PARALLEL AND DECENTRALIZED ALGORITHMS FOR BIG-DATA OPTIMIZATION OVER NETWORKS

by

Amir Daneshmand

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

School of Industrial Engineering

West Lafayette, Indiana

August 2021

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Gesualdo Scutari, Chair

School of Industrial Engineering

Dr. Andrew (Lu) Liu

School of Industrial Engineering

Dr. Shaoshuai Mou

School of Aeronautics and Astronautics

Dr. Shreyas Sundaram

School of Electrical and Computer Engineering

Approved by:

Dr. Abhijit Deshmukh

To my mother

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Recent decades have witnessed the rise of data deluge generated by heterogeneous sources, e.g., social networks, streaming, marketing services etc., which has naturally created a surge of interests in theory and applications of large-scale convex and non-convex optimization. For example, real-world instances of statistical learning problems such as deep learning, recommendation systems, etc. can generate sheer volumes of spatially/temporally diverse data (up to Petabytes of data in commercial applications) with millions of decision variables to be optimized. Such problems are often referred to as *Big-data* problems. Solving these problems by standard optimization methods demands intractable amount of *centralized* storage and computational resources which is infeasible and is the foremost purpose of *parallel* and *decentralized* algorithms developed in this thesis.

This thesis consists of two parts: (I) Distributed Nonconvex Optimization and (II) Distributed Convex Optimization.

In Part (I), we start by studying a winning paradigm in big-data optimization, Block Coordinate Descent (BCD) algorithm, which cease to be effective when problem dimensions grow overwhelmingly. In particular, we considered a general family of constrained non-convex composite large-scale problems defined on multicore computing machines equipped with shared memory. We design a hybrid deterministic/random parallel algorithm to efficiently solve such problems combining synergically Successive Convex Approximation (SCA) with greedy/random dimensionality reduction techniques. We provide theoretical and empirical results showing efficacy of the proposed scheme in face of huge-scale problems.

The next step is to broaden the network setting to general mesh networks modeled as directed graphs, and propose a class of gradient-tracking based algorithms with global convergence guarantees to critical points of the problem. We further explore the geometry of the landscape of the non-convex problems to establish second-order guarantees and strengthen our convergence to local optimal solutions results to global optimal solutions for a wide range of Machine Learning problems.

In Part (II), we focus on a family of distributed convex optimization problems defined over meshed networks. Relevant state-of-the-art algorithms often consider limited prob-

lem settings with pessimistic communication complexities with respect to the complexity of their centralized variants, which raises an important question: can one achieve the rate of centralized first-order methods over networks, and moreover, can one improve upon their communication costs by using higher-order local solvers? To answer these questions, we proposed an algorithm that utilizes surrogate objective functions in local solvers (hence going beyond first-order realms, such as proximal-gradient) coupled with a perturbed (push-sum) consensus mechanism that aims to track locally the gradient of the central objective function. The algorithm is proved to match the convergence rate of its centralized counterparts, up to multiplying network factors. When considering in particular, Empirical Risk Minimization (ERM) problems with statistically homogeneous data across the agents, our algorithm employing high-order surrogates provably achieves faster rates than what is achievable by first-order methods. Such improvements are made without exchanging any Hessian matrices over the network.

Finally, we focus on the ill-conditioning issue impacting the efficiency of decentralized first-order methods over networks which rendered them impractical both in terms of computation and communication cost. A natural solution is to develop distributed second-order methods, but their requisite for Hessian information incurs substantial communication overheads on the network. To work around such exorbitant communication costs, we propose a *"statistically informed"* preconditioned cubic regularized Newton method which provably improves upon the rates of first-order methods. The proposed scheme does not require communication of Hessian information in the network, and yet, achieves the iteration complexity of centralized second-order methods up to the statistical precision. In addition, (second-order) approximate nature of the utilized surrogate functions, improves upon the per-iteration computational cost of our earlier proposed scheme in this setting.

1. INTRODUCTION AND MOTIVATIONS

Over the recent decades, rapid advancement of social networks, digital systems, communication and sensing technologies have led to the rise of *distributed* systems including the Internet, mobile ad hoc networks, and wireless sensor networks. Consequently, these systems have given rise to new network application domains, such as sensor networks, data-based networks, robotic networks, unmanned aerial vehicle systems, and smart grid networks; see e.g. [1]–[6]. Such applications usually call for in-network control and optimization techniques to perform various operations, including resource allocation, coordination, learning, and estimation.

More specifically, these systems are composed of a large number of interconnected subsystems (nodes or agents) which are required to communicate and cooperate to accomplish a joint global objective. The topology of such interconnected networks can vary for different applications or could be imposed by connectivity restrictions which distinguishes the realm of distributed and *centralized* systems. If all the agents communicate through a main *center* node, we call it centralized system (equivalent to master-worker type networks); on the other hand, if the network is ad-hoc and all the agents are treated as identical *peers* who can only communicate with their immediate neighbors, we call it a distributed (or *decentralized*)[1] system (equivalent to ad-hoc or peer-to-peer systems termed in the literature).

Objective of above distributed applications are usually formulated as convex or non-convex optimization problems where the ultimate goal is that all the agents compute (an acceptable approximation of) the solution of such problem cooperatively over the network. However, designing solution methods to accomplish such goals faces multiple challenges, as outlined next.

1.1 Challenges and Motivations

• **Big-data:** Distributed optimization/problems in context of big-data applications usually deal with huge amount of data, partially stored in each of the network agents and, as well as, large number of optimization variables (see Fig. 1.1). Dealing with such enormity

[1]↑Throughout the thesis, we use "decentralized" and "distributed", interchangeably.

sets the path to developing *decentralized* and *parallel* algorithms that exploit *horizontal* and *vertical* scaling of the systems to cope with the curse of dimensionality and accommodate the need for fast (real-time) processing and optimization. For example, a properly designed parallel method can utilize hierarchical computational architectures (e.g., multicore systems, cluster computers, cloud-based networks), if available, to reduce the computation time. The challenge is that such optimization problems are in general not separable in the optimization variables, which makes the design of parallel and/or decentralized schemes not a trivial task.

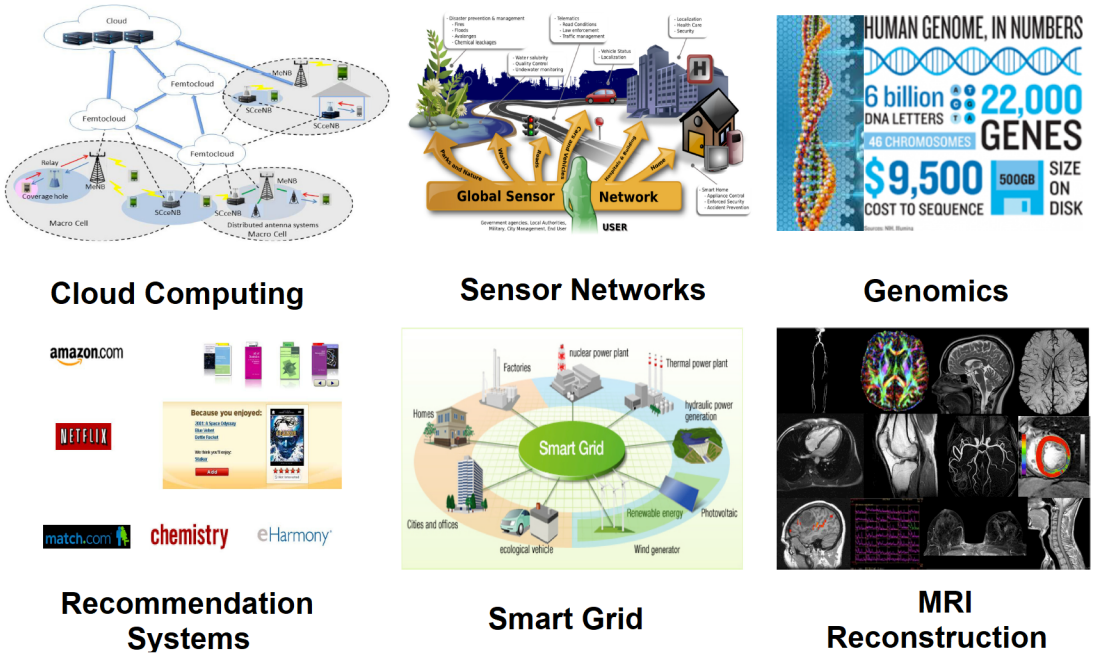


Figure 1.1. Big-data applications.

- **Non-convexity:** Many well-know applications lead to *non-convex* problems defined over networks. Such problems are in general NP-hard [7], [8], meaning that computing the global optimal solution might be computationally prohibitive in several practical applications. Such distributed non-convex problems have found a wide range of applications in several areas, including network information processing, machine learning, communications, and multi-agent control; see, e.g., [6]. The goal is to design parallel and distributed algorithms that are efficient and are guaranteed to converge to critical points (or more favorably to local or global minima) of the non-convex problem.

To this regard, two main ideas can be utilized, mainly: (i) the so-called *Successive Convex Approximation (SCA)* technique: as proxy of the non-convex problem, a sequence of "more tractable" (possibly convex) subproblems is solved, wherein the original non-convex functions are replaced by properly chosen "simpler" surrogates. By tailoring the choice of the surrogate functions to the specific structure of the optimization problem under consideration, SCA techniques offer remarkable freedom and flexibility in the algorithmic design; (ii) exploiting the geometric landscape of non-convex problems: it has been revealed that a wide range of Machine Learning problems, despite the non-convexity, possess some favorable geometry [7], [9], which enables many prominent algorithms such as gradient descent (GD) and alternating directions (AD) methods to be quite effective when applied to non-convex problems. Leveraging such properties, one may design well-tailored algorithms to solve non-convex problems to their local (or global) minima.

• **In-networked optimization:** Such networked systems are typically spatially distributed over a large area (or virtually distributed). Due to the network size (hundreds to millions of agents), and often to proprietary regulations, these systems do not possess a single central coordinator or access point with complete information, able to solve the entire optimization problem. Network/data information is instead distributed among the entities comprising the network (cf. Fig. 1.2). Furthermore, there are some networks such as surveillance networks or some cyber-physical systems where a centralized architecture is not desirable, as it makes the system prone to central entity failures and external attacks. Additional challenges are encountered from the network topology and connectivity that can be time-varying, due to, e.g., link failures, power outage, and agents' mobility. In this setting, the goal is to develop distributed solution methods that operate seamless in-network, by leveraging the network connectivity and local information (e.g., neighbor information) to cope with the lack of global knowledge on the optimization problem and offer robustness to possible failures/attacks of central units and/or to time-varying connectivity.

• **Ill-conditioning and communication burdens:** Ill-conditioned problems have been a long-time nemesis of first-order methods, rendering them inefficient in terms of communication cost when utilized over networks. A natural solution, as explored in the literature, is to utilize higher-order methods in distributed systems, but their requisite for the complete

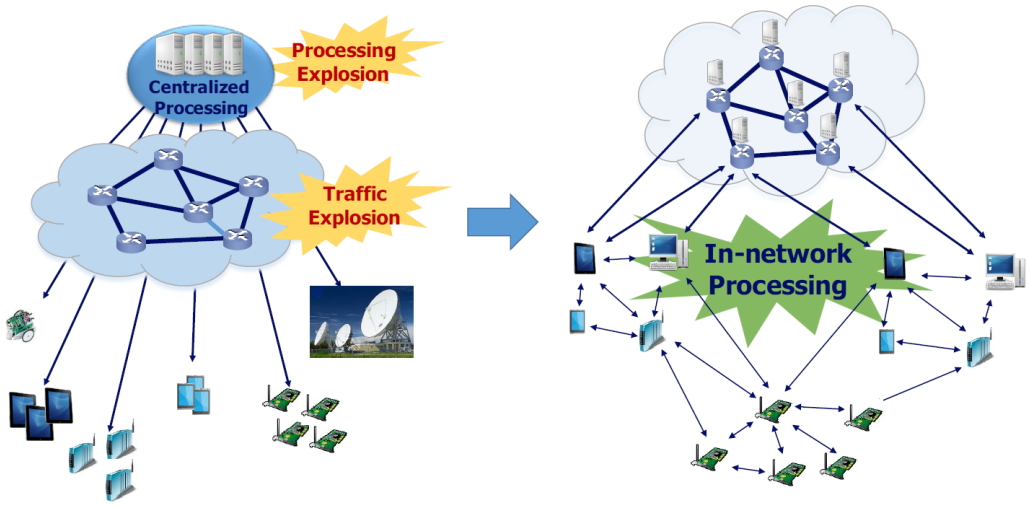


Figure 1.2. In-network optimization vs. centralized optimization.

Hessian information similarly imposes substantial communication burdens on the network (as well as more costly intermediate optimization steps), which is infeasible. These methods suffer since they are "statistically oblivious", meaning that they do not exploit statsitical properties of the loss functions, such as those due to homogeneity of data. Determining *statistical-computational error trade-offs* can bring new insights on how to efficiently solve such problems. Recent developments have been made in context of centralized optimization, e.g. [10], [11], but such methods are not applicable to general decentralized optimization problems.

In this thesis, we develop novel algorithmic frameworks in both centralized and decentralized system settings to work around the issues described above; see Sec. 1.2 for outline and contributions.

1.2 Thesis outline and contributions

This section highlights the contributions of this thesis, addressing the challenges posed in Sec. 1.1.

• Hybrid deterministic/random parallel algorithms for large-scale non-convex optimization (Chapter 2)

We propose a decomposition framework for the parallel optimization of the sum of a differentiable (possibly non-convex) function and a non-smooth (possibly non-separable), convex function. The latter term is usually employed to enforce structure in the solution, typically sparsity. The main contribution of this work is a novel *parallel, hybrid random/deterministic* decomposition scheme wherein, at each iteration, a subset of (block) variables is updated at the same time by minimizing a convex surrogate of the original non-convex function. To tackle huge-scale problems, the (block) variables to be updated are chosen according to a *mixed random and deterministic* procedure, which captures the advantages of both pure deterministic and random update-based schemes. Almost sure convergence of the proposed scheme is established. Numerical results show that on huge-scale problems the proposed hybrid random/deterministic algorithm outperforms random and deterministic schemes on both convex and non-convex problems.

The novel results of this chapter are published in:

1. Amir Daneshmand, Francisco Facchinei, Vyacheslav Kungurtsev, and Gesualdo Scutari. "Hybrid random/deterministic parallel algorithms for convex and non-convex big data optimization." IEEE Transactions on Signal Processing 63, no. 15 (2015): 3914-3929.

2. Amir Daneshmand, Francisco Facchinei, Vyacheslav Kungurtsev, and Gesualdo Scutari. "Flexible selective parallel algorithms for big data optimization." In proceedings of the 48th Asilomar Conference on Signals, Systems and Computers, pp. 3-7. IEEE, 2014.

• Decentralized first-order methods for non-convex optimization and second-order guarantees (Chapter 3)

We consider distributed smooth non-convex unconstrained optimization over networks, modeled as a connected graph. We examine the behavior of distributed gradient-based algorithms near strict saddle points. Specifically, we establish that (i) the renowned Distributed Gradient Descent (DGD) algorithm likely converges to a *neighborhood* of a Second-order Stationary (SoS) solution; and (ii) the more recent class of distributed algorithms based on gradient tracking–implementable also over digraphs–likely converges to *exact* SoS solutions,

thus avoiding strict saddle-points. Furthermore, new convergence rate results to first-order critical points is established for the latter class of algorithms.

The novel results of this chapter are published in:

1. Amir Daneshmand, Gesualdo Scutari, and Vyacheslav Kungurtsev. "Second-order guarantees of distributed gradient algorithms." SIAM Journal on Optimization 30, no. 4 (2020): 3029-3068.

2. Amir Daneshmand, Gesualdo Scutari, and Vyacheslav Kungurtsev. "Second-order guarantees of gradient algorithms over networks." In proceedings of the 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton), pp. 359-365. IEEE, 2018.

• **Decentralized first-order methods for (strongly) convex optimization under statistical similarity (Chapter 4)**

We study a class of multiagent optimization problems over (directed, time-varying) graphs. We consider the minimization of $F + G$ subject to convex constraints, where F is the smooth strongly convex sum of the agent's losses and G is a non-smooth convex function. The algorithm employs the use of surrogate objective functions in the agents' subproblems (going thus beyond linearization, such as proximal-gradient) coupled with a perturbed (push-sum) consensus mechanism that aims to track locally the gradient of F. The algorithm achieves precision $\epsilon > 0$ on the objective value in $\mathcal{O}(\kappa_g \log(1/\epsilon))$ gradient computations at each node and $\tilde{\mathcal{O}}\Big(\kappa_g(1-\rho)^{-1/2}\log(1/\epsilon)\Big)$ communication steps, where κ_g is the condition number of F and ρ characterizes the connectivity of the network. This is the first linear rate result for distributed composite optimization; it also improves on existing (non-accelerated) schemes just minimizing F, whose rate depends on much larger quantities than κ_g (e.g., the worst-case condition number among the agents). When considering in particular empirical risk minimization problems with statistically similar data across the agents, our algorithm employing high-order surrogates achieves precision $\epsilon > 0$ in $\mathcal{O}\Big((\beta/\mu)\log(1/\epsilon)\Big)$ iterations and $\tilde{\mathcal{O}}\Big((\beta/\mu)(1-\rho)^{-1/2}\log(1/\epsilon)\Big)$ communication steps, where β measures the degree of similarity of the agents' losses and μ is the strong convexity constant of F. Therefore, when

$\beta/\mu < \kappa_g$, the use of high-order surrogates yields provably faster rates than what achievable by first-order models; this is without exchanging any Hessian matrix over the network.

