nu=0,\dots,K}|s^{\nu+1}|z^{-\nu}\geq z\cdot\max_{\nu=0,\dots,K}|s^\nu|\,z^{-\nu}-z\cdot|s^0|$ for any sequence $\{s^\nu\}$ and $z\in(0,1)$. ∎

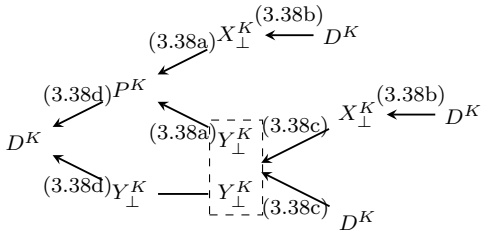

FIG. 1. *Chain of the inequalities in Proposition* 3.8 *leading to* (3.42).

Chaining the inequalities in Proposition 3.8 in the way shown in Figure 1, we can bound $D^K(z)$ as (see Appendix A for the proof)

$$D^K(z) \leq \mathcal{P}(\alpha, z) \cdot D^K(z) + \mathcal{R}(\alpha, z), \tag{3.42}$$

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

$$\begin{aligned} \mathcal{P}(\alpha, z) \triangleq\ & G_P(\alpha, z) \cdot G_X(z) \cdot C_1 \cdot 4L_{\mathrm{mx}}^2 \cdot \rho^2 \cdot \alpha^2 \\ & + (G_P(\alpha, z) \cdot 2C_1 + C_2) \cdot G_Y(z) \cdot 2L_{\mathrm{mx}}^2 \rho^2 \cdot \alpha^2 \\ & + (G_P(\alpha, z) \cdot 2C_1 + C_2) \cdot G_Y(z) \cdot 8L_{\mathrm{mx}}^2 \rho^2 \cdot G_X(z) \cdot \rho^2 \cdot \alpha^2, \end{aligned} \tag{3.43}$$

and $\mathcal{R}(\alpha, z)$ is a remainder, which is bounded under (3.39).

Therefore, as long as $\mathcal{P}(\alpha, z) < 1$, (3.42) implies

$$D^K(z) \leq \frac{\mathcal{R}(\alpha, z)}{1 - \mathcal{P}(\alpha, z)} < +\infty, \tag{3.44}$$

and thus $\{\|\mathbf{d}^\nu\|^2\}$ vanishes R-linearly 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 in showing that there exist $\alpha \in (0, 1]$ and $z \in (0, 1)$ satisfying (3.39), such that $\mathcal{P}(\alpha, z) < 1$. This is proved in Theorem 3.9 below.

THEOREM 3.9. *Consider problem* (P) *under Assumptions* A–B*, and the SONATA algorithm* (3.1a)–(3.1d)*, under Assumptions* C *and* D*, with* $\widetilde{\mu}_{\mathrm{mn}} \geq D_{\mathrm{mn}}^\ell$*. 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 consists of the following two steps: *Step* 1: We first consider the "marginal" stable case by letting $z = 1$, and we 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,

$$\epsilon^\star = \operatorname*{argmin}_{\epsilon > 0} \frac{1 + \epsilon^{-1}}{1 - \rho^2(1 + \epsilon)} = \frac{1 - \rho}{\rho}. \tag{3.45}$$

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

(3.22), the problem boils down to minimizing

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

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

$$\begin{aligned}\mathcal{P}^{\star}(\alpha,1)\triangleq\;&G_P^{\star}(\alpha)\cdot C_1\cdot 4L_{\rm mx}^2\cdot\frac{\rho^2}{(1-\rho)^2}\cdot\alpha^2\\&+(G_P^{\star}(\alpha)\cdot 2C_1+C_2)\cdot 2L_{\rm mx}^2\cdot\frac{\rho^2}{(1-\rho)^2}\cdot\alpha^2\\&+(G_P^{\star}(\alpha)\cdot 2C_1+C_2)\cdot 8L_{\rm mx}^2\cdot\frac{\rho^4}{(1-\rho)^4}\cdot\alpha^2,\end{aligned}\tag{3.46}$$

where

$$G_P^{\star}(\alpha)\triangleq\frac{\frac{D_{\rm mx}^2}{\mu}+\frac{1}{\mu}\cdot\left(\left(1-\frac{\alpha}{2}\right)\widetilde{\mu}_{\rm mn}+\frac{D_{\rm mn}^{\ell}}{2}\alpha\right)^2}{\left(\left(1-\frac{\alpha}{2}\right)\widetilde{\mu}_{\rm mn}+\frac{D_{\rm mn}^{\ell}}{2}\alpha\right)^2}.\tag{3.47}$$