The novel results of this chapter are published in:

1. Ying Sun, Amir Daneshmand, Gesualdo Scutari. "Distributed optimization based on gradient-tracking revisited: Enhancing convergence rate via surrogation." arXiv:1905.02637 (2019).
   Note: *Accepted under minor revision to SIAM Journal on Optimization (SIOPT).*

• **Decentralized Newton methods over networks (Chapter 5)**

We propose a distributed cubic regularization of the Newton method for solving (constrained) empirical risk minimization problems over a network of agents, modeled as undirected graph. The algorithm employs an *inexact, preconditioned* Newton step at each agent's side: the gradient of the centralized loss is iteratively estimated via a gradient-tracking consensus mechanism and the Hessian is subsampled over the local data sets. No Hessian matrices are thus exchanged over the network. We derive global complexity bounds for convex and strongly convex losses. Our analysis reveals an interesting interplay between sample and iteration/communication complexity: *statistically accurate* solutions are achievable roughly in the same number of iterations of the centralized cubic Newton, with a communication cost per iteration of the order of $\tilde{\mathcal{O}}\big(1/\sqrt{1-\rho}\big)$, where ρ characterizes the connectivity of the network. This represents a significant communication saving with respect to that of existing, statistically oblivious, distributed Newton-based methods over networks.

The novel results of this chapter are published in:

1. Amir Daneshmand, Gesualdo Scutari, Pavel Dvurechensky, and Alexander Gasnikov. "Newton Method over Networks is Fast up to the Statistical Precision." In proceedings of the 38th International Conference on Machine Learning (ICML), July 18-24, 2021.

• **Decentralized bi-convex optimization over time-varying directed networks** In this section, we briefly review other novel results that are not included in this thesis but are published in:

1. Amir Daneshmand, Ying Sun, Gesualdo Scutari, Francisco Facchinei, and Brian M. Sadler. "Decentralized dictionary learning over time-varying digraphs." Journal of Machine Learning Research 20 (2019).

2. Amir Daneshmand, Ying Sun, Gesualdo Scutari, and Francisco Facchinei. "D2L: Decentralized dictionary learning over dynamic networks." In proceedings of 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4084-4088. IEEE, 2017.

3. Amir Daneshmand, Gesualdo Scutari, and Francisco Facchinei. "Distributed dictionary learning." In proceedings of the 50th Asilomar Conference on Signals, Systems and Computers, pp. 1001-1005. IEEE, 2016.

We study a general family of Dictionary Learning problems over directed time-varying network, defined as the minimization of the sum of bi-convex functions (with *private* and *shared* variables coupling local cost functions) plus a non-smooth convex regularizer. We develop a unified decentralized algorithmic framework for this class of *non-convex* problems, which is proved to converge to stationary solutions at a sublinear rate. The new method hinges on Successive Convex Approximation techniques, coupled with a decentralized tracking mechanism aiming at locally estimating the gradient of the smooth part of the sum-utility. To the best of our knowledge, this is the first provably convergent decentralized algorithm for Dictionary Learning and, more generally, bi-convex problems over (time-varying) (di)graphs.

1.3 Notation

The set of nonnegative integers is denoted by $\mathbb{N}_+$ and we use $[n]$ as a shorthand for $\{1, 2, \ldots, n\}$. Given a vector x, $||x||$ denotes the ℓ_2 norm of x; any other specific vector norm is subscripted accordingly. x is called *stochastic* if all its components are nonnegative and sum to one; and 1 is the vector of all ones (we write 1_d for the d–dimensional vector, if the dimension is not clear from the context). Given sets $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}^d$, we denote $\mathcal{X} \setminus \mathcal{Y} \triangleq \{x \in \mathcal{X} : x \notin \mathcal{Y}\}$, $\overline{\mathcal{X}} \triangleq \mathbb{R}^d \setminus \mathcal{X}$ (complement of $\mathcal{X}$), and $x + \mathcal{X} = \{x + z : z \in \mathcal{X}\}$. $\mathcal{V}_x$ and $\mathcal{B}(x, r)^d$ denote a neighborhood of x and the d-dimensional closed ball of radius $r > 0$ centered at x,

respectively; when the ball is centered at 0, we will write $\mathcal{B}_r^d$. We further define an *annulus* by $\mathcal{S}_{r,\epsilon} \triangleq \mathcal{B}_r^d \setminus \mathcal{B}_{r-\epsilon}^d$, with some $r > \epsilon > 0$. The Euclidean projection of $x \in \mathbb{R}^d$ onto the convex closed set $\mathcal{X} \subseteq \mathbb{R}^d$ is $\text{proj}_{\mathcal{X}}(x) \triangleq \arg\min_{y \in \mathcal{X}} ||x - y||$. The sublevel set of a function U at u is denoted by $\mathcal{L}_U(u) \triangleq \{x : U(x) \leq u\}$.

Matrices are denoted by capital letters; A_{ij} is the the (i, j)-th element of A; $\mathcal{M}_d(\mathbb{R})$ is the set of all $d \times d$ real matrices; I is the identity matrix (if the dimension is not clear from the context, we write I_d for the $d \times d$ identity matrix); $A \geq 0$ denotes a nonnegative matrix; and $A \geq B$ stands for $A - B \geq 0$. The spectrum of a square real matrix M is denoted by $\text{spec}(M)$ and its spectral radius is $\text{spradii}(M) \triangleq \max\{|\lambda| : \lambda \in \text{spec}(M)\}$; the spectral norm is $||M|| \triangleq \max_{||x|| \neq 0} ||Mx||/||x||$, and any other matrix norm is subscripted accordingly. Finally, the minimum (resp. maximum) singular value are denoted by $\sigma_{\min}(M)$ (resp. $\sigma_{\max}(M)$) and minimum (resp. maximum) eigenvalue by $\lambda_{\min}(M)$ (resp. $\lambda_{\max}(M)$).

Part I

Distributed Non-convex Optimization

2. HYBRID RANDOM/DETERMINISTIC PARALLEL ALGORITHMS FOR CONVEX AND NON-CONVEX BIG-DATA OPTIMIZATION

In this chapter, we consider a general family of optimization problems, the minimization of the sum of a smooth (possibly *nonconvex*) function F and a nonsmooth (possibly *nonseparable*) convex function G:

$$\min_{x\in\mathcal{X}} V(x) \triangleq F(x) + G(x), \tag{2.1}$$

where $\mathcal{X}$ is a closed convex set with a cartesian product structure: $\mathcal{X} = \Pi_{i=1}^{B}\mathcal{X}_i \subseteq \mathbb{R}^d$. Our focus is on problems with a huge number of variables, as those that can be encountered, e.g., in machine learning, compressed sensing, data mining, tensor factorization and completion, network optimization, image processing, genomics, etc.. We refer the reader to [12]–[24] and the books [25], [26] as entry points to the literature.

Block Coordinate Descent (BCD) methods rapidly emerged as a winning paradigm to attack Big Data optimization, mainly due to their low-cost per-iteration and scalability; see e.g. [14]. At each iteration of a BCD method one block of variables is updated using first-order information, while keeping all other variables fixed. The choice of the block of variables to update at each iteration can accomplished in several ways, for example using a cyclic order or some greedy/opportunistic selection strategy, which aims at selecting the block leading to the largest decrease of the objective function. The cyclic order has the advantage of being extremely simple, but the greedy strategy usually provides faster convergence, at the cost of an increased computational effort at each iteration. However, no matter which block selection rule is adopted, as the dimensions of the optimization problems increase, even BCD methods may result inadequate. To alleviate the "curse of dimensionality", three different kind of strategies have been proposed, namely: (a) *parallelism*, where several blocks of variables are updated simultaneously in a multicore or distributed computing environment, see e.g. [16]–[21], [27]–[36]; (b) *random selection* of the block(s) of variables to update, see e.g. [31]–[41]; and (c) use of "*more-than-first-order*" information, for example (approximated) Hessians or (parts of) the original function itself, see e.g. [15], [29], [30], [42], [43]. Point (a) is self-

explanatory and rather intuitive; here we only remark that the vast majority of parallel BCD methods apply to *convex problems only*. Points (b) and (c) need further comments.
`Point (b)`: Random selection-based rules are essentially as cheap as cyclic selections while alleviating some of the pitfalls of cyclic updates. They are also relevant in distributed environments wherein data are not available in their entirety, but are acquired either in batches or over a network. In such scenarios, one might be interested in running the optimization at a certain instant even with the limited, randomly available information. The main limitation of random selection rules is that they remain disconnected from the status of the optimization process, which instead is exactly the kind of behavior that greedy-based updates try to avoid, in favor of faster convergence, but at the cost of more intensive computation.
`Point (c)`: The use of "more-than-first-order" information also has to do with the trade-off between cost-per-iteration and overall cost of the optimization process. Although using higher order or structural information may seem unreasonable in Big Data problems, recent studies, as those mentioned above, suggest that a judicious use of some kind of "more-than-first-order" information can lead to substantial improvements.

The above pros & cons analysis suggests that it would be desirable to design a parallel algorithm for nonconvex problems combining the benefits of random sketching *and* greedy updates, possibly using "more-than-first-order" information. To the best of our knowledge, no such algorithm exists in the literature. In this chapter, we propose a BCD-like scheme for the computation of stationary solutions of Problem (2.1) filling the gap and enjoying *all* the following features:

1. It uses a random selection rule for the blocks, followed by a deterministic subselection;
2. It can classically tackle separable convex function G, i.e., $G(x) = \sum_{\mathrm{i}} G_{\mathrm{i}}(x_{\mathrm{i}})$, but also *nonseparable* functions G;
3. It can deal with a nonconvex functions F;
4. It can use both first-order and higher-order information;
5. It is parallel;

6. It can use inexact updates;

7. It converges *almost surely*, i.e. our convergence results are of the form "with probability one".

The proposed scheme is the *first* algorithm enjoying all these properties, even in the convex case. Subsequent relevant parallel algorithms applicable to (2.1) appeared after this work was published among which, notably, they relax Lispchitz continuity assumption [44], synchronous updates assumption [45], and enable random data sample selection [46] and etc. The combination of all the features 1-7 in one single algorithm is a major achievement in itself, which offers great flexibility to develop tailored instances of solutions methods within the same framework (and thus all converging under the same unified conditions). Last but not least, our experiments show impressive performance of the proposed methods, outperforming state-of-the-art solution scheme (cf. Sec. 2.3). As a final remark, we underline that, at more methodological level, the combination of all features 1-7 and, in particular, the need to conciliate random and deterministic strategies, led to the development of a new type of convergence analysis (see Appendix 2.5.1) which is also of interest *per se* and could bring to further developments.

Below we further comment on some of features 1-7, compare to existing results, and detail our contributions.

`Feature 1`: As far as we are aware of, the idea of making a random selection and then perform a greedy subselection has been previously discussed only in [47]. However, results therein i) are only for *convex* problems with a *specific* structure; ii) are based on a regularized first-order model; iii) require a very stringent "spectral-radius-type" condition to guarantee convergence, which severely limits the degree of parallelism; and iv) convergence results are in terms of expected value of the objective function. The proposed algorithmic framework expands vastly on this setting, while enjoying also all properties 2-7. In particular, it is the first hybrid random/greedy scheme for *nonconvex nonseparable* functions, and it allows *any* degree of parallelism (i.e., the update of any number of variables); and all this is achieved under much weaker convergence conditions than those in [47], satisfied by most of practical problems. Numerical results show that the proposed hybrid schemes updating greedily just

some blocks within the pool of those selected by a random rule is very effective, and seems to preserve the advantages of both random and deterministic selection rules.

`Feature 2`: The ability of dealing with some classes of nonseparable convex functions has been documented in [48]–[50], *but only for deterministic and sequential schemes*; our approach extends also to *parallel*, *random* schemes.

`Feature 3`: The list of works dealing with BCD methods for nonconvex F's is short: [33], [40] for *random sequential* methods; and [18], [28]–[30] for *deterministic parallel* ones. Random parallel methods for nonconvex F's (not enjoying the key properties 1, 2, and 6) are studied, independently from this work but drawing on [29], [30], also in [51]. We observe that for certain classes of specific *additively separable* F's, dual ADMM-like schemes have been proposed for nonconvex problems shown to be convergent under strong conditions; see, e.g., [52] and references therein. However, for the scale and generality of problems we are interested in, they are computationally impractical.

`Feature 4`: We want to stress the ability of the proposed algorithm to exploit in a systematic way "more-than-first-order" information. Differently from BCD methods that use at each iteration a (possibly regularized) first-order model of the objective function, our method provides the flexibility of using more sophisticated models, including Newton-like surrogates as well as more structured functions as those described in the following example. Suppose that in (2.1) $F = F_1 + F_2$, where F_1 is convex and F_2 is not. Then, at iteration ν, one could base the update of the i-th block on the surrogate function $F_1(x_{\rm i}, x^\nu_{-\rm i}) + \nabla_{x_{\rm i}} F_2(x^\nu)^T (x_{\rm i} - x^\nu_{\rm i}) + G(x_{\rm i}, x^\nu_{-\rm i})$, where $x_{-\rm i}$ denotes the vector obtained from x by deleting $x_{\rm i}$. The rationale here is that instead of linearizing the whole function F we only linearize the difficult, nonconvex part F_2. In this light we can also better appreciate the importance of feature 6, since if we go for more complex surrogate functions, the ability to deal with inexact solutions becomes important.

`Feature 6`: Inexact solution methods have been little studied. Papers [14], [53], [54] (somewhat indirectly) consider some of these issues for ℓ_2-loss linear support vector machines problems. A more systematic treatment of inexactness of the solution of a first-order model is documented in [55], in the context of random sequential BCD methods for *convex* problems.

As a final remark, we note that a large portion of the aforementioned works focuses on (global) complexity analysis. Specifically, with the exception of [40], they all studied (regularized) *gradient-type* methods for *convex* problems. Complexity analysis is an important topic, but it is outside the scope of this thesis. Given our expanded setting, we believe it is more fruitful to concentrate on proving convergence and verifying the practical effectiveness of our algorithms.

This chapter is organized as follows. Section 2.1 formally introduces the optimization problem along with several motivating examples and also discusses some technical points. The proposed algorithmic framework and its convergence properties are introduced in Section 2.2, while numerical results are presented in Section 2.3. Section 2.4 draws some conclusions.

2.1 Problem Set-up and Preliminaries

We consider Problem (2.1), where the feasible set $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_B$ is a Cartesian product of lower dimensional convex sets $\mathcal{X}_\mathrm{i} \subseteq \mathbb{R}^{d_\mathrm{i}}$, and $x \in \mathbb{R}^d$ is partitioned accordingly: $x = (x_1, \ldots, x_B)$, with each $x_\mathrm{i} \in \mathbb{R}^{d_\mathrm{i}}$; we denote by $\mathcal{B} \triangleq \{1, \ldots, B\}$ the set of the B blocks. The function F is smooth (and not necessarily convex and separable) and G is convex, and possibly nondifferentiable and nonseparable. Problem (2.1) is very general and includes many popular Big Data formulations; some examples are listed next.

`Ex.#1-(group) LASSO:` $F(x) = \|Ax - b\|^2$ and $G(x) = c\|x\|_1$ (or $G(x) = c\sum_{\mathrm{i}=1}^{B} \|x_\mathrm{i}\|_2$, $\mathcal{X} = \mathbb{R}^d$), $\mathcal{X} = \mathbb{R}^d$, with $A \in \mathbb{R}^{m\times d}$, $b \in \mathbb{R}^m$, and $c \in \mathbb{R}_{++}$ given constants; (group) LASSO has long been used in many applications in signal processing and statistics [12].

`Ex.#2-linear regression:` $F(x) = 0$ and $G(x) = \|Ax - b\|_1$, $\mathcal{X} = \mathbb{R}^d$, with $A \in \mathbb{R}^{m\times d}$, and $b \in \mathbb{R}^m$ given constants; the ℓ_1-norm linear regression is widely used techniques in statistics [56]. Note that G is nonseparable.

`Ex.#3-The Fermat-Weber problem:` $F(x) = 0$ and $G(x) = \sum_{\mathrm{i}=1}^{I} \omega_\mathrm{i}\|A_\mathrm{i}x - b_\mathrm{i}\|_2$, $\mathcal{X} = \mathbb{R}^d$, with $A_\mathrm{i} \in \mathbb{R}^{m\times d}$, $b_\mathrm{i} \in \mathbb{R}^m$, and $\omega_\mathrm{i} > 0$ given constants, for all i; this problem, which consists in finding $x \in \mathbb{R}^d$ such that the weighted sum of distances between x and the I anchors $\omega_1, \omega_2, \ldots, \omega_I$, was widely investigated in the optimization as well as location communities; see, e.g., [57]. This is another example of nonseparable G.

`Ex.#4−The TV image reconstruction:` $F(X) = \|AX - V\|^2$ and $G(X) = c \cdot \mathrm{TV}(X)$, $\mathcal{X} = \mathbb{R}^{m\times m}$, where $A \in \mathbb{R}^{t\times m}$, $X \in \mathbb{R}^{m\times m}$, $V \in \mathbb{R}^{t\times m}$, $c \in \mathbb{R}_{++}$, and $\mathrm{TV}(X) \triangleq \sum_{i,j=1}^{m} \|\mathcal{D}_{ij}X\|_p$ is the discrete total variational semi-norm of X, with $p = 1$ or 2 and $\mathcal{D}_{ij}X$ being the discrete gradient of X defined as $\mathcal{D}_{ij}X \triangleq [(\mathcal{D}_{ij}X)^{(1)}, (\mathcal{D}_{ij}X)^{(2)}]$, with $(\mathcal{D}_{ij}X)^{(1)} = X_{i+1,j} - X_{i,j}$ if $i < m$ and $(\mathcal{D}_{ij}X)^{(1)} = 0$ otherwise, and $(\mathcal{D}_{ij}X)^{(2)} = X_{i,j+1} - X_{i,j}$ if $i < m$ and $(\mathcal{D}_{ij}X)^{(1)} = 0$ otherwise [58]. This is the well-known noise-free discrete TV model for compressing sensing image reconstruction [58]; TV minimizing models have become a successful methodology for image processing, including denoising, deconvolution, and restoration, to name a few.

`Ex.#5−Dictionary learning:` $F(X,Y) = \frac{1}{2}\|M - XY\|_F^2$ and $G(Y) = c\|Y\|_1$, $\mathcal{X} = \{(X,Y) \in \mathbb{R}^{s\times m} \times \mathbb{R}^{m\times t} : \|Xe_i\|^2 \le \alpha_i, \forall i = 1,\dots,m\}$, where X and Y are the (matrix) optimization variables, $M \in \mathbb{R}^{s\times t}$, $c > 0$, and $(\alpha_i)_{i=1}^m > 0$ are given constants, e_i is the m-dimensional vector with a 1 in the i-th coordinate and 0's elsewhere, and $\|X\|_F$ and $\|X\|_1$ denote the Frobenius norm and the ℓ_1 matrix norm of X, respectively; this is an example of the dictionary learning problem for sparse representation [59] that finds numerous applications in various fields such as computer vision and signal and image processing. Note that $F(X,Y)$ is not jointly convex in (X,Y)

`Ex.#6−Matrix completion:` $F(X,Y) = \sum_{i,j\in\Omega}(M_{ij} - (XY)_{ij})^2 + c\,(\|X\|_F^2 + \|Y\|_F^2)$, $G(X,Y) = 0$, $\mathcal{X} = \mathbb{R}^{s\times m} \times \mathbb{R}^{m\times t}$, where Ω is a given subset of $\{1,\dots,s\} \times \{1,\dots,t\}$. Matrix completion has found numerous applications in various fields such as recommender systems, computer vision, and system identification.

Other problems of interest that can be cast in the form (2.1) include the Logistic Regression, the Support Vector Machine, the Nuclear Norm Minimization, the Robust Principal Component Analysis, the Sparse Inverse Covariance Selection, and the Nonnegative Tensor Factorization; see, e.g., [60].

Assumption 2.1.1. *Given (2.1), we make the following blanket assumptions:*

2.1.1.1 Each $\mathcal{X}_i$ is nonempty, closed, and convex;

2.1.1.2 F is C^1 on an open set containing $\mathcal{X}$;

2.1.1.3 ∇F is Lipschitz continuous on $\mathcal{X}$ with constant L_F;

2.1.1.4 G is continuous and convex on $\mathcal{X}$ (possibly nondifferentiable and nonseparable);

2.1.1.5 V is coercive, i.e., $\lim_{x\in\mathcal{X},\,\|x\|\to\infty} V(x) = +\infty$.

The above assumptions are standard and are satisfied by many practical problems. For instance, 2.1.1.3 holds automatically if $\mathcal{X}$ is bounded, whereas 2.1.1.5 guarantees the existence of a solution.

With the advances of multi-core architectures, it is desirable to develop *parallel* solution methods for Problem (2.1) whereby operations can be carried out on some or (possibly) all (block) variables x_{i} at the *same* time. The most natural parallel (Jacobi-type) method one can think of is updating *all* blocks simultaneously: given x^ν, each (block) variable x_{i} is updated by solving the following subproblem

$$x_{\mathrm{i}}^{\nu+1} \in \operatorname*{argmin}_{x_{\mathrm{i}}\in\mathcal{X}_{\mathrm{i}}} \left\{ F(x_{\mathrm{i}}, x_{-\mathrm{i}}^\nu) + G(x_{\mathrm{i}}, x_{-\mathrm{i}}^\nu) \right\}. \tag{2.2}$$

Unfortunately this method converges only under very restrictive conditions [61] that are seldom verified in practice (even in the absence of the nonsmooth part G). Furthermore, the exact computation of $x_{\mathrm{i}}^{\nu+1}$ may be difficult and computationally too expensive. To cope with these issues, a natural approach is to replace the (nonconvex) function $F(\bullet, x_{-\mathrm{i}}^\nu)$ by a suitably chosen local convex surrogate $\widetilde{F}_{\mathrm{i}}(x_{\mathrm{i}}; x^\nu)$, and solve instead the convex problems (one for each block)

$$x_{\mathrm{i}}^{\nu+1} \in \operatorname*{argmin}_{x_{\mathrm{i}}\in\mathcal{X}_{\mathrm{i}}} \left\{ \tilde{h}_{\mathrm{i}}(x_{\mathrm{i}}; x^\nu) \triangleq \widetilde{F}_{\mathrm{i}}(x_{\mathrm{i}}; x^\nu) + G(x_{\mathrm{i}}; x_{-\mathrm{i}}^\nu) \right\}, \tag{2.3}$$

with the understanding that the minimization in (2.3) is simpler than that in (2.2). Note that the function G has not been touched; this is because i) it is generally much more difficult to find a "good" surrogate of a nondifferentiable function than of a differentiable one; ii) G is already convex; and iii) the functions G encountered in practice do not make the optimization problem (2.3) difficult (a closed form solution is available for a large classes of G's, if $\widetilde{F}_{\mathrm{i}}$ are properly chosen). In this work we assume that the surrogate functions $\widetilde{F}_{\mathrm{i}}(z; w) : \mathcal{X}_{\mathrm{i}} \times \mathcal{X} \to \mathbb{R}$, have the following properties:

(F1) $\widetilde{F}_{\mathrm{i}}(\bullet; w)$ is uniformly strongly convex with constant $q > 0$ on $\mathcal{X}_{\mathrm{i}}$;

(F2) $\nabla_{x_{\mathrm{i}}} \widetilde{F}_{\mathrm{i}}(x_{\mathrm{i}}; x) = \nabla_{x_{\mathrm{i}}} F(x)$ for all $x \in \mathcal{X}$;

(F3) $\nabla_{x_{\mathrm{i}}} \widetilde{F}_{\mathrm{i}}(z; \bullet)$ is Lipschitz continuous on $\mathcal{X}$ for all $z \in \mathcal{X}_{\mathrm{i}}$;

where $\nabla_{x_{\mathrm{i}}} \widetilde{F}_{\mathrm{i}}$ is the partial gradient of $\widetilde{F}_{\mathrm{i}}$ with respect to (w.r.t.) its first argument z. Function $\widetilde{F}_{\mathrm{i}}$ should be regarded as a (simple) convex surrogate of F at the point x w.r.t. the block of variables x_{i} that preserves the first order properties of F w.r.t. x_{i}. Note that, contrary to most of the works in the literature (e.g., [50]), we do not require $\widetilde{F}_{\mathrm{i}}$ to be a global *upper* surrogate of F, which significantly enlarges the range of applicability of the proposed solution methods.

The most popular choice for $\widetilde{F}_{\mathrm{i}}$ satisfying F1-F3 is

$$\widetilde{F}_{\mathrm{i}}(x_{\mathrm{i}}; x^{\nu}) = F(x^{\nu}) + \nabla_{x_{\mathrm{i}}} F(x^{\nu})^T (x_{\mathrm{i}} - x_{\mathrm{i}}^{\nu}) + \frac{\tau_{\mathrm{i}}}{2} \|x_{\mathrm{i}} - x_{\mathrm{i}}^{\nu}\|^2, \tag{2.4}$$

with $\tau_{\mathrm{i}} > 0$. This is essentially the way a new iteration is computed in most (block-)BCDs for the solution of LASSO problems and its generalizations. When $G \equiv 0$, this choice gives rise to a gradient-type scheme; in fact we obtain $x_{\mathrm{i}}^{\nu+1}$ simply by a shift along the antigradient. As we discussed in Sec. I, this is a first-order method, so it seems advisable, at least in some situations, to use more informative $\widetilde{F}_{\mathrm{i}}$-s. If $F(x_{\mathrm{i}}, x_{-\mathrm{i}}^{\nu})$ is convex, an alternative is to take $\widetilde{F}_{\mathrm{i}}(x_{\mathrm{i}}; x^{\nu})$ as a second order expansion of $F(x_{\mathrm{i}}, x_{-\mathrm{i}}^{\nu})$ around x_{i}^{ν}, i.e.,

$$\widetilde{F}_{\mathrm{i}}(x_{\mathrm{i}}; x^{\nu}) = F(x^{\nu}) + \nabla_{x_{\mathrm{i}}} F(x^{\nu})^T (x_{\mathrm{i}} - x_{\mathrm{i}}^{\nu}) + \frac{1}{2} (x_{\mathrm{i}} - x_{\mathrm{i}}^{\nu})^T \left(\nabla^2_{x_{\mathrm{i}} x_{\mathrm{i}}} F(x^{\nu}) + qI \right) (x_{\mathrm{i}} - x_{\mathrm{i}}^{\nu}), \tag{2.5}$$

where q is nonnegative and can be taken to be zero if $F(x_{\mathrm{i}}, x_{-\mathrm{i}}^{\nu})$ is actually strongly convex. When $G \equiv 0$, this essentially corresponds to taking a Newton step in minimizing the "reduced" problem $\min_{x_{\mathrm{i}} \in \mathcal{X}_{\mathrm{i}}} F(x_{\mathrm{i}}, x_{-\mathrm{i}}^{\nu})$. Still in the case of a uniformly strongly convex $F(x_{\mathrm{i}}, x_{-\mathrm{i}}^{\nu})$, one could also take just $\widetilde{F}_{\mathrm{i}}(x_{\mathrm{i}}; x^{\nu}) = F(x_{\mathrm{i}}, x_{-\mathrm{i}}^{\nu})$, which preserves the structure of the function. Other valuable choices tailored to specific applications are discussed in [30], [62]. As a guideline, note that our method, as we shall describe in details shortly, is based on the iterative (approximate) solution of problem (2.3) and therefore a balance should be

aimed at between the accuracy of the surrogate $\tilde{F}$ and the ease of solution of (2.3). Needless to say, the option (2.4) is the less informative one, but usually it makes the computation of the solution of (2.3) a cheap task.

Best-response map: Associated with each i and point $x^\nu \in \mathcal{X}$, under F1-F3, we can define the following optimal block solution map:

$$\widehat{x}_{\mathrm{i}}(x^\nu) \triangleq \underset{x_{\mathrm{i}} \in \mathcal{X}_{\mathrm{i}}}{\operatorname{argmin}} \tilde{h}_{\mathrm{i}}(x_{\mathrm{i}}; x^\nu). \tag{2.6}$$

Note that $\widehat{x}_{\mathrm{i}}(x^\nu)$ is always well-defined, since the optimization problem in (2.6) is strongly convex. Given (2.6), we can then introduce the solution map

$$\mathcal{X} \ni y \mapsto \widehat{x}(y) \triangleq (\widehat{x}_{\mathrm{i}}(y))_{\mathrm{i}=1}^{B}. \tag{2.7}$$

Our algorithmic framework is based on solving in parallel a suitable selection of sub-problems (2.6), converging thus to *fixed-points* of $\widehat{x}(\bullet)$ (of course the selection varies at each iteration). It is then natural to ask which relation exists between these fixed points and the stationary solutions of Problem (2.1). To answer this key question, we recall first two basic definitions.

Stationarity: A point x^* is a stationary point of (2.1) if a subgradient $\xi \in \partial G(x^*)$ exists such that $(\nabla F(x^*) + \xi)^T (y - x^*) \geq 0$ for all $y \in \mathcal{X}$.

Coordinate-wise stationarity: A point x^* is a coordinate-wise stationary point of (2.1) if subgradients $\xi_{\mathrm{i}} \in \partial_{\xi_{\mathrm{i}}} G(x^*)$, with $\mathrm{i} \in \mathcal{B}$, exist such that $(\nabla_{x_{\mathrm{i}}} F(x^*) + \xi_{\mathrm{i}})^T (y_{\mathrm{i}} - x_{\mathrm{i}}^*) \geq 0$, for all $y_{\mathrm{i}} \in \mathcal{X}_{\mathrm{i}}$ and $\mathrm{i} \in \mathcal{B}$.

In words, a coordinate-wise stationary solution is a point for which x^* is stationary w.r.t. every block of variables. Coordinate-wise stationarity is a weaker form of stationarity. It is the standard property of a limit point of a convergent coordinate-wise scheme (see, for example [48]–[50]).

It is clear that a stationary point is always a coordinate-wise stationary point; the converse however is not always true, unless extra conditions on G are satisfied. **Regularity:** Problem

(2.1) is *regular* at a coordinate-wise stationary point x^* if x^* is also a stationary point of the problem.

The following two simple cases imply the regularity condition,

(a) G is separable (still nonsmooth), i.e., $G(x) = \sum_i G_i(x_i)$;

(b) G is continuously differentiable around x^*.

This is evident from the fact that in the first case, $\partial_{\xi_i} G(x^*) = \partial G_i(x^*)$ and in the second case, $\partial_{\xi_i} G(x^*) = \nabla_i G(x^*) = (\partial G(x^*))_i$.

Of course these two cases are not at all inclusive of situations for which regularity holds. As an example of a nonseparable function for which regularity holds at a point at which G is not continuously differentiable, consider the function arising in logistic regression problems $F(x) = \sum_{j=1}^m \log(1 + e^{-a_{ij} y_j^T x})$, with $\mathcal{X} = \mathbb{R}^d$, and $y_j \in \mathbb{R}^d$ and $a_j \in \{-1, 1\}$ being given constants. Now, choose $G(x) = c\|x\|_2$; the resulting function is continuously differentiable, and therefore regular, at any stationary point but $x^* \neq 0$. It is easy to verify that V is also regular at $x = 0$, if $c < \log 2$.

The algorithm we present in this chapter expands upon the literature in presenting the first (deterministic or random) *parallel* coordinate-wise scheme that converges to coordinate-wise stationary points. Under the regularity condition these points are also stationary, and so among the class of parallel algorithms, the method we present enlarges the class of problems for which convergence to stationary points is achieved for Problem (2.1) to include some classes of nonseparable G. Certainly, proximal gradient-like algorithms can converge to stationary points for any nonseparable G, but such schemes are inherently incapable of parallelization, and thus are typically much slower in practice. Thus, our algorithm is a step towards, if not complete fulfillment of, the desiderata of a parallel algorithm that converges to stationary points for all classes of Problem (2.1) with arbitrary nonsmooth convex G.

The following proposition is elementary and elucidates the connections between stationarity conditions of Problem (2.1) and fixed-points of $\hat{x}(\bullet)$.

Proposition 2.1.1. *Given Problem (2.1) under 2.1.1.1-2.1.1.5 and F1-F3, the following hold:*

i) *The set of fixed-points of* $\widehat{x}(\bullet)$ *coincides with the coordinate-wise stationary points of Problem* (2.1)*;*

ii) *If, in addition, Problem* (2.1) *is regular at a fixed-point of* $\widehat{x}(\bullet)$*, then such a fixed-point is also a stationary point of the problem.*

Other properties of the best-response map $\widehat{x}(\bullet)$ that are instrumental to prove convergence of the proposed algorithm are introduced in Appendix 2.5.2.

2.2 Algorithmic Framework and convergence guarantees

We begin introducing a formal description of the salient characteristic of the proposed algorithmic framework−the novel hybrid random/greedy block selection rule.

The random block selection works as follows: at each iteration k, a random set $\mathcal{S}^\nu \subseteq \mathcal{B}$ is generated, and the blocks $\mathrm{i} \in \mathcal{S}^\nu$ are the potential candidate variables to update in parallel. The set $\mathcal{S}^\nu$ is a realization of a random set-valued mapping $\boldsymbol{\mathcal{S}}^\nu$ with values in the power set of $\mathcal{B}$. To keep the proposed scheme as general as possible, we do not constraint $\boldsymbol{\mathcal{S}}^\nu$ to any specific distribution; we only require that, at each iteration k, each block i has a positive probability (possibly nonuniform) to be selected. Thus we append the following additional assumption to Assumption 2.1.1:

Assumption 2.1.1. *Given* (2.1)*, we make the following additional assumption:*

2.1.1.6 The sets $\mathcal{S}^\nu$ *are realizations of independent random set-valued mappings* $\boldsymbol{\mathcal{S}}^\nu$ *such that* $\mathbb{P}(\mathrm{i} \in \boldsymbol{\mathcal{S}}^\nu) \geq p$*, for all* $\mathrm{i} = 1, \dots, B$ *and* $\nu \in \mathbb{N}_+$*, and some* $p > 0$*.*

A random selection rule $\boldsymbol{\mathcal{S}}^\nu$ satisfying 2.1.1.6 will be called *proper sampling*. Several proper sampling rules will be discussed in details shortly.

The proposed hybrid random/greedy block selection rule consists in combining random and greedy updates in the following form. First, a random selection is performed−the set $\mathcal{S}^\nu$ is generated. Second, a greedy procedure is run to select *in the pool* $\mathcal{S}^\nu$ only the subset of blocks, say $\hat{\mathcal{S}}^\nu$, that are "promising" (according to a prescribed criterion). Finally all the blocks in $\hat{\mathcal{S}}^\nu$ are updated in parallel. The notion of "promising" block is made formal next.