• *Step* 2. Since $\mathcal{P}^{\star}(\bullet,1)$ is continuous and monotonically increasing on $(0,2\widetilde{\mu}_{\rm mn}/(\widetilde{\mu}_{\rm mn}-D_{\rm mn}^{\ell})$, with $\mathcal{P}^{\star}(0,1)=0$, there exists some $\bar{\alpha}<2\widetilde{\mu}_{\rm mn}/(\widetilde{\mu}_{\rm mn}-D_{\rm mn}^{\ell})$ such that $\mathcal{P}^{\star}(\alpha,1)<1$, for all $\alpha\in(0,\bar{\alpha})$. One can verify that, for any $\alpha\in(0,2\widetilde{\mu}_{\rm mn}/(\widetilde{\mu}_{\rm mn}-D_{\rm mn}^{\ell}))$, $\mathcal{P}^{\star}(\alpha,z)$ 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 $\alpha\leq\widetilde{\mu}_{\rm mn}/(\widetilde{\mu}_{\rm mn}-D_{\rm mn}^{\ell})$, we upper bound $G_P^{\star}(\alpha)$ by $G_P^{\star}(\widetilde{\mu}_{\rm mn}/(\widetilde{\mu}_{\rm mn}-D_{\rm mn}^{\ell}))$. Using for $G_P^{\star}(\alpha)$ this upper bound in (3.46) and solving the resulting $\mathcal{P}^{\star}(\alpha,1)<1$ for α yields

$$\begin{aligned}\alpha<\alpha_1\triangleq\;&\left(G_P^{\star}\left(\frac{\widetilde{\mu}_{\rm mn}}{\widetilde{\mu}_{\rm mn}-D_{\rm mn}^{\ell}}\right)\cdot C_1\cdot 4L_{\rm mx}^2\cdot\frac{\rho^2}{(1-\rho)^2}\right.\\&+\left(G_P^{\star}\left(\frac{\widetilde{\mu}_{\rm mn}}{\widetilde{\mu}_{\rm mn}-D_{\rm mn}^{\ell}}\right)\cdot 2C_1+C_2\right)\cdot 2L_{\rm mx}^2\cdot\frac{\rho^2}{(1-\rho)^2}\\&\left.+\left(G_P^{\star}\left(\frac{\widetilde{\mu}_{\rm mn}}{\widetilde{\mu}_{\rm mn}-D_{\rm mn}^{\ell}}\right)\cdot 2C_1+C_2\right)\cdot 8L_{\rm mx}^2\cdot\frac{\rho^4}{(1-\rho)^4}\right)^{-\frac{1}{2}}.\end{aligned}\tag{3.48}$$

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

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 (B.7), (B.5) with $\theta=1/2$, and (B.9), respectively.

THEOREM 3.10. *In the setting of Theorem* 3.9*, suppose that the step-size* α *satisfies* $\alpha\in(0,\alpha_{\rm mx})$*, with* $\alpha_{\rm mx}\triangleq\min\{(1-\rho)^2/A_{\frac{1}{2}},\widetilde{\mu}_{\rm mn}/(\widetilde{\mu}_{\rm mn}-D_{\rm mn}),1\}$*. Then,*

$U(\mathbf{x}_i^\nu) - U^\star = \mathcal{O}(z^\nu)$, *for all* $i \in [m]$, *where*

$$z = \begin{cases} 1 - J \cdot \alpha & \text{for } \alpha \in (0, \min\{\alpha^*, \alpha_{\rm mx}\}), \\ \left(\rho + \sqrt{\alpha A_{\frac{1}{2}}}\right)^2 & \text{for } \alpha \in [\min\{\alpha^*, \alpha_{\rm mx}\}, \alpha_{\rm mx}). \end{cases} \tag{3.49}$$

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. section 3.4.1) and then move to general graph topologies with no master node (cf. section 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.*

$$\widetilde{f}_i(\mathbf{x}_i; \mathbf{x}_i^\nu) \triangleq \nabla f_i(\mathbf{x}_i^\nu)^\top (\mathbf{x}_i - \mathbf{x}_i^\nu) + \frac{L}{2}\|\mathbf{x}_i - \mathbf{x}_i^\nu\|^2; \tag{3.50}$$

- *Local* f_i.

$$\widetilde{f}_i(\mathbf{x}_i; \mathbf{x}_i^\nu) \triangleq f_i(\mathbf{x}_i) + \frac{\beta}{2}\|\mathbf{x}_i - \mathbf{x}_i^\nu\|^2. \tag{3.51}$$

3.4.1. Star-networks: SONATA-Star. Convergence of SONATA-Star (Algorithm 3.2) 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, based on the surrogate functions satisfying Assumption* C *and step-size* $\alpha \in (0, \min(2\widetilde{\mu}_{\rm mn}/(\widetilde{\mu}_{\rm mn} - D_{\rm mn}^\ell), 1)]$. *Then, for all* $i = 1, \ldots, m$,

$$U(\mathbf{x}^\nu) - U^\star = \mathcal{O}(z^\nu), \quad \text{with} \quad z = 1 - \alpha \cdot \frac{\left(1 - \frac{\alpha}{2}\right)\widetilde{\mu}_{\rm mn} + \frac{\alpha D_{\rm mn}^\ell}{2}}{\frac{D_{\rm mx}^2}{2\mu} + \left(1 - \frac{\alpha}{2}\right)\widetilde{\mu}_{\rm mn} + \frac{\alpha D_{\rm mn}^\ell}{2}}. \tag{3.52}$$

In particular, when the surrogates (3.50) *and* (3.51) *are employed along with* $\alpha = 1$, *the rate above reduces to the following expressions:*

- Linearization (3.50). $z \le 1 - \kappa_g^{-1}$. *Therefore,* $U(\mathbf{x}^\nu) - U^\star \le \epsilon$ *in at most* $\mathcal{O}(\kappa_g \log(1/\epsilon))$ *iterations (communications).*
- Local f_i (3.51).

$$z \le 1 - \frac{1}{1 + 4 \cdot \frac{\beta}{\mu} \cdot \min\{1, \frac{\beta}{\mu}\}}. \tag{3.53}$$

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

$$\begin{cases} \mathcal{O}\left(1 \cdot \log\left(1/\epsilon\right)\right) & \text{if } \beta \le \mu, \\ \mathcal{O}\left(\frac{\beta}{\mu} \cdot \log\left(1/\epsilon\right)\right) & \text{if } \beta > \mu \end{cases} \tag{3.54}$$

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, (3.53)–(3.54) proves that faster rates can be achieved if surrogates (3.51) are chosen over first-order approximations. As a case study, consider the problem discussed in section 2.1.2: plugging (2.8) into Corollary 3.11 shows that using the surrogates (3.51) 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 and CEASE. For quadratic losses (and $G=0$), the rate of DANE, $\mathcal{O}\big((\beta/\mu)^2\cdot\log(1/\epsilon)\big)$, is worse than (3.54). For nonquadratic losses, [38] 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 (3.54). Comparing to the rate achievable by CEASE, $\mathcal{O}\big((\beta/\mu)^2\cdot\log(1/\epsilon)\big)$, SONATA improves by a factor β/μ, which matches the order of the mirror-descent algorithm.

3.4.2. The general case. The convergence rate of SONATA over general graphs is summarized in Corollary 3.12 for the linearization surrogates (3.50), while Corollaries 3.13 and 3.14 consider the surrogates (3.51) based on local f_i, with Corollary 3.13 addressing the case $\beta\le\mu$ 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* (3.50) *and step-size* $\alpha=c\cdot\alpha_{\rm mx}$*,* $c\in(0,1)$*, with* $\alpha_{\rm mx}=\min\{1,(1-\rho)^2/(\rho\cdot 110\kappa_g(1+\beta/L)^2)\}$*. The number of iterations (communications) needed for* $U(\mathbf{x}_i^\nu)-U^\star\le\epsilon$*,* $i\in[m]$*, is*

(3.55)

$$\text{Case I.}\qquad \mathcal{O}\left(\kappa_g\log(1/\epsilon)\right),\qquad \text{if}\quad \frac{\rho}{(1-\rho)^2}\le\frac{1}{110\,\kappa_g\,\left(1+\frac{\beta}{L}\right)^2},$$

(3.56)

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

Proof. See Appendix D. ∎

COROLLARY 3.13 (local f_i, $\beta\le\mu$). *Instate the assumptions of Theorem* 3.10 *and suppose* $\beta\le\mu$*. Consider SONATA using the surrogates* (3.51) *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* $M=193(1+\frac{\beta}{\mu})^2(\kappa_g+\frac{\beta}{\mu})^2$*. The number of iterations (communications) needed for* $U(\mathbf{x}_i^\nu)-U^\star\le\epsilon$*,* $i\in[m]$*, is*

(3.57)

$$\text{Case I.}\qquad \mathcal{O}\left(1\cdot\log(1/\epsilon)\right),\qquad \text{if}\quad \frac{\rho}{(1-\rho)^2}\le\frac{1}{193\left(1+\frac{\beta}{\mu}\right)^2\left(\kappa_g+\frac{\beta}{\mu}\right)^2},$$

(3.58)

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

COROLLARY 3.14 (local f_i, $\beta > \mu$). *Instate the assumptions of Theorem* 3.10 *and suppose* $\beta > \mu$. *Consider SONATA using the surrogates* (3.51) *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* $M = 253(1+\frac{L}{\beta})(\kappa_g+\frac{\beta}{\mu})$. *The number of iterations (communications) needed for* $U(\mathbf{x}_i^\nu) - U^\star \le \epsilon$, $i \in [m]$, *is*

(3.59)

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

(3.60)

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

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

Several comments are in order.

• *Rate of centralized (nonaccelerated) methods (Case* I*).* For a fixed optimization problem, if the network is sufficiently connected (ρ "small"), the bottleneck on the rate is given by the optimization process; 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. (3.55)) and by the mirror-descent scheme when the local f_i's are used in the surrogates (cf. (3.57) and (3.59)).

• *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, 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 local quantities (2.3). The rate dependence on $(1-\rho)^{-2}$ can be improved leveraging accelerated consensus protocols, as discussed below.

• *Linearization* (3.50) *versus local* f_i (3.51) *surrogates.* As already observed in the setting of star-networks, the use of local losses as surrogates employs a form of preconditioning in 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 (3.51) provably reduces the iteration/communication complexity over linear models (3.50)—contrast (3.55) with (3.57) and (3.59). 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 (3.51) are no longer informative of the average-loss F, and using linearization might yield better rates. These design recommendations are supported by numerical results; see [42].