Since $x_{\rm i}^\nu$ is an optimal solution of (2.6) if and only if $\widehat{x}_{\rm i}(x^\nu) = x_{\rm i}^\nu$, a natural distance of $x_{\rm i}^\nu$ from the optimality is $d_{\rm i}^\nu \triangleq \|\widehat{x}_{\rm i}(x^\nu) - x_{\rm i}^\nu\|$. The blocks in $\mathcal{S}^\nu$ to be updated can be then chosen based on such an optimality measure (e.g., opting for blocks exhibiting larger $d_{\rm i}^\nu$'s). Note that in some applications, including some of those discussed in Sec. II, given a proper block decomposition, $\widehat{x}_{\rm i}(x^\nu)$ can be computed easily in closed form, see Sec. IV for three different examples. However, this is not always the case, and on some problems, the computation of $\widehat{x}_{\rm i}(x^\nu)$ might be too expensive. In these cases it might be useful to introduce alternative, less expensive metrics by replacing the distance $\|\widehat{x}_{\rm i}(x^\nu) - x_{\rm i}^\nu\|$ with a computationally cheaper *error bound*, i.e., a function $E_{\rm i}(x)$ such that

$$\underline{s}_{\rm i}\|\widehat{x}_{\rm i}(x^\nu) - x_{\rm i}^\nu\| \leq E_{\rm i}(x^\nu) \leq \bar{s}_{\rm i}\|\widehat{x}_{\rm i}(x^\nu) - x_{\rm i}^\nu\|, \tag{2.8}$$

for some $0 < \underline{s}_{\rm i} \leq \bar{s}_{\rm i}$. We refer the interested reader to [30] for some more details, and to [63] as an entry point to the vast literature on error bounds. As an example, if problem (2.1) is unconstrained, $G(x) \equiv 0$, and we are using the surrogate function given by (2.4), a suitable error bound is the function $E_{\rm i}(x) = \|\nabla_{x_{\rm i}} F(x^\nu) + \tau_{\rm i}(x_{\rm i} - x_{\rm i}^\nu)\|$ with $\underline{s}_{\rm i} = \frac{\tau_{\rm i}}{2}$ and $\bar{s}_{\rm i} = L_F$.

The proposed hybrid random/greedy scheme capturing all the features 1)-6) discussed in Sec. I is formally given in Algorithm 1. Note that in step S.3 inexact calculations of $\widehat{x}_{\rm i}$ are allowed, which is another noticeable and useful feature: one can reduce the cost per iteration without affecting too much, experience shows, the empirical convergence speed. In step S.5 we introduced a memory in the variable updates: the new point $x^{\nu+1}$ is a convex combination via γ^ν of x^ν and $\widehat{z}^\nu$.

The convergence properties of Algorithm 1 are given next.

Theorem 2.2.1. *Let $\{x^\nu\}$ be the sequence generated by Algorithm 1, under 2.1.1.1-2.1.1.6. Suppose that $\{\gamma^\nu\}$ and $\{\varepsilon_{\rm i}^\nu\}$ satisfy the following conditions: i) $\gamma^\nu \in (0, 1]$; ii) $\gamma^\nu \to 0$; iii) $\sum_\nu \gamma^\nu = +\infty$; iv) $\sum_\nu (\gamma^\nu)^2 < +\infty$; and v) $\varepsilon_{\rm i}^\nu \leq \gamma^\nu \alpha_1 \min\{\alpha_2, 1/\|\nabla_{x_{\rm i}} F(x^\nu)\|\}$ for all ${\rm i} \in \mathcal{B}$ and some nonnegative constants α_1 and α_2. Additionally, if inexact solutions are used in Step 3, i.e., $\varepsilon_{\rm i}^\nu > 0$ for some ${\rm i}$ and infinite ν, then assume also that G is globally Lipschitz on $\mathcal{X}$. Then, either Algorithm 1 converges in a finite number of iterations to a fixed-point*

Algorithm 1: Hybrid Random/Deterministic Flexible Parallel Algorithm (HyFLEXA)

Data : $\{\varepsilon_{\mathrm{i}}^{\nu}\}$ for $\mathrm{i} \in \mathcal{B}$, $\{\gamma^{\nu}\} > 0$, $x^0 \in \mathcal{X}$, $\rho \in (0,1]$.
Iterate: ν=1, 2, ...

[S.1]: If x^{ν} satisfies a termination criterion: STOP;

[S.2]: Randomly generate a set of blocks $\mathcal{S}^{\nu} \subseteq \{1, \ldots, B\}$

[S.3]: Set $M^{\nu} \triangleq \max_{\mathrm{i} \in \mathcal{S}^{\nu}} \{E_{\mathrm{i}}(x^{\nu})\}$. Choose a subset $\hat{\mathcal{S}}^{\nu} \subseteq \mathcal{S}^{\nu}$ that contains at least one index i for which $E_{\mathrm{i}}(x^{\nu}) \geq \rho M^{\nu}$.

[S.4]: For all $\mathrm{i} \in \hat{\mathcal{S}}^{\nu}$, solve (2.6) with accuracy $\varepsilon_{\mathrm{i}}^{\nu}$:

find $z_{\mathrm{i}}^{\nu} \in \mathcal{X}_{\mathrm{i}}$ s.t. $\|z_{\mathrm{i}}^{\nu} - \widehat{x}_{\mathrm{i}}(x^{\nu})\| \leq \varepsilon_{\mathrm{i}}^{\nu}$;

Set $\widehat{z}_{\mathrm{i}}^{\nu} = z_{\mathrm{i}}^{\nu}$ for $\mathrm{i} \in \hat{\mathcal{S}}^{\nu}$ and $\widehat{z}_{\mathrm{i}}^{\nu} = x_{\mathrm{i}}^{\nu}$ for $\mathrm{i} \notin \hat{\mathcal{S}}^{\nu}$

[S.5]: Set $x^{\nu+1} \triangleq x^{\nu} + \gamma^{\nu}(\widehat{z}^{\nu} - x^{\nu})$;

[S.6]: $\nu \leftarrow \nu + 1$, and go to (S.1).

of $\hat{x}(\bullet)$ of (2.1) or there exists at least one limit point of $\{x^{\nu}\}$ that is a fixed-point of $\hat{x}(\bullet)$ w.p.1.

Proof. See Appendix 2.5.3. □

Remark 2.2.1. Note that the conditions on $\{\epsilon_{\mathrm{i}}^{\nu}\}$ imply that $\epsilon_{\mathrm{i}}^{\nu} \to 0$ for all i. The Theorem provides minimal conditions under which convergence can be guaranteed. Practically, of course the choice of $\epsilon_{\mathrm{i}}^{\nu}$ will affect the practical performance of the algorithm and the appropriate choice is problem dependent and given by practical experience.

The convergence results in Theorem 2.2.1 can be strengthened when G is separable.

Theorem 2.2.2. *In the setting of Theorem 2.2.1, suppose in addition that $G(x)$ is separable, i.e., $G(x) = \sum_{\mathrm{i} \in \mathcal{B}} G_{\mathrm{i}}(x_{\mathrm{i}})$. Then, either Algorithm 1 converges in a finite number of iterations to a stationary solution of Problem (2.1) or* every *limit point of $\{x^{\nu}\}$ is a stationary solution of Problem (2.1) w.p.1.*

Proof. See Appendix 2.5.4. □

On the random choice of $\mathcal{S}^\nu$. We discuss next some proper sampling rules $\boldsymbol{\mathcal{S}}^\nu$ that can be used in Step 3 of the algorithm to generate the random sets $\mathcal{S}^\nu$; for notational simplicity the iteration index ν will be omitted. The sampling rule $\boldsymbol{\mathcal{S}}$ is uniquely characterized by the probability mass function

$$\mathbb{P}(\mathcal{S}) \triangleq \mathbb{P}\left(\boldsymbol{\mathcal{S}} = \mathcal{S}\right), \ \mathcal{S} \subseteq \mathcal{B},$$

which assign probabilities to the subsets $\mathcal{S}$ of $\mathcal{B}$. Associated with $\boldsymbol{\mathcal{S}}$, define the probabilities $q_{\mathrm{j}} \triangleq \mathbb{P}(|\boldsymbol{\mathcal{S}}| = \mathrm{j})$, for $\mathrm{j} = 1, \ldots, B$. The following proper sampling rules, proposed in [36] for convex problems with separable G, are instances of rules satisfying 2.1.1.6, and are used in our computational experiments.

– *Uniform (U) sampling.* All blocks get selected with the same (non zero) probability:

$$\mathbb{P}(\mathrm{i} \in \boldsymbol{\mathcal{S}}) = \mathbb{P}(\mathrm{j} \in \boldsymbol{\mathcal{S}}) = \frac{\mathbb{E}\left[|\boldsymbol{\mathcal{S}}|\right]}{B}, \quad \forall \mathrm{i} \neq \mathrm{j} \in \mathcal{B}.$$

– *Doubly Uniform (DU) sampling.* All sets $\mathcal{S}$ of equal cardinality are generated with equal probability, i.e., $\mathbb{P}(\mathcal{S}) = \mathbb{P}(\mathcal{S})$, for all $\mathcal{S}, \mathcal{S} \subseteq \mathcal{B}$ such that $|\mathcal{S}| = |\mathcal{S}|$. The density function is then

$$\mathbb{P}(\mathcal{S}) = \frac{q_{|\mathcal{S}|}}{\left(\begin{array}{c} d \\ |\mathcal{S}| \end{array}\right)}.$$

– *Nonoverlapping Uniform (NU) sampling.* It is a uniform sampling assigning positive probabilities only to sets forming a partition of $\mathcal{B}$. Let $\mathcal{S}^1, \ldots, \mathcal{S}^P$ be a partition of $\mathcal{B}$, with each $|\mathcal{S}^{\mathrm{i}}| > 0$, the density function of the NU sampling is:

$$\mathbb{P}(\mathcal{S}) = \begin{cases} \dfrac{1}{P}, & \text{if } \mathcal{S} \in \left\{\mathcal{S}^1, \ldots, \mathcal{S}^P\right\} \\ 0 & \text{otherwise} \end{cases}$$

which corresponds to $\mathbb{P}(\mathrm{i} \in \mathcal{S}) = B/P$, for all $\mathrm{i} \in \mathcal{B}$.

A special case of the DU sampling that we found very effective in our experiments is the so called "nice sampling".

– *Nice Sampling (NS).* Given an integer $0 \leq \tau \leq B$, a τ-nice sampling is a DU sampling with $q_\tau = 1$ (i.e., each subset of τ blocks is chosen with the same probability).

The NS allows us to control the degree of parallelism of the algorithm by tuning the cardinality τ of the random sets generated at each iteration, which makes this rule particularly appealing in a multi-core environment. Indeed, one can set τ equal to the number of available cores/processors, and assign each block coming out from the greedy selection (if implemented) to a dedicated processor/core.

As a final remark, note that the DU/NU rules contain as special cases fully parallel and sequential updates, wherein at each iteration a *single* block is updated uniformly at random, or *all* blocks are updated.

– *Sequential sampling*: It is a DU sampling with $q_1 = 1$, or a NU sampling with $P = B$ and $\mathcal{S}^{\mathrm{j}} = \mathrm{j}$, for $\mathrm{j} = 1, \ldots, P$.

– *Fully parallel sampling*: It is a DU sampling with $q_B = 1$, or a NU sampling with $P = 1$ and $\mathcal{S}^1 = \mathcal{B}$.

Other interesting uniform and nonuniform practical rules (still satisfying 2.1.1.6) can be found in [36], [64].

On the choice of the step-size γ^ν. An example of step-size rule satisfying Theorem 2.2.1i)-iv) is: given $0 < \gamma^0 \leq 1$, let

$$\gamma^\nu = \gamma^{\nu-1}\left(1 - \theta\,\gamma^{\nu-1}\right), \quad \nu = 1, 2, \ldots, \tag{2.9}$$

where $\theta \in (0, 1)$ is a given constant. Numerical results in Section 2.3 show the effectiveness of (2.9) on specific problems. We remark that it is possible to prove convergence of Algorithm 1 also using other step-size rules, including a standard Armijo-like line-search procedure or a (suitably small) constant step-size. Note that differently from most of the schemes in the literature, the tuning of the step-size does not require the knowledge of the problem parameters (e.g., the Lipschitz constants of ∇F and G).

2.3 Numerical Results

In this section we present some preliminary experiments providing a solid evidence of the viability of our approach; they clearly show that our framework leads to practical methods that exploit well parallelism and compare favorably to existing schemes, both deterministic and random.

Because of space limitation, we present results only for (synthetic) LASSO problems, one of the most studied instances of (the convex version of) Problem (2.1), corresponding to $F(x) = \|Ax - v\|^2$, $G(x) = c\|x\|_1$, and $\mathcal{X} = \mathbb{R}^d$. Extensive experiments on more varied (nonconvex) classes of Problem (2.1) are the subject of a separate work.

All codes have been written in C++ and use the Message Passing Interface for parallel operations. All algebra is performed by using the Intel Math Kernel Library (MKL). The algorithms were tested on the General Compute Cluster of the Center for Computational Research at the SUNY Buffalo. In particular for our experiments we used a partition composed of 372 DELL 32x2.13GHz Intel E7-4830 Xeon Processor nodes with 512 GB of DDR4 main memory and QDR InfiniBand 40Gb/s network card.

Tuning of Algorithm 1: The most successful class of random and deterministic methods for LASSO problem are (proximal) gradient-like schemes, based on a linearization of F. As a major departure from current schemes, here we propose to better exploit the structure of F and use in Algorithm 1 the following best-response: given a scalar partition of the variables (i.e., $d_{\mathrm{i}} = 1$ for all i), let

$$\widehat{x}_{\mathrm{i}}(x^k) \triangleq \operatorname*{argmin}_{x_{\mathrm{i}} \in \mathbb{R}} \left\{ F(x_{\mathrm{i}}, x^k_{-\mathrm{i}}) + \frac{\tau_{\mathrm{i}}}{2}(x_{\mathrm{i}} - x^k_{\mathrm{i}})^2 + \lambda |x_{\mathrm{i}}| \right\}. \tag{2.10}$$

Note that $\widehat{x}_{\mathrm{i}}(x^k)$ has a closed form expression (using a soft-thresholding operator [19]).

The free parameters of Algorithm 1 are chosen as follows. The proximal gains τ_{i} and the step-size γ are tuned as in [30, Sec. VI.A]. The error bound function is chosen as $E_{\mathrm{i}}(x^k) = \|\widehat{x}_{\mathrm{i}}(x^k) - x^k_{\mathrm{i}}\|$, and, for any realization $\mathcal{S}^k$, the subsets $\hat{\mathcal{S}}^k$ in S.3 of the algorithm are chosen as

$$\hat{\mathcal{S}}^k = \{\mathrm{i} \in \mathcal{S}^k : E_{\mathrm{i}}(x^k) \geq \sigma M^k\}. \tag{2.11}$$

We denote by $\mathsf{c}_{\mathcal{S}^k}$ the cardinality of $\mathcal{S}^k$ normalized to the overall number of variables (in our experiments, all sets $\mathcal{S}^k$ have the same cardinality, i.e., $\mathsf{c}_{\mathcal{S}^k} = \mathsf{c}_{\mathcal{S}}$, for all k). We considered the following options for σ and $\mathsf{c}_{\mathcal{S}}$: i) $\mathsf{c}_{\mathcal{S}} = 0.01, 0.1, 0.2, 0.5, 0.8$; ii) $\sigma = 0$, which leads to a *fully parallel* pure random scheme wherein at each iteration *all* variables in $\hat{\mathcal{S}}^k$ are updated; and iii) different positive values of σ ranging from 0.01 to 0.5, which corresponds to updating in a greedy manner only a subset of the variables in $\hat{\mathcal{S}}^k$ (the smaller the σ the larger the number of potential variables to be updated at each iteration). We termed Algorithm 1 with $\sigma = 0$ "Random FLEXible parallel Algorithm" (RFLEXA), whereas the other instances with $\sigma > 0$ as "Hybrid FLEXA" (HyFLEXA).

Algorithms in the literature: We compared our versions of (Hy)FLEXA with the most representative *parallel* random and deterministic algorithms proposed in the literature to solve the *convex* instance of Problem (1) (and thus also LASSO). More specifically, we consider the following schemes.

• **PCDM** & **PCDM2**: These are (proximal) gradient-like parallel randomized BCD methods proposed in [36] for convex optimization problems. Since the authors recommend to use PCDM instead of PCDM2 for LASSO problems, we do so (indeed, our experiments show that PCDM outperforms PCDM2). We simulated PCDM under different sampling rules and we set the parameters β and ω as in [36, Table 4], which guarantees convergence of the algorithm in *expected value*.

• **Hydra** & **Hydra**^{2}: Hydra is a parallel and distributed random gradient-like CDM, proposed in [65], wherein different cores in parallel update a randomly chosen subset of variables from those they own; a closed form solution of the scalar updates is available. Hydra^2 [31] is the accelerated version of Hydra; indeed, in all our experiments, it outperformed Hydra; therefore, we will report the results only for Hydra^2. The free parameter β is set to $\beta = 2\beta_1^*$ (cf. Eq. (15) in [65]), with σ given by Eq. (12) in [65] (according to the authors, this seems one of the best choices for β).

• **FLEXA**: This is the parallel deterministic scheme we proposed in [29], [30]. We use FLEXA as a benchmark of deterministic algorithms, since it has been shown in [29], [30] that it outperforms current (parallel) first-order (accelerated) gradient-like schemes, including FISTA [19], SparRSA [20], GRock [21], parallel BCD [18], and parallel ADMM. The free

parameters of FLEXA, $\tau_{\rm i}$ and γ, are tuned as in [30, Sec. VI.A], whereas the set $\mathcal{S}^k$ is chosen as in (2.11).