• *Multiple communications rounds and acceleration.* When the network is not sufficiently connected (as in Case I), one can still achieve iteration complexity of the order of that of centralized methods, at the cost of multiple, finite, rounds of communications per iteration. Specifically, let ρ_0 be the connectivity of the given network, associated with a given weight matrix $\mathbf{W}$; suppose we run K steps of communications per iteration (computation) in (3.31a)–(3.31b), each time using the weight matrix $\mathbf{W}$; 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 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 (3.50) 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 [3] based on a symmetric weight matrix $\mathbf{W}$: the final communication complexity of SONATA reads

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

We refer the reader to [42] for details on the implementation of Chebyshev polynomials in the communication steps of SONATA.

(2) *Local f_i surrogates.* Considering the case $\beta \geq \mu$ (Corollary 3.14), we can show that SONATA using the surrogates (3.51) 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. Concluding remarks. We studied convergence of the SONATA algorithm, solving composite optimization problems over mesh networks. For strongly convex sum-loss functions, the algorithm was proved to converge at a linear rate, which depends on the global condition number κ_g of the sum-loss; this improves on existing results showing a much more pessimistic dependence on optimization parameters. When the local losses are β-similar, faster rates (and thus communication savings) are provably achievable—scaling with β/μ—at the cost of higher local computations.

Some extensions of the SONATA framework worth mentioning include (i) the application to directed, time-varying digraphs [42]; (ii) the use of preconditioned Newton steps as local agents' updates, in substitution of surrogates (3.51) [5]; (iii) the acceleration of the plain algorithm in both unrelated scenarios and β-related scenarios, matching (up to log-factors) lower complexity bounds [44]; and (iv) the generalization to asynchronous modus operandi (still preserving linear convergence) [45].

Appendix A. Proof of (3.42). Chaining the inequalities in (3.38) as shown in Figure 1, we have

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

$$+ (C_1 \cdot G_P(\alpha, z) \cdot 2 + C_2) \cdot G_Y(z) \cdot 8L_{\rm mx}^2 \rho^2 \cdot \omega_x.$$

Notice that, under (3.39), $G_P(\alpha, z)$, $G_X(z)$, $G_Y(z)$ and ω_p, ω_x, ω_y are all bounded, which implies that the reminder $\mathcal{R}(\alpha, z)$ in (3.38) is bounded as well.

Appendix B. Proof of Theorem 3.10. We find the smallest z satisfying (3.39) 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 (3.39). To simplify the analysis, we impose instead the following stronger version:

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

for some $\theta \in (0, 1)$, which will be chosen to tighten the bound. Notice that the right-hand side of (B.1) is strictly larger than $\sigma(\alpha)$ but still strictly less than one, for any $\alpha \in (0, (2\widetilde{\mu}_{\rm mn} - \epsilon_{opt})/(\widetilde{\mu}_{\rm mn} - D_{\rm mn}^{\ell}))$, 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 (B.1), $G_P(\alpha, z)$ can be upper bounded as

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

where the minimum is attained at $\epsilon_{opt}^{\star} \triangleq \widetilde{\mu}_{\rm mn} - \frac{\alpha}{2}(\widetilde{\mu}_{\rm mn} - D_{\rm mn}^{\ell})$, and $G_P^{\star}(\alpha)$ is defined as in (3.47). Substituting the upper bound (B.2) into $\mathcal{P}(\alpha, z)$ and setting therein $\epsilon_{opt} = \epsilon_{opt}^{\star}$, we get the following sufficient condition for $\mathcal{P}(\alpha, z) < 1$:

$$\begin{aligned} &G_P^{\star}(\alpha) \cdot \theta^{-1} \cdot C_1 \cdot 4L_{\rm mx}^2 \cdot G_X(z) \cdot \rho^2 \cdot \alpha^2 \\ &\quad + \left(G_P^{\star}(\alpha) \cdot \theta^{-1} \cdot 2C_1 + C_2\right) \cdot G_Y(z) \cdot 2L_{\rm mx}^2 \rho^2 \cdot \alpha^2 \\ &\quad + \left(G_P^{\star}(\alpha) \cdot \theta^{-1} \cdot 2C_1 + C_2\right) \cdot G_Y(z) \cdot 8L_{\rm mx}^2 \rho^2 \cdot G_X(z) \cdot \rho^2 \cdot \alpha^2 < 1. \end{aligned} \tag{B.3}$$

To minimize the left-hand side, we set $\epsilon_x = \epsilon_y = (\sqrt{z} - \rho)/\rho$. Furthermore, using the fact that $G_P^{\star}(\alpha)$ is monotonically increasing on $\alpha \in (0, 2\widetilde{\mu}_{\rm mn}/(\widetilde{\mu}_{\rm mn} - D_{\rm mn}^{\ell}))$, and restricting $\alpha \in (0, \widetilde{\mu}_{\rm mn}/(\widetilde{\mu}_{\rm mn} - D_{\rm mn}^{\ell})]$, a sufficient condition for (B.3) is

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

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

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

Condition (B.4) shows the rate z must satisfy

$$z \geq \left(\rho + \sqrt{A_\theta \alpha}\right)^2, \quad \text{with} \quad A_\theta \triangleq \sqrt{A_{1,\theta} + A_{2,\theta} + A_{3,\theta}}. \tag{B.5}$$

Notice that, under $\epsilon_x = \epsilon_y = (\sqrt{z} - \rho)/\rho$, (B.5) implies $z > \rho^2(1+\epsilon_x) = \rho^2(1+\epsilon_y) = \rho\sqrt{z}$, which are the other two conditions on z in (3.39). Therefore, overall, z must satisfy (B.1) and (B.5). Letting $\epsilon_{opt} = \epsilon_{opt}^\star$ in (B.1), 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

(B.6)
$$\bar{z} = \inf_{\theta\in(0,1)} \max\left\{ \left(\rho + \sqrt{A_\theta \alpha}\right)^2, 1 - \frac{\widetilde{\mu}_{\rm mn} - \frac{\alpha}{2}(\widetilde{\mu}_{\rm mn} - D^\ell_{\rm mn})}{\frac{2D^2_{\rm mx}}{\mu} + \widetilde{\mu}_{\rm mn} - \frac{\alpha}{2}(\widetilde{\mu}_{\rm mn} - D^\ell_{\rm mn})} \cdot (1-\theta)\alpha \right\}.$$

Finally, we further simplify (B.6). 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 (B.6) can be upper bounded by

(B.7)
$$1 - \underbrace{\frac{\widetilde{\mu}_{\rm mn}\mu}{4D^2_{\rm mx} + \widetilde{\mu}_{\rm mn}\mu} \cdot \frac{1}{2}}_{\triangleq J} \alpha.$$

The condition $\bar{z} < 1$ imposes the following upper bound on α: $\alpha < \alpha_{\rm mx} = \min\{(1-\rho)^2/A_{\frac{1}{2}}, \widetilde{\mu}_{\rm mn}/(\widetilde{\mu}_{\rm mn} - D^\ell_{\rm mn}), 1\}$. Equation (B.6) then simplifies to

(B.8)
$$\bar{z} = \max\left\{ \left(\rho + \sqrt{\alpha A_{\frac{1}{2}}}\right)^2, 1 - J\alpha \right\}.$$

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

(B.9)
$$\alpha^* = \left(\frac{-\rho\sqrt{A_{\frac{1}{2}}} + \sqrt{A_{\frac{1}{2}} + J(1-\rho^2)}}{A_{\frac{1}{2}} + J} \right)^2.$$

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 as given in (3.49).

Appendix C. Proof of Corollary 3.11. Since $\mathbf{W} = \mathbf{J}$, we have $\boldsymbol{\delta}^\nu = \mathbf{0}$; then (3.21a) and (3.23) reduce to

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

and

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

respectively. Combining (C.1) and (C.2) and using $\alpha < 2\widetilde{\mu}_{\rm mn}/(\widetilde{\mu}_{\rm mn} - D_{\rm mn})$ yields

(C.3)
$$p^{\nu+1} \leq \left(1 - \alpha \cdot \frac{\left(1 - \frac{\alpha}{2}\right) \widetilde{\mu}_{\rm mn} + \frac{\alpha D^\ell_{\rm mn}}{2}}{\frac{D^2_{\rm mx}}{2\mu} + \left(1 - \frac{\alpha}{2}\right) \widetilde{\mu}_{\rm mn} + \frac{\alpha D^\ell_{\rm mn}}{2}} \right) p^\nu,$$

which proves (3.52).

We next customize (3.52) to the specific choices of the surrogate functions.

• *Linearization.* Consider the choice of $\widetilde{f}_i$ as in (3.50). We have $\widetilde{\mu}_{\rm mn} = L$; and we can set $D^{\ell}_{\rm mn} = 0$, $D_{\rm mx} = L - \mu$, and $\alpha = 1$. Substituting these values into (3.52), we obtain $z \leq 1 - \kappa_g^{-1}$.

• *Local f_i.* Consider now $\widetilde{f}_i$ as in (3.51). By $\nabla^2 f_i(\mathbf{x}) \succeq \mathbf{0}$ for all $\mathbf{x} \in \mathcal{K}$, and by 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 $D^{\ell}_{\rm mn} = 0$, $D_{\rm mx} = 2\beta$, and $\widetilde{\mu}_{\rm mn} = \beta + (\mu - \beta)_+$. Using these values in (3.52) yields

$$z \begin{cases} = 1 - \alpha \cdot \frac{\beta\left(1 - \frac{\alpha}{2}\right)}{\frac{2\beta^2}{\mu} + \beta\left(1 - \frac{\alpha}{2}\right)} & \text{if } \mu \leq \beta, \\ \leq 1 - \alpha \cdot \frac{\mu\left(1 - \frac{\alpha}{2}\right)}{\frac{2\beta^2}{\mu} + \mu\left(1 - \frac{\alpha}{2}\right)} & \text{if } \mu > \beta. \end{cases} \tag{C.4}$$

Setting $\alpha = \min\{1, 2\widetilde{\mu}_{\rm mn}/((\mu - \beta)_+ + \beta)\} = 1$ in the expression above yields (3.53).

Appendix D. Proof of Corollary 3.12. According to Theorem 3.10, the rate z can be bounded as

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

where J and $A_{\frac{1}{2}}$ are defined as in (B.7) and (B.5), respectively.