• **Other algorithms**: We tested also other random algorithms, including *sequential* random BCD-like methods and Shotgun [27]. However, since they were not competitive, to not overcrowd the figures, we do not report results for these algorithms.

In all the experiments, the data matrix $A = [A_1 \cdots A_P]$ of the LASSO problem is stored in a column-block manner, uniformly across the P parallel processes. Thus the computation of each product Ax (required to evaluate ∇F) and the norm $\|x\|_1$ (that is G) is divided into the parallel jobs of computing $A_{\rm i}x_{\rm i}$ and $\|x_{\rm i}\|_1$, followed by a reduce operation. Also, for all the algorithms, the initial point was set to the zero vector.

Numerical Tests: We generated synthetic LASSO problems using the random generation technique proposed by Nesterov [17], which we properly modified following [36] to generate instances of the problem with different levels of sparsity of the solution as well as density of the data matrix $A \in \mathbb{R}^{m\times d}$; we introduce the following two control parameters: $\mathtt{s}_A$ = average % of nonzeros in each column of A (out of m); and $\mathtt{s}_{\rm sol}$ = % of nonzeros in the solution (out of d). We tested the algorithms on two groups of LASSO problems, $A \in \mathbb{R}^{10^4\times 10^5}$ and $A \in \mathbb{R}^{10^5\times 10^6}$, and several degrees of density of A and sparsity of the solution, namely $\mathtt{s}_{\rm sol} = 0.1\%, 1\%, 5\%, 15\%, 30\%$, and $\mathtt{s}_A = 10\%, 30\%, 50\%, 70\%, 90\%$. Because of the space limitation, we report next only the most representative results. Results for the LASSO instance with 100,000 variables are reported in Fig. 2.1 and 2.2. Fig. 2.1 shows the behavior of HyFLEXA as a function of the design parameters σ and $\mathtt{c}_{\mathcal{S}}$, for different values of the solution sparsity ($\mathtt{s}_{\rm sol}$), whereas in Fig. 2.2 we compare the proposed RFLEXA and HyFLEXA with FLEXA, PCDM, and Hydra^2, for different values of $\mathtt{s}_{\rm sol}$ and $\mathtt{s}_A$ (ranging from "low" dense matrices and "high" sparse solutions to "high" dense matrices and "low" sparse solutions). Finally, in Fig. 2.3 we consider larger problems with 1M variables. In all the figures, we plot the relative error $\mathtt{re}(x) \triangleq (V(x) - V^*)/V^*$ versus the CPU time, where V^* is the optimal value of the objective function V (in our experiments V^* is known). All the curves are averaged over ten independent random realizations. Note that the CPU time includes communication times and the initial time needed by the methods to perform all pre-iterations computations (this explains why the curves associated with Hydra^2 start after

the others; in fact Hydra[2] requires some nontrivial computations to estimates β). Given Fig. 2.1-2.3, the following comments are in order.

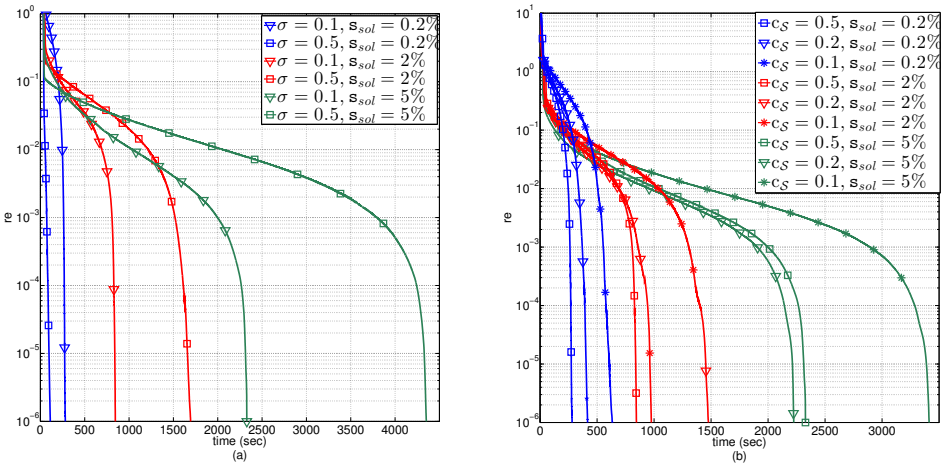


Figure 2.1. HyFLEXA for different values of $\mathtt{c}_{\mathcal{S}}$ and σ: Relative error vs. time; $\mathtt{s}_{\rm sol} = 0.2\%, 2\%, 5\%$, $\mathtt{s}_A = 70\%$, 100.000 variables, NU sampling, 8 cores; (a) $\mathtt{c}_{\mathcal{S}} = 0.5$, and $\sigma = 0.1, 0.5$ - (b) $\sigma = 0.5$, and $\mathtt{c}_{\mathcal{S}} = 0.1, 0.2, 0.5$.

HyFLEXA: On the choice of $(\mathtt{c}_{\mathcal{S}}, \sigma)$*, and the sampling strategy.* All the experiments (including those that we cannot report here because of lack of space) show the following trend in the behavior of HyFLEXA as a function of $(\mathtt{c}_{\mathcal{S}}, \sigma)$. For "low" density problems ("low" $\mathtt{s}_{\rm sol}$ and $\mathtt{s}_A$), "large" pairs $(\mathtt{c}_{\mathcal{S}}, \sigma)$ are preferable, which corresponds to updating at each iteration only *some* variables by performing a (heavy) greedy search over a *sizable* amount of variables. This is in agreement with [30] (cf. Remark 5): by the greedy selection, Algorithm 1 is able to identify those variables that will be zero at the a solution; therefore updating only variables that we have "strong" reason to believe will not be zero at a solution is a better strategy than updating them all, especially if the solutions are very sparse. Note that this behavior can be obtained using either "large" or "small" $(\mathtt{c}_{\mathcal{S}}, \sigma)$. However, in the case of "low" dense problems, the former strategy outperforms the latter. We observed that this is mainly due to the fact that when $\mathtt{s}_A$ is "small", estimating $\hat{x}_{\rm i}$ (computing the products $A^T A$) is computationally affordable, and thus performing a greedy search over more variables enhances the practical convergence. When the sparsity of the solution decreases and/or the density of A increases ("large" $\mathtt{s}_A$ and/or $\mathtt{s}_{\rm sol}$), one can see from the figures that "smaller" values

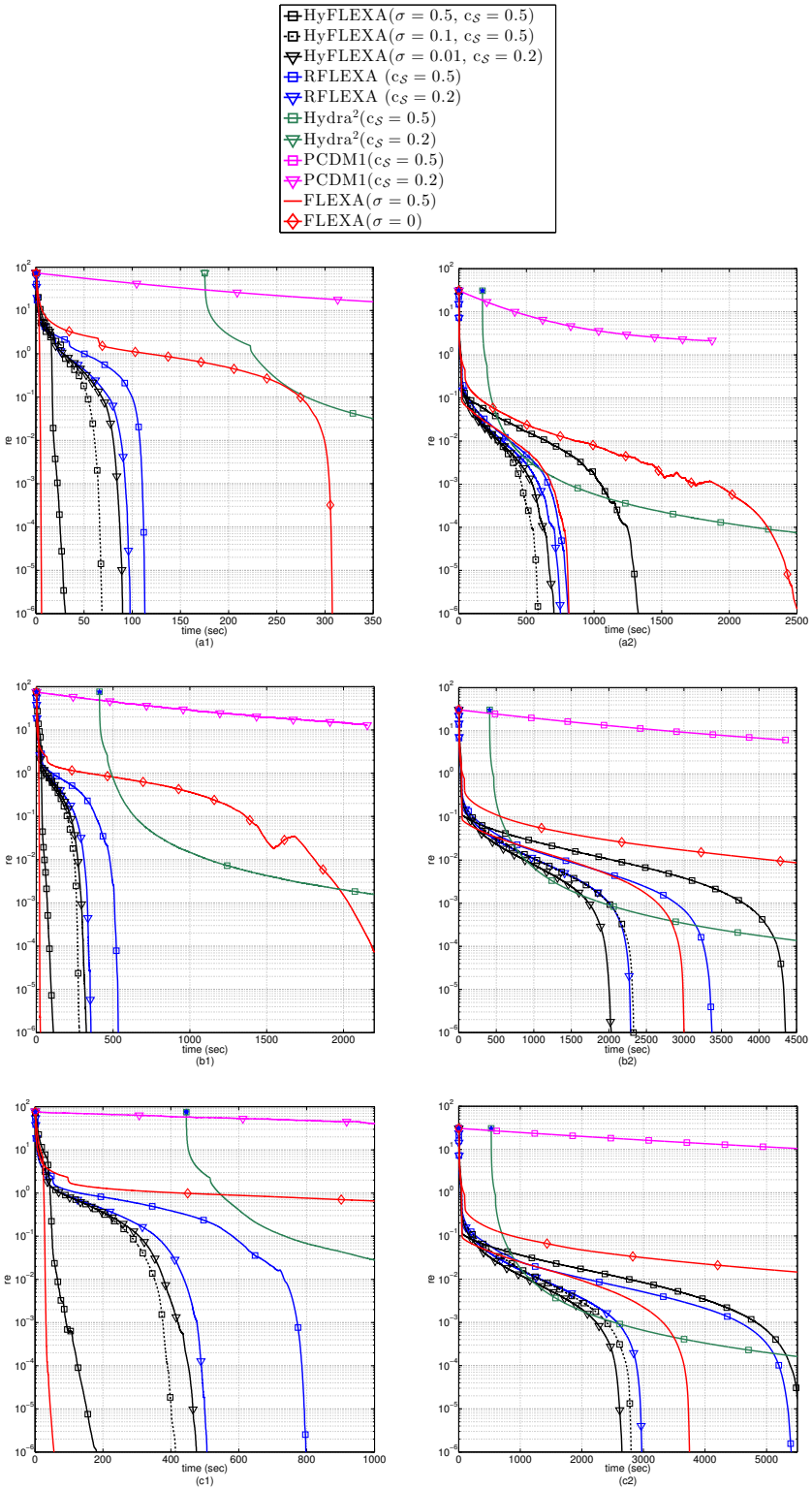


Figure 2.2. LASSO with 100.000 variables, 8 cores; Relative error vs. time for: (a1) $\mathbf{s}_A = 30\%$ and $\mathbf{s}_{\text{sol}} = 0.2\%$ - (a2) $\mathbf{s}_A = 30\%$ and $\mathbf{s}_{\text{sol}} = 5\%$ - (b1) $\mathbf{s}_A = 70\%$ and $\mathbf{s}_{\text{sol}} = 0.2\%$ - (b2) $\mathbf{s}_A = 70\%$ and $\mathbf{s}_{\text{sol}} = 5\%$ - (c1) $\mathbf{s}_A = 90\%$ and $\mathbf{s}_{\text{sol}} = 0.2\%$ - (c2) $\mathbf{s}_A = 90\%$ and $\mathbf{s}_{\text{sol}} = 5\%$.

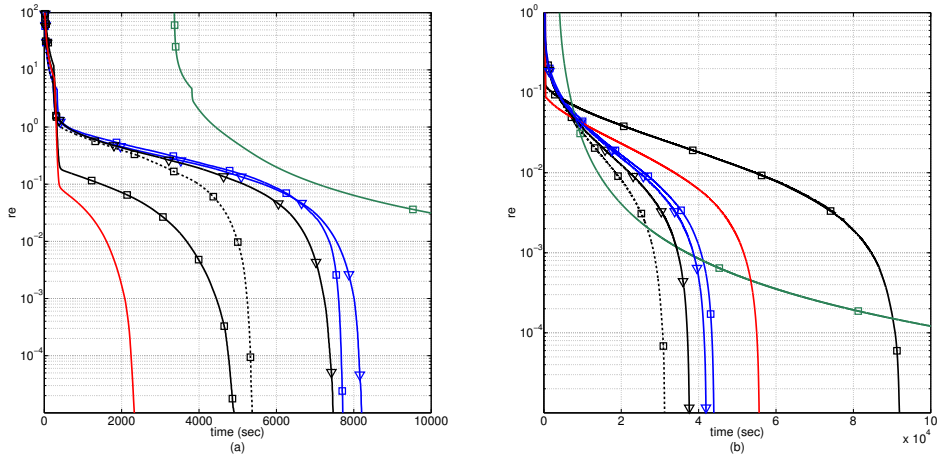


Figure 2.3. LASSO with 1M variables, $\mathtt{s}_A = 10\%$, 16 cores; Relative error vs. time for: (a)$\mathtt{s}_{\text{sol}} = 1\%$ - (b) $\mathtt{s}_{\text{sol}} = 5\%$. The legend is as in Fig. 2.2.

of $(\mathtt{c}_{\mathcal{S}}, \sigma)$ are more effective than larger ones, which corresponds to using a "less aggressive" greedy selection while searching over a smaller pool of variables. In fact, when A is dense, computing all $\hat{x}_{\text{i}}$ might be prohibitive and thus nullify the potential benefits of a greedy procedure. For instance, it follows from Fig. 2.1-2.3 that, as the density of the solution ($\mathtt{s}_{\text{sol}}$) increases the preferable choice for $(\mathtt{c}_{\mathcal{S}}, \sigma)$ progressively moves from $(0.5, 0.5)$ to $(0.2, 0.01)$, with both $\mathtt{c}_{\mathcal{S}}$ and σ decreasing. Interesting, a tuning that works quite well in practice for all the classes of problems we simulated (different densities of A, solution sparsity, number of cores, etc.) is $(\mathtt{c}_{\mathcal{S}}, \sigma) = (0.5, 0.1)$, which seems to strike a good balance between not updating variables that are probably zero at the optimum and nevertheless update a sizable amount of variables when needed in order to enhance convergence..

As a final remark, we report that, according to our experiments, the most effective sampling rule among U, DU, NU, and NS is the NU (which is actually the one the figures refers to); NS becomes competitive only when the solutions are very sparse.

Comparison of the algorithms. For low dense matrices A and very sparse solutions, FLEXA $\sigma = 0.5$ is faster than its random counterparts (RFLEXA and HyFLEXA) as well as its fully parallel version, FLEXA $\sigma = 0$ [see Fig. 2.2 a1), b1) c1) and Fig. 2.3a)]. Nevertheless, HyFLEXA [with $(\mathtt{c}_{\mathcal{S}}, \sigma) = (0.5, 0.5)$] remains close. As already pointed out, this is mainly due to the fact that in these scenarios i) estimating *all* $\hat{x}_{\text{i}}$ is computationally cheap (and

thus performing a greedy selection over a sizable set of variable is beneficial, see Fig. 2.1); and ii) updating only some variables at each iteration is more effective than updating all (FLEXA $\sigma = 0.5$ outperforms FLEXA $\sigma = 0$). However, as the density of A and/or the size of the problem increase, computing all the products $[A^T A]_{\mathrm{ii}}$ (required to estimate $\hat{x}_{\mathrm{i}}$) becomes too costly; this is when a random selection of the variables becomes beneficial: indeed, RFLEXA and HyFLEXA consistently outperform FLEXA [see Fig 2.2 a2), b2) c2) and Fig. 2.3b)]. Among the random algorithms, Hydra^2 is capable to approach relatively fast low accuracy, especially when the solution is not too sparse, but has difficulties in reaching high accuracy. RFLEXA and HyFLEXA are always much faster than current state-of-the-art schemes (PCDM and Hydra^2), especially if high accuracy of the solutions is required. Between RFLEXA and HyFLEXA (with the same $\mathsf{c}_{\mathcal{S}}$), the latter consistently outperforms the former (about up to five time faster), with a gap that is more significant when solutions are sparse. This provides a solid evidence of the effectiveness of the proposed hybrid random/greedy selection method.

In conclusion, our experiments indicate that the proposed framework leads to very efficient and practical solution methods for large and very large-scale (LASSO) problems, with the flexibility to adapt to many different problem characteristics.

2.4 Conclusions

We proposed a highly parallelizable hybrid random/deterministic decomposition algorithm for the minimization of the sum of a possibly noncovex differentiable function F and a possibily nonsmooth nonseparable convex function G. The proposed framework is the first scheme enjoying all the following features: i) it allows for pure greedy, pure random, or mixed random/greedy updates of the variables, all converging under the same unified set of convergence conditions; ii) it can tackle via parallel updates also nonseparable convex functions G; iii) it can deal with nonconvex nonseparable F; iv) it is parallel; v) it can incorporate both first-order or higher-order information; and vi) it can use inexact solutions. Our preliminary experiments on LASSO and few selected nonconvex ones showed a very promising behavior with respect to state-of-the-art random and deterministic algorithms. Of course, a more

complete assesment, especially in the nonconvex case, require much more experiments and is the subject of current research.

2.5 Appendix: Proof of Theorem 2.2.1 and 2.2.2

We first introduce some preliminary results instrumental to prove both Theorem 2.2.1 and Theorem 2.2.2. Given $\hat{\mathcal{S}}^\nu \subseteq \mathcal{B}$ and $x \triangleq (x_\mathrm{i})_{\mathrm{i}\in\mathcal{B}}$, for notational simplicity, we will denote by $(x)^\nu_{\hat{\mathcal{S}}^\nu}$ (or interchangeably $x^\nu_{\hat{\mathcal{S}}^\nu}$) the vector whose component i is equal to x_i if $\mathrm{i} \in \hat{\mathcal{S}}^\nu$, and zero otherwise. With a slight abuse of notation we will also use $(x_\mathrm{i}, y_{-\mathrm{i}})$ to denote the ordered tuple $(y_1, \ldots, y_{\mathrm{i}-1}, x_\mathrm{i}, y_{\mathrm{i}+1}, \ldots, y_B)$; similarly $(x_\mathrm{i}, x_\mathrm{j}, y_{-(\mathrm{i},\mathrm{j})})$, with $\mathrm{i} < \mathrm{j}$ stands for $(y_1, \ldots, y_{\mathrm{i}-1}, x_\mathrm{i}, y_{\mathrm{i}+1}, \ldots, y_{\mathrm{j}-1}, x_\mathrm{j}, y_{\mathrm{j}+1}, \ldots, y_B)$.