The proof consists in bounding properly z_1 and z_2 based upon the surrogate (3.50) 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 $\widetilde{\mu}_{\rm mn} = L$, and (3.4) holds with $D^{\ell}_{\rm mn} = 0$ and $D_{\rm mx} = L - \mu$. 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 $L_{\rm mx} = L + \beta$. Next, we will substitute the above values into the expressions of J and $A_{\frac{1}{2}}$.

To do so, we need to first particularize the quantities $G_P^{\star}(\frac{\widetilde{\mu}_{\rm mn}}{\widetilde{\mu}_{\rm mn} - D^{\ell}_{\rm mn}})$ (cf. (3.47)), C_1, and C_2 (cf. (3.40d)):

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

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

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

and

(D.3)
$$\begin{aligned}
(A_{\frac{1}{2}})^2 &= G_P^\star(1)\cdot 2\cdot C_1\cdot 4L_{\rm mx}^2\cdot\rho^2 + (G_P^\star(1)\cdot 4\cdot C_1 + C_2)\cdot 2L_{\rm mx}^2\rho^2 \\
&\quad + (G_P^\star(1)\cdot 4\cdot C_1 + C_2)\cdot 8L_{\rm mx}^2\rho^4 \\
&= (24G_P^\star(1)\cdot C_1 + 5C_2)\cdot 2L_{\rm mx}^2\rho^2 \\
&= \left(24\cdot\frac{4(L-\mu)^2+L^2}{\mu L^2}\cdot\frac{6}{\mu L^2}\left((2L-\mu)^2+4(L+\beta)^2\right)+20L^{-2}\right)\cdot 2(L+\beta)^2\rho^2 \\
&\le \left(24\cdot\frac{5}{\mu}\cdot\frac{24}{\mu L^2}\left(L^2+(L+\beta)^2\right)+20L^{-2}\right)\cdot 2(L+\beta)^2\rho^2 \\
&= \left(24\cdot 24\cdot 5\left(1+\left(1+\frac{\beta}{L}\right)^2\right)\left(1+\frac{\beta}{L}\right)^2\kappa_g^2 + 20\left(1+\frac{\beta}{L}\right)^2\right)\cdot 2\rho^2 \\
&\le 110^2\cdot\kappa_g^2\left(1+\frac{\beta}{L}\right)^4\rho^2,
\end{aligned}$$

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

Using the above expressions, in what follows we upper bound z_1 and z_2.

By (D.3), we have

$$z_2 \le \bar{z}_2 \triangleq \left(\rho+\sqrt{\alpha M\rho}\right)^2, \quad \text{with} \quad M \triangleq 110\cdot\kappa_g(1+\beta/L)^2. \tag{D.4}$$

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

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

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

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

Therefore, z can be bounded as

$$\begin{aligned}
z \le \max\{z_1,\bar{z}_2\} &\le 1 - c\cdot\left(1-\sqrt{c}\right)^2\cdot(1-1/110)^2\cdot J \\
&\le 1 - c\cdot\left(1-\sqrt{c}\right)^2\cdot(1-1/110)^2\cdot\frac{1}{8\kappa_g}.
\end{aligned} \tag{D.5}$$

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

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

We claim that $(J\,c)/(M\rho) < 1$. Suppose this is not the case, that is, $M\rho \le Jc$. Since $Jc < 1/2$ (cf. (D.2)) and $M \ge 110\,\kappa$, $M\rho \le Jc$ would imply $\rho < 1/(220\kappa_g)$.

This however is in contradiction to 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 prove the two corollaries together. We follow steps similar to those in Appendix D but customized to the surrogate (3.51). We begin by 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 $D^{\ell}_{\rm mn} = 0$, $D_{\rm mx} = 2\beta$, $\widetilde{\mu}_{\rm mn} = \beta + (\mu-\beta)_+ = \max\{\beta,\mu\}$, and $L_{\rm mx} = L + \beta$.

Using these values, $G^{\star}_P(\frac{\widetilde{\mu}_{\rm mn}}{\widetilde{\mu}_{\rm mn} - D^{\ell}_{\rm mn}})$, C_1, and C_2 can be simplified as follows:

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

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

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

and

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

where

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

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

$$z \leq \max\{z_1, \bar{z}_2\}, \quad \text{with} \quad z_1 \triangleq 1 - \alpha \cdot J \ \text{ and } \ \bar{z}_2 \triangleq \left(\rho + \sqrt{\alpha\, M\, \rho}\right)^2, \tag{E.3}$$

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

(1) $\beta > \mu$. In this case we have

$$M = 253\left(1+\frac{L}{\beta}\right)\left(\kappa_g + \frac{\beta}{\mu}\right) \quad \text{and} \quad J = \frac{1}{2}\frac{1}{1+16\,(\beta/\mu)} \ge \frac{1}{34(\beta/\mu)}. \tag{E.4}$$

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

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

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

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

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

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

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

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

$$\begin{aligned} z \le \max\{z_1, \bar{z}_2\} &\le 1 - \left(1-\sqrt{c}\right)^2 \cdot c\,J \cdot \frac{(1-\rho)^2}{M\rho} \\ &\overset{\text{(E.4)}}{\le} 1 - c\cdot\left(1-\sqrt{c}\right)^2 \cdot \frac{1}{34}\cdot\frac{(1-\rho)^2}{253\left(\kappa_g+\frac{\beta}{\mu}\right)^2\rho}. \end{aligned}$$