2.5.1 On the random sampling and its properties

We introduce some properties associated with the random sampling rules $\boldsymbol{\mathcal{S}}^\nu$ satisfying assumption 2.1.1.6. A key role in our proofs is played by the following random set: let $\{x^\nu\}$ be the sequence generated by Algorithm 1, and

$$\mathrm{i}^\nu_{\mathrm{mx}} = \operatorname*{argmax}_{\mathrm{i}\in\{1,\ldots,B\}} ||\widehat{x}_\mathrm{i}(x^\nu) - x^\nu_\mathrm{i}||, \tag{2.12}$$

define the set $\mathcal{K}_{\mathrm{mx}}$ as

$$\mathcal{K}_{\mathrm{mx}} \triangleq \{\nu \in \mathbb{N}_+ \,:\, \mathrm{i}^\nu_{\mathrm{mx}} \in \boldsymbol{\mathcal{S}}^\nu\}\,. \tag{2.13}$$

The key properties of this set are summarized in the following two lemmata.

Lemma 2.5.1 (Infinite cardinality)**.** *Given the set $\mathcal{K}_{mx}$ as in (2.13), it holds that*

$$\mathbb{P}\left(|\mathcal{K}_{\mathrm{mx}}| = \infty\right) = 1,$$

where $|\mathcal{K}_{\mathrm{mx}}|$ denotes the cardinality of $\mathcal{K}_{\mathrm{mx}}$.

Proof. Suppose that the statement of the lemma is not true. Then, with positive probability, there must exist some $\bar{\nu}$ such that for $\nu \geq \bar{\nu}$, $\mathrm{i}_{\mathrm{mx}}^{\nu} \notin \boldsymbol{\mathcal{S}}^{\nu}$. But we can write

$$\begin{aligned}\mathbb{P}\left(\{\mathrm{i}_{\mathrm{mx}}^{\nu} \notin \boldsymbol{\mathcal{S}}^{\nu}\}_{\nu \geq \bar{\nu}}\right) &= \prod_{\nu \geq \bar{\nu}} \mathbb{P}\left(\mathrm{i}_{\mathrm{mx}}^{\nu} \notin \boldsymbol{\mathcal{S}}^{\nu} \mid (\mathrm{i}_{\mathrm{mx}}^{\bar{\nu}} \notin \boldsymbol{\mathcal{S}}^{\bar{\nu}}),\ \ldots,\ (\mathrm{i}_{\mathrm{mx}}^{\nu-1} \notin \boldsymbol{\mathcal{S}}^{\nu-1})\right) \\ &\leq \lim_{\nu \to \infty} (1-p)^{\nu-\bar{\nu}} = 0.\end{aligned}$$

where the inequality follows by 2.1.1.6 and the independence of the events. But this obviously gives a contradiction and concludes the proof. □

Lemma 2.5.2. *Let $\{\gamma^{\nu}\}$ be a sequence satisfying assumptions i)-iii) of Theorem 2.2.1. Then it holds that*

$$\mathbb{P}\left(\sum_{\nu \in \mathcal{K}_{\mathrm{mx}}} \gamma^{\nu} < \infty\right) = 0. \tag{2.14}$$

Proof. It holds that,

$$\mathbb{P}\left(\sum_{\nu \in \mathcal{K}_{\mathrm{mx}}} \gamma^{\nu} < \infty\right) \ \leq\ \mathbb{P}\left(\bigcup_{n \in \mathbb{N}} \sum_{\nu \in \mathcal{K}_{\mathrm{mx}}} \gamma^{\nu} < n\right) \leq \sum_{n \in \mathbb{N}} \mathbb{P}\left(\sum_{\nu \in \mathcal{K}_{\mathrm{mx}}} \gamma^{\nu} < n\right).$$

To prove the lemma, it is then sufficient to show that $\mathbb{P}\left(\sum_{\nu \in \mathcal{K}_{\mathrm{mx}}} \gamma^{\nu} < n\right) = 0$, as proved next.

Define $\hat{K}_{\mathrm{i}}$, with $\mathrm{i} \in \mathbb{N}_+$, as the smallest index $\hat{K}_{\mathrm{i}}$ such that

$$\sum_{\mathrm{j}=0}^{\hat{K}_{\mathrm{i}}} \gamma^{\mathrm{j}} \geq \mathrm{i} \cdot n. \tag{2.15}$$

Note that since $\sum_{\nu=0}^{\infty} \gamma^{\nu} = +\infty$, $\hat{K}_{\mathrm{i}}$ is well-defined for all i and $\lim_{\mathrm{i} \to \infty} \hat{K}_{\mathrm{i}} = +\infty$. Hence,

$$\begin{aligned}&\mathbb{P}\left(\sum_{\nu \in \mathcal{K}_{\mathrm{mx}}} \gamma^{\nu} < n\right) = \mathbb{P}\left(\bigcap_{m \in \mathbb{N}} \left(\sum_{\nu \in \mathcal{K}_{\mathrm{mx}}}^{m} \gamma^{\nu} < n\right)\right) \qquad = \lim_{m \to \infty} \mathbb{P}\left(\sum_{\nu \in \mathcal{K}_{\mathrm{mx}}}^{m} \gamma^{\nu} < n\right) = \lim_{\mathrm{i} \to \infty} \mathbb{P}\left(\sum_{\nu \in \mathcal{K}_{\mathrm{mx}}}^{\hat{K}_{\mathrm{i}}} \gamma^{\nu} < n\right) \\ &= \lim_{\mathrm{i} \to \infty} \left[\mathbb{P}\left(\sum_{\nu \in \mathcal{K}_{\mathrm{mx}}}^{\hat{K}_{\mathrm{i}}} \gamma^{\nu} < n,\ \ |\mathcal{K}_{\mathrm{mx}} \cap [0, \hat{K}_{\mathrm{i}}]| < \frac{\hat{K}_{\mathrm{i}}}{\sqrt{\mathrm{i}}}\right) + \mathbb{P}\left(\sum_{\nu \in \mathcal{K}_{\mathrm{mx}}}^{\hat{K}_{\mathrm{i}}} \gamma^{\nu} < n,\ \ |\mathcal{K}_{\mathrm{mx}} \cap [0, \hat{K}_{\mathrm{i}}]| \geq \frac{\hat{K}_{\mathrm{i}}}{\sqrt{\mathrm{i}}}\right)\right] \\ &\leq \lim_{\mathrm{i} \to \infty} \Bigg[\underbrace{\mathbb{P}\left(|\mathcal{K}_{\mathrm{mx}} \cap [0, \hat{K}_{\mathrm{i}}]| < \frac{\hat{K}_{\mathrm{i}}}{\sqrt{\mathrm{i}}}\right)}_{\text{term I}} + \underbrace{\mathbb{P}\left(\sum_{\nu \in \mathcal{K}_{\mathrm{mx}}}^{\hat{K}_{\mathrm{i}}} \gamma^{\nu} < n,\ \ |\mathcal{K}_{\mathrm{mx}} \cap [0, \hat{K}_{\mathrm{i}}]| \geq \frac{\hat{K}_{\mathrm{i}}}{\sqrt{\mathrm{i}}}\right)}_{\text{term II}}\Bigg].\end{aligned} \tag{2.16}$$

holds for any $n \in \mathbb{N}$. Let us bound next "term I" and "term II" separately.

Term I: We have

$$\begin{aligned}
&\mathbb{P}\left(|\mathcal{K}_{\mathrm{mx}} \cap [0, \hat{K}_{\mathrm{i}}]| < \frac{\hat{K}_{\mathrm{i}}}{\sqrt{\mathrm{i}}}\right) \overset{(a)}{=} \mathbb{P}\left(\sum_{\nu=0}^{\hat{K}_{\mathrm{i}}} X_\nu < \frac{\hat{K}_{\mathrm{i}}}{\sqrt{\mathrm{i}}}\right) \leq \mathbb{P}\left(\left|\sum_{\nu=0}^{\hat{K}_{\mathrm{i}}} X_\nu - \sum_{\nu=0}^{\hat{K}_{\mathrm{i}}} p_\nu\right| > \sum_{\nu=0}^{\hat{K}_{\mathrm{i}}} p_\nu - \frac{\hat{K}_{\mathrm{i}}}{\sqrt{\mathrm{i}}}\right) \\
&\overset{(b)}{\leq} \left(\frac{\sqrt{\sum_{\nu=0}^{\hat{K}_{\mathrm{i}}} p_\nu(1-p_\nu)}}{\sum_{\nu=0}^{\hat{K}_{\mathrm{i}}} p_\nu - \frac{\hat{K}_{\mathrm{i}}}{\sqrt{\mathrm{i}}}}\right)^2 \overset{(c)}{\leq} \left(\frac{\sqrt{\hat{K}_{\mathrm{i}}}}{\hat{K}_{\mathrm{i}}\left(p - \frac{1}{\sqrt{\mathrm{i}}}\right)}\right)^2 = \left(\frac{1}{\sqrt{\hat{K}_{\mathrm{i}}}\left(p - \frac{1}{\sqrt{\mathrm{i}}}\right)}\right)^2 \underset{\mathrm{i}\to\infty}{\longrightarrow} 0
\end{aligned} \tag{2.17}$$

where:

(a): $X_0, \ldots, X_{\hat{K}_{\mathrm{i}}}$ are independent Bernoulli random variables, with parameter $p_\nu \triangleq \mathbb{P}(\nu \in \mathcal{K}_{\mathrm{mx}})$. Note that, due to 2.1.1.6, $p_\nu \geq p$, for all ν;

(b): it follows from Chebyshev's inequality;

(c): we used the bounds $\sum_{\nu=0}^{\hat{K}_{\mathrm{i}}} p_\nu(1-p_\nu) \leq \hat{K}_{\mathrm{i}}$ and $\sum_{\nu=0}^{\hat{K}_{\mathrm{i}}} p_\nu \geq p\hat{K}_{\mathrm{i}}$.

Term II: Let us rewrite term II as

$$\begin{aligned}
&\mathbb{P}\left(\frac{\sum_{\nu\in\mathcal{K}_{\mathrm{mx}}}^{\hat{K}_{\mathrm{i}}} \gamma^\nu}{|\mathcal{K}_{\mathrm{mx}} \cap [0, \hat{K}_{\mathrm{i}}]|} < \frac{n}{|\mathcal{K}_{\mathrm{mx}} \cap [0, \hat{K}_{\mathrm{i}}]|} \,\middle|\, |\mathcal{K}_{\mathrm{mx}} \cap [0, \hat{K}_{\mathrm{i}}]| \geq \frac{\hat{K}_{\mathrm{i}}}{\sqrt{\mathrm{i}}}\right) \cdot \mathbb{P}\left(|\mathcal{K}_{\mathrm{mx}} \cap [0, \hat{K}_{\mathrm{i}}]| \geq \frac{\hat{K}_{\mathrm{i}}}{\sqrt{\mathrm{i}}}\right) \\
&\overset{(a)}{\leq} \mathbb{P}\left(\frac{\sum_{\nu\in\mathcal{K}_{\mathrm{mx}}}^{\hat{K}_{\mathrm{i}}} \gamma^\nu}{|\mathcal{K}_{\mathrm{mx}} \cap [0, \hat{K}_{\mathrm{i}}]|} < \frac{n\sqrt{\mathrm{i}}}{\hat{K}_{\mathrm{i}}} \,\middle|\, |\mathcal{K}_{\mathrm{mx}} \cap [0, \hat{K}_{\mathrm{i}}]| \geq \frac{\hat{K}_{\mathrm{i}}}{\sqrt{\mathrm{i}}}\right) \cdot \mathbb{P}\left(|\mathcal{K}_{\mathrm{mx}} \cap [0, \hat{K}_{\mathrm{i}}]| \geq \frac{\hat{K}_{\mathrm{i}}}{\sqrt{\mathrm{i}}}\right) \\
&\overset{(b)}{\leq} \mathbb{P}\left(\frac{\sum_{\nu\in\mathcal{K}_{\mathrm{mx}}}^{\hat{K}_{\mathrm{i}}} \gamma^\nu}{|\mathcal{K}_{\mathrm{mx}} \cap [0, \hat{K}_{\mathrm{i}}]|} < \frac{\sum_{\nu=0}^{\hat{K}_{\mathrm{i}}} \gamma^\nu}{\hat{K}_{\mathrm{i}}\sqrt{\mathrm{i}}}\right) \overset{(c)}{\leq} \mathbb{P}\left(\frac{\sum_{\nu=0}^{\hat{K}_{\mathrm{i}}} \gamma^\nu X_\nu}{\hat{K}_{\mathrm{i}}} < \frac{\sum_{\nu=0}^{\hat{K}_{\mathrm{i}}} \gamma^\nu}{\hat{K}_{\mathrm{i}}} \frac{1}{\sqrt{\mathrm{i}}}\right) \\
&\leq \mathbb{P}\left(\left|\frac{\sum_{\nu=0}^{\hat{K}_{\mathrm{i}}} \gamma^\nu X_\nu}{\hat{K}} - \frac{\sum_{\nu=0}^{\hat{K}_{\mathrm{i}}} \gamma^\nu p_\nu}{\hat{K}_{\mathrm{i}}}\right| > \frac{\sum_{\nu=0}^{\hat{K}_{\mathrm{i}}} \gamma^\nu p_\nu}{\hat{K}_{\mathrm{i}}} - \frac{\sum_{\nu=0}^{\hat{K}_{\mathrm{i}}} \gamma^\nu}{\hat{K}_{\mathrm{i}}} \frac{1}{\sqrt{\mathrm{i}}}\right) \\
&\leq \mathbb{P}\left(\left|\frac{\sum_{\nu=0}^{\hat{K}_{\mathrm{i}}} \gamma^\nu X_\nu}{\hat{K}_{\mathrm{i}}} - \frac{\sum_{\nu=0}^{\hat{K}_{\mathrm{i}}} \gamma^\nu p_\nu}{\hat{K}_{\mathrm{i}}}\right| > \left(p - \frac{1}{\sqrt{\mathrm{i}}}\right) \frac{\sum_{\nu=0}^{\hat{K}_{\mathrm{i}}} \gamma^\nu}{\hat{K}_{\mathrm{i}}}\right) \overset{(d)}{\leq} \left(\frac{\sqrt{\sum_{\nu=0}^{\hat{K}_{\mathrm{i}}} (\gamma^\nu)^2\, p\,(1-p)}}{\left(p - \frac{1}{\sqrt{\mathrm{i}}}\right) \sum_{\nu=0}^{\hat{K}_{\mathrm{i}}} \gamma^\nu}\right)^2 \\
&\leq \left(\frac{\sqrt{\sum_{\nu=0}^{\hat{K}_{\mathrm{i}}} \gamma^\nu}}{\left(p - \frac{1}{\sqrt{\mathrm{i}}}\right) \sum_{\nu=0}^{\hat{K}_{\mathrm{i}}} \gamma^\nu}\right)^2 = \left(\frac{1}{\left(p - \frac{1}{\sqrt{\mathrm{i}}}\right) \sqrt{\sum_{\nu=0}^{\hat{K}_{\mathrm{i}}} \gamma^\nu}}\right)^2 \underset{\mathrm{i}\to\infty}{\longrightarrow} 0,
\end{aligned} \tag{2.18}$$

where:

(a): we used $|\mathcal{K}_{\mathrm{mx}} \cap [0, \hat{K}_{\mathrm{i}}]| \geq \frac{\hat{K}_{\mathrm{i}}}{\sqrt{\mathrm{i}}}$, by the conditioning event;

(b): it follows from (2.15), and $\mathbb{P}(A \bigcap B) \leq \mathbb{P}(A)$;

(c): $X_0, \ldots, X_{\hat{K}_\mathrm{i}}$ are independent Bernoulli random variables, with parameter p_ν. The bound is due to $|\mathcal{K}_\mathrm{mx} \cap [0, \hat{K}_\mathrm{i}]| \leq \hat{K}_\mathrm{i}$;

(d): it follows from the Chebyshev's inequality.

The desired result (2.14) follows readily combining (2.16), (2.17), and (2.18). □

2.5.2 On the best-response map $\widehat{x}(\bullet)$ and its properties

We introduce now some key properties of the mapping $\widehat{x}(\bullet)$ defined in (2.6). We also derive some bounds involving $\widehat{x}(\bullet)$ along with the sequence $\{x^\nu\}$ generated by Algorithm 1.

Lemma 2.5.3 ([30]). *Consider Problem (2.1) under 2.1.1.1-2.1.1.5, and F1-F3. Suppose that $G(x)$ is separable, i.e., $G(x) = \sum_\mathrm{i} G_\mathrm{i}(x_\mathrm{i})$, with each G_i convex on $\mathcal{X}_\mathrm{i}$. Then the mapping $\mathcal{X} \ni y \mapsto \widehat{x}(y)$ is Lipschitz continuous on $\mathcal{X}$, i.e., there exists a positive constant $\hat{L}$ such that*

$$\|\widehat{x}(y) - \widehat{x}(z)\| \leq \hat{L}\, \|y - z\|\,, \quad \forall y, z \in \mathcal{X}. \tag{2.19}$$

Lemma 2.5.4. *Let $\{x^\nu\}$ be the sequence generated by Algorithm 1. For every $\nu \in \mathcal{K}_{mx}$ and $\hat{\mathcal{S}}^\nu$ generated as in step S.3 of Algorithm 1, the following holds: there exists a positive constant c_1 such that,*

$$||\hat{x}_{\hat{\mathcal{S}}^\nu}(x^\nu) - x^\nu_{\hat{\mathcal{S}}^\nu}|| \geq c_1\, ||\hat{x}(x^\nu) - x^\nu||. \tag{2.20}$$

Proof. The following chain of inequalities holds:

$$\begin{aligned}
\left(\max_{\mathrm{i}\in\mathcal{B}} \bar{s}_\mathrm{i}\right) \left\|\hat{x}_{\hat{\mathcal{S}}^\nu}(x^\nu) - x^\nu_{\hat{\mathcal{S}}^\nu}\right\| &\overset{(a)}{\geq} \bar{s}_{\mathrm{i}^\nu_\rho} \left\|\hat{x}_{\mathrm{i}^\nu_\rho}(x^\nu) - x^\nu_{\mathrm{i}^\nu_\rho}\right\| \\
&\overset{(b)}{\geq} E_{\mathrm{i}^\nu_\rho}(x^\nu) \overset{(c)}{\geq} \rho\, E_{\mathrm{i}^\nu_\mathrm{mx}}(x^\nu) \\
&\overset{(d)}{\geq} \rho \left(\min_{\mathrm{i}\in\mathcal{B}} \underline{s}_\mathrm{i}\right) \left(\max_{\mathrm{i}\in\mathcal{B}} \|\hat{x}_\mathrm{i}(x^\nu) - x^\nu_\mathrm{i}\|\right) \\
&\geq \frac{\rho}{B} \left(\min_{\mathrm{i}\in\mathcal{B}} \underline{s}_\mathrm{i}\right) \|\hat{x}(x^\nu) - x^\nu\|
\end{aligned}$$

where: in (a) i^ν_ρ is any index in $\hat{\mathcal{S}}^\nu$ such that $E_{\mathrm{i}^\nu_\rho}(x^\nu) \geq \rho \max_{\mathrm{i}\in\mathcal{S}^\nu} E_\mathrm{i}(x^\nu)$. Note that by definition of $\hat{\mathcal{S}}^\nu$ (cf. step S.3 of Algorithm 1), such a index always exists; (b) is due to (2.8); (c) follows from the definition of i^ν_ρ, and $\max_{\mathrm{i}\in\mathcal{S}^\nu} E_\mathrm{i}(x^\nu) = E_{\mathrm{i}^\nu_\mathrm{mx}}(x^\nu)$, the latter due to $\mathrm{i}^\nu_\mathrm{mx} \in \mathcal{S}^\nu \supseteq \hat{\mathcal{S}}^\nu$ (recall that $\nu \in \mathcal{K}_\mathrm{mx}$); and (d) follows from (2.8). □

Lemma 2.5.5. *Let $\{x^\nu\}$ be the sequence generated by Algorithm 1. For every $\nu \in \mathbb{N}_+$, and $\hat{\mathcal{S}}^\nu$ generated as in step S.3, the following holds:*

$$(\nabla_x F(x^\nu))^T_{\hat{\mathcal{S}}^\nu} \left(\widehat{x}(x^\nu) - x^\nu\right)_{\hat{\mathcal{S}}^\nu} \leq -q \left\| \left(\widehat{x}(x^\nu) - x^\nu\right)_{\hat{\mathcal{S}}^\nu} \right\|^2 + \sum_{\mathrm{i} \in \hat{\mathcal{S}}^\nu} \left[G(x^\nu) - G(\widehat{x}_\mathrm{i}(x^\nu), x^\nu_{-\mathrm{i}}) \right]. \tag{2.21}$$

Proof. Optimality of $\widehat{x}_\mathrm{i}(x^\nu)$ for the subproblem i implies

$$\left(\nabla_{x_\mathrm{i}} \widetilde{F}_\mathrm{i}(\widehat{x}_\mathrm{i}(x^\nu); x^\nu) + \xi_\mathrm{i}(\widehat{x}_\mathrm{i}(x^\nu), x^\nu_{-\mathrm{i}}) \right)^T (y_\mathrm{i} - \widehat{x}_\mathrm{i}(x^\nu)) \geq 0,$$

for all $y_\mathrm{i} \in \mathcal{X}_\mathrm{i}$, and some $\xi_\mathrm{i}(\widehat{x}_\mathrm{i}(x^\nu), x^\nu_{-\mathrm{i}}) \in \partial_{x_\mathrm{i}} G(\widehat{x}_\mathrm{i}(x^\nu), x^\nu_{-\mathrm{i}})$. Therefore,

$$0 \geq \nabla_{x_\mathrm{i}} \widetilde{F}_\mathrm{i}(\widehat{x}_\mathrm{i}(x^\nu); x^\nu)^T \left(\widehat{x}_\mathrm{i}(x^\nu) - x^\nu_\mathrm{i}\right) + \xi_\mathrm{i}(\widehat{x}_\mathrm{i}(x^\nu), x^\nu_{-\mathrm{i}})^T \left(\widehat{x}_\mathrm{i}(x^\nu) - x^\nu_\mathrm{i}\right). \tag{2.22}$$

Let us (lower) bound next the two terms on the RHS of (2.22). The uniform strong monotonicity of $\widetilde{F}_\mathrm{i}(\bullet; x^\nu)$ (cf. F1),

$$\left(\nabla_{x_\mathrm{i}} \widetilde{F}_\mathrm{i}(\widehat{x}_\mathrm{i}(x^\nu); x^\nu) - \nabla_{x_\mathrm{i}} \widetilde{F}_\mathrm{i}(x^\nu_\mathrm{i}; x^\nu) \right)^T \left(\widehat{x}_\mathrm{i}(x^\nu) - x^\nu_\mathrm{i}\right) \geq q \, ||\widehat{x}_\mathrm{i}(x^\nu) - x^\nu_\mathrm{i}||^2, \tag{2.23}$$

along with the gradient consistency condition (cf. F2) $\nabla_{x_\mathrm{i}} \widetilde{F}_\mathrm{i}(x^\nu_\mathrm{i}; x^\nu) = \nabla_{x_\mathrm{i}} F(x^\nu)$ imply

$$\begin{aligned}
&\nabla_{x_\mathrm{i}} \widetilde{F}_\mathrm{i}(\widehat{x}_\mathrm{i}(x^\nu); x^\nu)^T \left(\widehat{x}_\mathrm{i}(x^\nu) - x^\nu_\mathrm{i}\right) \\
=& \left(\nabla_{x_\mathrm{i}} \widetilde{F}_\mathrm{i}(\widehat{x}_\mathrm{i}(x^\nu); x^\nu) - \nabla_{x_\mathrm{i}} \widetilde{F}_\mathrm{i}(x^\nu_\mathrm{i}; x^\nu) \right)^T \left(\widehat{x}_\mathrm{i}(x^\nu) - x^\nu_\mathrm{i}\right) \\
&+ \nabla_{x_\mathrm{i}} \widetilde{F}_\mathrm{i}(x^\nu_\mathrm{i}; x^\nu)^T \left(\widehat{x}_\mathrm{i}(x^\nu) - x^\nu_\mathrm{i}\right) \\
\geq& \nabla_{x_\mathrm{i}} F(x^\nu)^T \left(\widehat{x}_\mathrm{i}(x^\nu) - x^\nu_\mathrm{i}\right) + q \, ||\widehat{x}_\mathrm{i}(x^\nu) - x^\nu_\mathrm{i}||^2.
\end{aligned} \tag{2.24}$$

To bound the second term on the RHS of (2.22), let us invoke the convexity of $G(\bullet, x^\nu_{-\mathrm{i}})$:

$$G(x^\nu_\mathrm{i}, x^\nu_{-\mathrm{i}}) - G(\widehat{x}_\mathrm{i}(x^\nu), x^\nu_{-\mathrm{i}}) \geq \xi_\mathrm{i}(\widehat{x}_\mathrm{i}(x^\nu), x^\nu_{-\mathrm{i}})^T \left(x^\nu_\mathrm{i} - \widehat{x}_\mathrm{i}(x^\nu)\right),$$

which yields

$$\xi_\mathrm{i}(\widehat{x}_\mathrm{i}(x^\nu), x^\nu_{-\mathrm{i}})^T \left(\widehat{x}_\mathrm{i}(x^\nu) - x^\nu_\mathrm{i}\right) \geq G(\widehat{x}_\mathrm{i}(x^\nu), x^\nu_{-\mathrm{i}}) - G(x^\nu). \tag{2.25}$$

The desired result (2.21) is readily obtained by combining (2.22) with (2.24) and (2.25), and summing over $\mathrm{i} \in \hat{\mathcal{S}}^\nu$. □

Lemma 2.5.6. *Let $\{x^\nu\}$ be the sequence generated by Algorithm 1, and $\{\gamma^\nu\} \downarrow 0$. For every $\nu \in \mathbb{N}_+$ sufficiently large, and $\hat{\mathcal{S}}^\nu$ generated as in step S.3, the following holds:*

$$G(x^{\nu+1}) \leq G(x^\nu) + \gamma^\nu L_G \sum_{\mathrm{i} \in \hat{\mathcal{S}}^\nu} \varepsilon_\mathrm{i}^\nu + \gamma^\nu \sum_{\mathrm{i} \in \hat{\mathcal{S}}^\nu} \left[G(\widehat{x}_\mathrm{i}(x^\nu), x_{-\mathrm{i}}^\nu) - G(x^\nu) \right]. \tag{2.26}$$

Proof. Given $\nu \geq 0$ and $\hat{\mathcal{S}}^\nu$, define $\bar{x}^\nu \triangleq (\bar{x}_\mathrm{i}^\nu)_{\mathrm{i} \in \mathcal{B}}$, with

$$\bar{x}_\mathrm{i}^\nu \triangleq \begin{cases} x_\mathrm{i}^\nu + \gamma^\nu \left(\widehat{x}_\mathrm{i}(x^\nu) - x_\mathrm{i}^\nu \right), & \text{if } \mathrm{i} \in \hat{\mathcal{S}}^\nu \\ x_\mathrm{i}^\nu & \text{otherwise.} \end{cases}$$

By the convexity and Lipschitz continuity of G, it follows

$$\begin{aligned} G(x^{\nu+1}) &= G(x^\nu) + (G(x^{\nu+1}) - G(\bar{x}^\nu)) + (G(\bar{x}^\nu) - G(x^\nu)) \\ &\leq G(x^\nu) + \gamma^\nu L_G \textstyle\sum_{\mathrm{i} \in \hat{\mathcal{S}}^\nu} \varepsilon_\mathrm{i}^\nu + (G(\bar{x}^\nu) - G(x^\nu)), \end{aligned} \tag{2.27}$$

where L_G is a (global) Lipschitz constant of G. We bound next the last term on the RHS of (2.27).

Let $\bar{\gamma}^\nu = \gamma^\nu B$, for ν large enough so that $0 < \bar{\gamma}^\nu < 1$. Define $\check{x}^\nu \triangleq (\check{x}_\mathrm{i}^\nu)_{\mathrm{i} \in \mathcal{B}}$, with $\check{x}_\mathrm{i}^\nu = x_\mathrm{i}^\nu$ if $\mathrm{i} \notin \hat{\mathcal{S}}^\nu$, and

$$\check{x}_\mathrm{i}^\nu \triangleq \bar{\gamma}^\nu \, \widehat{x}_\mathrm{i}(x^\nu) + (1 - \bar{\gamma}^\nu) \, x_\mathrm{i}^\nu \tag{2.28}$$

otherwise. Using the definition of $\bar{x}^\nu$ it is not difficult to see that

$$\bar{x}^\nu = \frac{B-1}{B} \, x^\nu + \frac{1}{B} \, \check{x}^\nu. \tag{2.29}$$

Using (2.29) and invoking the convexity of G, the following recursion holds for sufficiently large ν:

$$\begin{aligned}
G(\bar{x}^\nu) =& G\left(\frac{1}{B}(\check{x}_1^\nu, x_{-1}^\nu) + \frac{1}{B}(x_1^\nu, \check{x}_{-1}^\nu) + \frac{B-2}{B}x^\nu\right) \\
=& G\left(\frac{1}{B}(\check{x}_1^\nu, x_{-1}^\nu) + \frac{B-1}{B}\left(x_1^\nu, \frac{1}{B-1}\check{x}_{-1}^\nu + \frac{B-2}{B-1}x_{-1}^\nu\right)\right) \\
\le& \frac{1}{B}G\left(\check{x}_1^\nu, x_{-1}^\nu\right) + \frac{B-1}{B}G\left(x_1^\nu, \frac{1}{B-1}\check{x}_{-1}^\nu + \frac{B-2}{B-1}x_{-1}^\nu\right) \\
=& \frac{1}{B}G\left(\check{x}_1^\nu, x_{-1}^\nu\right) + \frac{B-1}{B}G\left(\frac{1}{B-1}\left(x_1^\nu, \check{x}_{-1}^\nu\right) + \frac{B-2}{B-1}x^\nu\right) \\
=& \frac{1}{B}G\left(\check{x}_1^\nu, x_{-1}^\nu\right) + \frac{B-1}{B}G\left(\frac{1}{B-1}\left(\check{x}_2^\nu, x_{-2}^\nu\right)\right. \\
&\left. + \frac{1}{B-1}\left(x_1^\nu, x_2^\nu, \check{x}_{-(1,2)}^\nu\right) + \frac{B-3}{B-1}x^\nu\right) \\
=& \frac{1}{B}G\left(\check{x}_1^\nu, x_{-1}^\nu\right) + \frac{B-1}{B}G\left(\frac{1}{B-1}\left(\check{x}_2^\nu, x_{-2}^\nu\right)\right. \\
&\left. + \frac{B-2}{B-1}\left(x_1^\nu, x_2^\nu, \frac{1}{B-2}\check{x}_{-(1,2)}^\nu + \frac{B-3}{B-2}x_{-(1,2)}^\nu\right)\right) \\
\le& \frac{1}{B}G\left(\check{x}_1^\nu, x_{-1}^\nu\right) + \frac{1}{B}G\left(\check{x}_2^\nu, x_{-2}^\nu\right) \\
&+ \frac{B-2}{B-1}G\left(x_1^\nu, x_2^\nu, \frac{1}{B-2}\check{x}_{-(1,2)}^\nu + \frac{B-3}{B-2}x_{-(1,2)}^\nu\right) \\
\le& \quad \ldots \quad \le \frac{1}{B}\sum_{\mathrm{i}\in\mathcal{B}} G(\check{x}_{\mathrm{i}}^\nu, x_{-\mathrm{i}}^\nu).
\end{aligned} \tag{2.30}$$

Using (2.30), the last term on the RHS of (2.27) can be upper bounded for ν sufficiently large as

$$\begin{aligned}
G(\bar{x}^\nu) - G(x^\nu) \le& \frac{1}{B}\sum_{\mathrm{i}\in\mathcal{B}}\left[G(\check{x}_{\mathrm{i}}^\nu, x_{-\mathrm{i}}^\nu) - G(x^\nu)\right] \\
=& \frac{1}{B}\sum_{\mathrm{i}\in\hat{\mathcal{S}}^\nu}\left[G(\check{x}_{\mathrm{i}}^\nu, x_{-\mathrm{i}}^\nu) - G(x^\nu)\right] \\
\overset{(a)}{\le}& \frac{1}{B}\sum_{\mathrm{i}\in\hat{\mathcal{S}}^\nu}\left[\bar{\gamma}^\nu G(\widehat{x}_{\mathrm{i}}(x^\nu), x_{-\mathrm{i}}^\nu) + (1-\bar{\gamma}^\nu)G(x^\nu) - G(x^\nu)\right] \\
=& \gamma^\nu \sum_{\mathrm{i}\in\hat{\mathcal{S}}^\nu}\left[G(\widehat{x}_{\mathrm{i}}(x^\nu), x_{-\mathrm{i}}^\nu) - G(x^\nu)\right],
\end{aligned} \tag{2.31}$$

where (a) is due to the convexity of $G(\bullet, x_{-\mathrm{i}}^\nu)$ and the definition of $\check{x}_{\mathrm{i}}^\nu$ [cf. (2.28)].

The desired inequality (2.26) follows readily by combining (2.27) with (2.31). □

Lemma 2.5.7. [66, Lemma 3.4, p.121] *Let* $\{X^\nu\}$, $\{Y^\nu\}$, *and* $\{Z^\nu\}$ *be three sequences of numbers such that* $Y^\nu \geq 0$ *for all* ν. *Suppose that*

$$X^{\nu+1} \leq X^\nu - Y^\nu + Z^\nu, \quad \forall \nu = 0, 1, \ldots$$

and $\sum_{\nu=0}^{\infty} Z^\nu < \infty$. *Then either* $X^\nu \to -\infty$ *or else* $\{X^\nu\}$ *converges to a finite value and* $\sum_{\nu=0}^{\infty} Y^\nu < \infty$.