(2) $\beta \le \mu$. In this case we have

$$M = 193\left(1+\frac{\beta}{\mu}\right)^2\left(\kappa_g+\frac{\beta}{\mu}\right)^2 \quad \text{and} \quad J = \frac{1}{2}\frac{1}{1+16\,(\beta/\mu)^2}. \tag{E.5}$$

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

$$z \le \max\{z_1, \bar{z}_2\} \le 1 - c\cdot\left(1-\sqrt{c}\right)^2 \cdot \left(1-\frac{1}{193}\right)^2 \cdot \frac{1}{2+32\left(\frac{\beta}{\mu}\right)^2}. \tag{E.6}$$

- *Case* II. $\alpha_{\rm mx} = (1-\rho)^2/(M\rho)$. 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 of 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*, preprint, https://arxiv.org/abs/1905.07996, 2019.

[2] Y. Arjevani and O. Shamir, *Communication complexity of distributed convex learning and optimization*, in Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS), Vol. 1, MIT Press, Cambridge, MA, 2015, pp. 1756–1764.

[3] A. Auzinger and J. M. Melenk, *Iterative Solution of Large Linear Systems*, Lecture notes, TU Wien, 2011, https://www.asc.tuwien.ac.at/~winfried/teaching/106.079/SS2017/downloads/iter.pdf.

[4] A. Berahas, R. Bollapragada, N. S. Keskar, and E. Wei, *Balancing communication and computation in distributed optimization*, IEEE Trans. Automat. Control, 64 (2019), pp. 3141–3155.

[5] A. Daneshmand, G. Scutari, P. Dvurechensky, and A. Gasnikov, *Newton method over networks is fast up to the statistical precision*, in Proceedings of the 38th International Conference on Machine Learning (ICML 2021), 2021, pp. 2398–2409.

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

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

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

[9] J. Fan, Y. Guo, and K. Wang, *Communication-Efficient Accurate Statistical Estimation*, preprint, https://arxiv.org/abs/1906.04870, 2019.

[10] H. Hendrikx, L. Xiao, S. Bubeck, F. Bach, and L. Massoulie, *Statistically preconditioned accelerated gradient method for distributed optimization*, in International Conference on Machine Learning, PMLR, 2020, pp. 4203–4227.

[11] D. Jakovetic, *A unification and generalization of exact distributed first-order methods*, IEEE Trans. Signal Inform. Process. Netw., 5 (2019), pp. 31–46.

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

[13] 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.

[14] X. Jinming, Y. Tian, Y. Sun, and G. Scutari, *Distributed algorithms for composite optimization: Unified and tight convergence analysis*, IEEE Trans. Signal Process., 69 (2020), pp. 3555–3570.

[15] B. Li, S. Cen, Y. Chen, and Y. Chi, *Communication-Efficient Distributed Optimization in Networks with Gradient Tracking and Variance Reduction*, preprint, https://arxiv.org/abs/1909.05844v3, 2019.

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

[17] 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.

[18] 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.

[19] H. Lu, R. M. Freund, and Y. Nesterov, *Relatively smooth convex optimization by first-order methods, and applications*, SIAM J. Optim., 28 (2018), pp. 333–354, https://doi.org/10.1137/16M1099546.

[20] M. Maros and J. Jalden, *PANDA: A dual linearly converging method for distributed optimization over time-varying undirected graphs*, in Proceedings of the 2018 IEEE Conference on Decision and Control (CDC), IEEE, Washington, DC, 2018, pp. 6520–6525.

[21] M. Maros and J. Jalden, *On the Q-linear convergence of distributed generalized ADMM under non-strongly convex function components*, IEEE Trans. Signal Inform. Process. Netw., 5 (2019), pp. 442–453.

[22] A. Mokhtari, W. Shi, Q. Ling, and A. Ribeiro, *DQM: Decentralized quadratically approxi-*

mated alternating direction method of multipliers, IEEE Trans. Signal Process., 64 (2016), pp. 5158–5173.

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

[24] A. Nedić, A. Olshevsky, and W. Shi, *Achieving geometric convergence for distributed optimization over time-varying graphs*, SIAM J. Optim., 27 (2017), pp. 2597–2633, https://doi.org/10.1137/16M1084316.

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

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

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

[28] S. Pu, W. Shi, J. Xu, and A. Nedić, *A push-pull gradient method for distributed optimization in networks*, in Proceedings of the 2018 IEEE Conference on Decision and Control (CDC), IEEE, Washington, DC, 2018, pp. 3385–3390.

[29] G. Qu and N. Li, *Accelerated distributed Nesterov gradient descent for smooth and strongly convex functions*, in Proceedings of the 54th Annual Allerton Conference on Communication, Control, and Computing, Allerton, IL, 2016, pp. 209–216.

[30] G. Qu and N. Li, *Harnessing smoothness to accelerate distributed optimization*, IEEE Trans. Control Netw. Syst., 5 (2018), pp. 1245–1260.

[31] A. Rogozin and A. Gasnikov, *Projected Gradient Method for Decentralized Optimization over Time-Varying Networks*, preprint, https://arxiv.org/abs/1911.08527, 2019.

[32] F. Saadatniaki, R. Xin, and U. A. Khan, *Decentralized optimization over time-varying directed graphs with row and column-stochastic matrices*, IEEE Trans. Automat. Control, 65 (2020), pp. 4769–4780.

[33] K. Scaman, F. Bach, S. Bubeck, Y. T. Lee, and L. Massoulié, *Optimal algorithms for smooth and strongly convex distributed optimization in networks*, in Proceedings of the 34th International Conference on Machine Learning, Vol. 70, Sydney, Australia, 2017, pp. 3027–3036.

[34] G. Scutari, F. Facchinei, and L. Lampariello, *Parallel and distributed methods for constrained nonconvex optimization—part* I: *Theory*, IEEE Trans. Signal Process., 65 (2017), pp. 1929–1944.

[35] G. Scutari and Y. Sun, *Parallel and distributed successive convex approximation methods for big-data optimization*, in Multi-agent Optimization, Lecture Notes in Math. 2224, Springer, Cham, 2018, pp. 141–308.

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

[37] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan, *Stochastic convex optimization*, in Proceedings of the 22nd Annual Conference on Learning Theory (COLT), Montreal, Canada, 2009.

[38] O. Shamir, N. Srebro, and T. Zhang, *Communication-efficient distributed optimization using an approximate newton-type method*, in Proceedings of the 31st International Conference on Machine Learning (PMLR), Vol. 32, 2014, pp. 1000–1008.

[39] 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, https://doi.org/10.1137/14096668X.

[40] 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.

[41] 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.

[42] Y. Sun, A. Daneshmand, and G. Scutari, *Distributed Optimization Based on Gradient-Tracking Revisited: Enhancing Convergence Rate via Surrogation*, preprint, https://arxiv.org/abs/1905.02637, 2019.

[43] Y. Sun, G. Scutari, and D. Palomar, *Distributed nonconvex multiagent optimization over time-varying networks*, in Proceedings of the Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, 2016.

[44] Y. Tian, G. Scutari, T. Cao, and A. Gasnikov, *Acceleration in distributed optimization under similarity*, in Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS 2022), March 28–30, 2022 (virtual conference).

[45] Y. Tian, Y. Sun, and G. Scutari, *Achieving linear convergence in distributed asynchronous multi-agent optimization*, IEEE Trans. Automat. Control, 65 (2020), pp. 5264–5279.
[46] J. Tsitsiklis, *Problems in Decentralized Decision Making and Computation*, Ph.D. dissertation, Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA, 1984.
[47] C. Xi and U. A. Khan, *ADD-OPT: Accelerated distributed directed optimization*, IEEE Trans. Automat. 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 Control Syst. Lett., 2 (2018), pp. 315–320.
[49] 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. Automat. Control, 63 (2018), pp. 3558–3565.
[50] L. Xiao, S. Boyd, and S. Lall, *A scheme for robust distributed sensor fusion based on average consensus*, in Proceedings of the 4th International Symposium on Information Processing in Sensor Networks, Los Angeles, CA, 2005, pp. 63–70.
[51] R. Xin, D. Jakovetic, and U. A. Khan, *Distributed Nesterov Gradient Methods over Arbitrary Graphs*, preprint, https://arxiv.org/abs/1901.06995, 2019.
[52] R. Xin and U. A. Khan, *Distributed Heavy-Ball: A Generalization and Acceleration of First-Order Methods with Gradient Tracking*, preprint, https://arxiv.org/abs/1808.02942, 2018.
[53] J. Xu, S. Zhu, Y. C. Soh, and L. Xie, *Augmented distributed gradient methods for multi-agent optimization under uncoordinated constant stepsizes*, in Proceedings of the 54th IEEE Conference on Decision and Control (CDC 2015), Osaka, Japan, 2015, pp. 2055–2060.
[54] J. Xu, S. Zhu, Y. C. Soh, and L. Xie, *Convergence of asynchronous distributed gradient methods over stochastic networks*, IEEE Trans. Automat. Control, 63 (2018), pp. 434–448.
[55] K. Yuan, Q. Ling, and W. Yin, *On the convergence of decentralized gradient descent*, SIAM J. Optim., 26 (2016), pp. 1835–1854, https://doi.org/10.1137/130943170.
[56] K. Yuan, B. Ying, X. Zhao, and A. H. Sayed, *Exact diffusion for distributed optimization and learning—Part* II: *Convergence analysis*, IEEE Trans. Signal Process., 67 (2019), pp. 724–739.
[57] J. Zeng and W. Yin, *ExtraPush for convex smooth decentralized optimization over directed networks*, J. Comput. Math., 35 (2017), pp. 383–396.
[58] Y. Zhang and X. Lin, *DISCO: Distributed optimization for self-concordant empirical loss*, in Proceedings of the 32nd International Conference on Machine Learning (PMLR), Vol. 37, 2015, pp. 362–370.
[59] Y. Zhang and L. Xiao, *Communication-efficient distributed optimization of self-concordant empirical loss*, in Large-Scale and Distributed Optimization, Lecture Notes in Math. 2227, Springer, Cham, 2018, pp. 289–341.