2.5.3 Proof of Theorem 2.2.1

For any given $\nu \geq 0$, the Descent Lemma [61] yields: with $\widehat{z}^\nu \triangleq (\widehat{z}_{\mathrm{i}}^\nu)_{\mathrm{i}\in\mathcal{B}}$ and $z^\nu \triangleq (z_{\mathrm{i}}^\nu)_{\mathrm{i}\in\mathcal{B}}$ defined in step S.4 of Algorithm 1,

$$F\left(x^{\nu+1}\right) \leq F\left(x^\nu\right) + \gamma^\nu \nabla_x F\left(x^\nu\right)^T \left(\widehat{z}^\nu - x^\nu\right) + \frac{\left(\gamma^\nu\right)^2 L_{\nabla F}}{2} \left\|\widehat{z}^\nu - x^\nu\right\|^2. \tag{2.32}$$

We bound next the second and third terms on the RHS of (2.32). Denoting by $\overline{\hat{\mathcal{S}}}^\nu$ the complement of $\hat{\mathcal{S}}^\nu$, we have,

$$\begin{aligned}
&\nabla_x F\left(x^\nu\right)^T \left(\widehat{z}^\nu - x^\nu\right) \\
=&\nabla_x F\left(x^\nu\right)^T \left(\widehat{z}^\nu - \widehat{x}(x^\nu) + \widehat{x}(x^\nu) - x^\nu\right) \\
\overset{(a)}{=}&\nabla_x F\left(x^\nu\right)_{\hat{\mathcal{S}}^\nu}^T \left(z^\nu - \widehat{x}(x^\nu)\right)_{\hat{\mathcal{S}}^\nu} + \nabla_x F\left(x^\nu\right)_{\overline{\hat{\mathcal{S}}}^\nu}^T \left(x^\nu - \widehat{x}(x^\nu)\right)_{\overline{\hat{\mathcal{S}}}^\nu} \\
&+ \nabla_x F\left(x^\nu\right)_{\hat{\mathcal{S}}^\nu}^T \left(\widehat{x}(x^\nu) - x^\nu\right)_{\hat{\mathcal{S}}^\nu} + \nabla_x F\left(x^\nu\right)_{\overline{\hat{\mathcal{S}}}^\nu}^T \left(\widehat{x}(x^\nu) - x^\nu\right)_{\overline{\hat{\mathcal{S}}}^\nu} \\
=&\nabla_x F\left(x^\nu\right)_{\hat{\mathcal{S}}^\nu}^T \left(z^\nu - \widehat{x}(x^\nu)\right)_{\hat{\mathcal{S}}^\nu} + \nabla_x F\left(x^\nu\right)_{\hat{\mathcal{S}}^\nu}^T \left(\widehat{x}(x^\nu) - x^\nu\right)_{\hat{\mathcal{S}}^\nu} \\
\overset{(b)}{\leq}&\sum_{\mathrm{i}\in\hat{\mathcal{S}}^\nu} \varepsilon_{\mathrm{i}}^\nu \left\|\nabla_{x_{\mathrm{i}}} F(x^\nu)\right\| + \nabla_x F\left(x^\nu\right)_{\hat{\mathcal{S}}^\nu}^T \left(\widehat{x}(x^\nu) - x^\nu\right)_{\hat{\mathcal{S}}^\nu} \\
\overset{(c)}{\leq}&\sum_{\mathrm{i}\in\hat{\mathcal{S}}^\nu} \varepsilon_{\mathrm{i}}^\nu \left\|\nabla_{x_{\mathrm{i}}} F(x^\nu)\right\| - q \left\|\left(\widehat{x}(x^\nu) - x^\nu\right)_{\hat{\mathcal{S}}^\nu}\right\|^2 + \sum_{\mathrm{i}\in\hat{\mathcal{S}}^\nu} \left[G(x^\nu) - G(\widehat{x}_{\mathrm{i}}(x^\nu), x_{-\mathrm{i}}^\nu)\right]
\end{aligned} \tag{2.33}$$

where in (a) we used the definition of $\widehat{z}^\nu$ and of the set $\hat{\mathcal{S}}^\nu$; in (b) we used $\|z_{\mathrm{i}}^\nu - \widehat{x}_{\mathrm{i}}(x^\nu)\| \leq \varepsilon_{\mathrm{i}}^\nu$; and (c) follows from (2.21) (cf. Lemma 2.5.5).

The third term on the RHS of (2.32) can be bounded as

$$\begin{aligned}\|\widehat{z}^{\nu}-x^{\nu}\|^2 &\leq 2\,\|(z^{\nu}-\hat{x}(x^{\nu}))_{\hat{\mathcal{S}}^{\nu}}\|^2+2\,\|(\hat{x}(x^{\nu})-x^{\nu})_{\hat{\mathcal{S}}^{\nu}}\|^2\\ &= 2\sum_{\mathrm{i}\in\hat{\mathcal{S}}^{\nu}}\|z_{\mathrm{i}}^{\nu}-\widehat{x}_{\mathrm{i}}(x^{\nu})\|^2+2\,\|(\hat{x}(x^{\nu})-x^{\nu})_{\hat{\mathcal{S}}^{\nu}}\|^2\\ &\leq 2\sum_{\mathrm{i}\in\hat{\mathcal{S}}^{\nu}}(\varepsilon_{\mathrm{i}}^{\nu})^2+2\,\|(\hat{x}(x^{\nu})-x^{\nu})_{\hat{\mathcal{S}}^{\nu}}\|^2,\end{aligned}\tag{2.34}$$

where the first inequality follows from the definition of z^{ν} and $\widehat{z}^{\nu}$, and in the last inequality we used $\|z_{\mathrm{i}}^{\nu}-\widehat{x}_{\mathrm{i}}(x^{\nu})\|\leq\varepsilon_{\mathrm{i}}^{\nu}$.

Now, we combine the above results to get the descent property of V along $\{x^{\nu}\}$. For sufficiently large $\nu\in\mathbb{N}_+$, it holds

$$\begin{aligned}&V(x^{\nu+1})=F(x^{\nu+1})+G(x^{\nu+1})\\ \leq&V\left(x^{\nu}\right)-\gamma^{\nu}\left(q-\gamma^{\nu}L_{\nabla F}\right)\|(\widehat{x}(x^{\nu})-x^{\nu})_{\hat{\mathcal{S}}^{\nu}}\|^2+T^{\nu},\end{aligned}\tag{2.35}$$

where the inequality follows from (2.21), (2.32), (2.33), and (2.34), and T^{ν} is given by

$$T^{\nu}\triangleq\gamma^{\nu}\sum_{\mathrm{i}\in\mathcal{B}}\varepsilon_{\mathrm{i}}^{\nu}\left(L_G+\|\nabla_{x_{\mathrm{i}}}F(x^{\nu})\|\right)+(\gamma^{\nu})^2\,L_{\nabla F}\sum_{\mathrm{i}\in\mathcal{B}}(\varepsilon_{\mathrm{i}}^{\nu})^2.$$

By assumption (iv) in Theorem 2.2.1, it is not difficult to show that $\sum_{\nu=0}^{\infty}T^{\nu}<\infty$. Since $\gamma^{\nu}\to 0$, it follows from (2.35) that there exist some positive constant β_1 and a sufficiently large ν, say $\bar{\nu}$, such that

$$V(x^{\nu+1})\leq V(x^{\nu})-\gamma^{\nu}\beta_1\,\|(\widehat{x}(x^{\nu})-x^{\nu})_{\hat{\mathcal{S}}^{\nu}}\|^2+T^{\nu},\tag{2.36}$$

for all $\nu\geq\bar{\nu}$. Invoking Lemma 2.5.7 while using $\sum_{\nu=0}^{\infty}T^{\nu}<\infty$ and the coercivity of V, we deduce from (2.36) that

$$\lim_{t\to\infty}\sum_{\nu=\bar{\nu}}^{t}\gamma^{\nu}\,\|(\widehat{x}(x^{\nu})-x^{\nu})_{\hat{\mathcal{S}}^{\nu}}\|^2<+\infty,\tag{2.37}$$

and thus also

$$\lim_{t\to\infty} \sum_{\mathcal{K}_{\rm mx}\ni \nu \geq \bar{\nu}}^{t} \gamma^{\nu} \left\| (\widehat{x}(x^{\nu}) - x^{\nu})_{\hat{\mathcal{S}}^{\nu}} \right\|^2 < +\infty. \tag{2.38}$$

Lemma 2.5.2 together with (2.38) imply

$$\liminf_{\nu\in\mathcal{K}_{\rm mx}} \left\| (\widehat{x}(x^{\nu}) - x^{\nu})_{\hat{\mathcal{S}}^{\nu}} \right\| = 0, \qquad \text{w.p. 1},$$

which by Lemma 2.5.4 implies

$$\liminf_{\nu\to\infty} \left\| \widehat{x}(x^{\nu}) - x^{\nu} \right\| = 0, \qquad \text{w.p. 1}. \tag{2.39}$$

Therefore, the limit point of the infimum sequence is a fixed point of $\widehat{x}(\cdot)$ w.p.1.

2.5.4 Proof of Theorem 2.2.2

The proof follows similar ideas as the one of Theorem 1 in our recent work [30], but with the nontrivial complication of dealing with randomness in the block selection.

Given (2.39), we show next that, under the separability assumption on G, it holds that $\lim_{\nu\to\infty} \|\widehat{x}(x^{\nu}) - x^{\nu}\| = 0$ w.p.1. For notational simplicity, let us define $\triangle\widehat{x}(x^{\nu}) \triangleq \widehat{x}(x^{\nu}) - x^{\nu}$.

Note first that for any finite but arbitrary sequence $\{\nu, \nu+1, ..., \mathrm{i}_{\nu}-1\}$, it holds that

$$\mathbb{E}\left[\sum_{\mathcal{K}_{\rm mx}\ni t=\nu}^{\mathrm{i}_{\nu}-1} \gamma^t \right] = \sum_{t=\nu}^{\mathrm{i}_{\nu}-1} \gamma^t \left[\mathbb{P}(t \in \mathcal{K}_{\rm mx}) \right] \geq p \sum_{t=\nu}^{\mathrm{i}_{\nu}-1} \gamma^t,$$

and thus

$$\mathbb{P}\left(\sum_{\mathcal{K}_{\rm mx}\ni t=\nu}^{\mathrm{i}_{\nu}-1} \gamma^t > \beta \sum_{t=\nu}^{\mathrm{i}_{\nu}-1} \gamma^t \right) > 0,$$

for all $\nu \in \mathcal{K}$ and $0 < \beta < p$. This implies that, w.p.1, there exists an infinite sequence of indexes, say $\mathcal{K}_1 \subseteq \mathcal{K}$, such that

$$\sum_{\mathcal{K}_{\rm mx}\ni t=\nu}^{\mathrm{i}_{\nu}-1} \gamma^t > \beta \sum_{t=\nu}^{\mathrm{i}_{\nu}-1} \gamma^t, \quad \forall \nu \in \mathcal{K}_1. \tag{2.40}$$

Suppose now, by contradiction, that $\limsup_{\nu\to\infty} \|\triangle\widehat{x}(x^\nu)\| > 0$ with a positive probability. Then we can find a realization such that at the same time (2.40) holds for some $\mathcal{K}_1$ and $\limsup_{\nu\to\infty} \|\triangle\widehat{x}(x^\nu)\| > 0$. In the rest of the proof we focus on this realization and get a contradiction, thus proving that $\limsup_{\nu\to\infty} \|\triangle\widehat{x}(x^\nu)\| = 0$ w.p.1.

If $\limsup_{\nu\to\infty} \|\triangle\widehat{x}(x^\nu)\| > 0$ then there exists a $\delta > 0$ such that $\|\triangle\widehat{x}(x^\nu)\| > 2\delta$ for infinitely many ν and also $\|\triangle\widehat{x}(x^\nu)\| < \delta$ for infinitely many ν. Therefore, one can always find an infinite set of indexes, say $\mathcal{K}$, having the following properties: for any $\nu \in \mathcal{K}$, there exists an integer $\mathrm{i}_\nu > \nu$ such that

$$\|\triangle\widehat{x}(x^\nu)\| < \delta, \qquad \left\|\triangle\widehat{x}(x^{\mathrm{i}_\nu})\right\| > 2\delta \tag{2.41}$$

$$\delta \le \left\|\triangle\widehat{x}(x^{\mathrm{j}})\right\| \le 2\delta \qquad \nu < \mathrm{j} < \mathrm{i}_\nu. \tag{2.42}$$

Proceeding now as in the proof of Theorem 2.2.1 in [30], we have: for $\nu \in \mathcal{K}_1$,

$$\begin{aligned}
\delta &\overset{(a)}{<} \left\|\triangle\widehat{x}(x^{\mathrm{i}_\nu})\right\| - \|\triangle\widehat{x}(x^\nu)\| \\
&\le \left\|\widehat{x}(x^{\mathrm{i}_\nu}) - \widehat{x}(x^\nu)\right\| + \left\|x^{\mathrm{i}_\nu} - x^\nu\right\| \qquad (2.43)\\
&\overset{(b)}{\le} (1+\hat{L})\left\|x^{\mathrm{i}_\nu} - x^\nu\right\| \qquad (2.44)\\
&\overset{(c)}{\le} (1+\hat{L})\sum_{t=\nu}^{\mathrm{i}_\nu-1}\gamma^t\left(\left\|\triangle\widehat{x}(x^t)_{S^t}\right\| + \left\|(z^t - \widehat{x}(x^t))_{S^t}\right\|\right)\\
&\overset{(d)}{\le} (1+\hat{L})\,(2\delta + \varepsilon^{\max})\sum_{t=\nu}^{\mathrm{i}_\nu-1}\gamma^t, \qquad (2.45)
\end{aligned}$$

where (a) follows from (2.41); (b) is due to Lemma 2.5.3; (c) comes from the triangle inequality, the updating rule of the algorithm and the definition of $\widehat{z}^\nu$; and in (d) we used (2.41), (2.42), and $\|z^t - \widehat{x}(x^t)\| \le \sum_{\mathrm{i}\in\mathcal{B}}\varepsilon_\mathrm{i}^t$, where $\varepsilon^{\max} \triangleq \max_\nu \sum_{\mathrm{i}\in\mathcal{B}}\varepsilon_\mathrm{i}^\nu < \infty$. It follows from (2.45) that

$$\liminf_{\mathcal{K}_1\ni\nu\to\infty}\sum_{t=\nu}^{\mathrm{i}_\nu-1}\gamma^t \ge \frac{\delta}{(1+\hat{L})(2\delta+\varepsilon^{\max})} > 0. \tag{2.46}$$

We show next that (2.46) is in contradiction with the convergence of $\{V(x^\nu)\}$. To do that, we preliminary prove that, for sufficiently large $\nu \in \mathcal{K}$, it must be $\|\triangle\widehat{x}(x^\nu)\| \geq \delta/2$. Proceeding as in (2.45), we have: for any given $\nu \in \mathcal{K}$,

$$\left\|\triangle\widehat{x}(x^{\nu+1})\right\| - \|\triangle\widehat{x}(x^\nu)\| \leq (1+\hat{L})\left\|x^{\nu+1}-x^\nu\right\| \leq (1+\hat{L})\gamma^\nu\left(\|\triangle\widehat{x}(x^\nu)\| + \varepsilon^{\max}\right). \quad (2.47)$$

It turns out that for sufficiently large $\nu \in \mathcal{K}_1$ so that $(1+\hat{L})\gamma^\nu < \delta/(\delta + 2\varepsilon^{\max})$, it must be

$$\|\triangle\widehat{x}(x^\nu)\| \geq \delta/2; \quad (2.48)$$

otherwise the condition $\|\triangle\widehat{x}(x^{\nu+1})\| \geq \delta$ would be violated [cf. (2.42)]. Hereafter we assume without loss of generality that (2.48) holds for all $\nu \in \mathcal{K}_1$ (in fact, one can always restrict $\{x^\nu\}_{\nu\in\mathcal{K}_1}$ to a proper subsequence).

We can show now that (2.46) is in contradiction with the convergence of $\{V(x^\nu)\}$. Using (2.36) (possibly over a subsequence), we have: for sufficiently large $\nu \in \mathcal{K}_1$,

$$\begin{aligned} V(x^{\mathrm{i}_\nu}) &\leq V(x^\nu) - \beta_1 \sum_{\mathcal{K}_{\mathrm{mx}}\ni t=\nu}^{\mathrm{i}_\nu-1} \gamma^t \left\|\left(\triangle\widehat{x}(x^t)\right)_{\hat{\mathcal{S}}^t}\right\|^2 + \sum_{\mathcal{K}_{\mathrm{mx}}\ni t=\nu}^{\mathrm{i}_\nu-1} T^t \\ &\overset{(a)}{\leq} V(x^\nu) - \beta_2 \sum_{\mathcal{K}_{\mathrm{mx}}\ni t=\nu}^{\mathrm{i}_\nu-1} \gamma^t \left\|\triangle\widehat{x}(x^t)\right\|^2 + \sum_{t=\nu}^{\mathrm{i}_\nu-1} T^t \\ &\overset{(b)}{\leq} V(x^\nu) - \beta_3 \sum_{t=\nu}^{\mathrm{i}_\nu-1} \gamma^t + \sum_{t=\nu}^{\mathrm{i}_\nu-1} T^t, \end{aligned} \quad (2.49)$$

where (a) follows from Lemma 2.5.4 and $\beta_2 = c_1\,\beta_1 > 0$; and (b) is due to (2.48) and (2.40), with $\beta_3 = \beta\,\beta_2\,(\delta^2/4)$.

Since $\{V(x^\nu)\}$ converges and $\sum_{\nu=0}^{\infty} T^\nu < \infty$, it holds that $\lim_{\mathcal{K}_1\ni\nu\to\infty}\sum_{t=\nu}^{\mathrm{i}_\nu-1}\gamma^t = 0$, contradicting (2.46). Therefore $\lim_{\nu\to\infty}\|\widehat{x}(x^\nu) - x^\nu\| = 0$ w.p.1. Since $\{x^\nu\}$ is bounded by the coercivity of V and the convergence of $\{V(x^\nu)\}$, it has at least one limit point $\bar{x} \in \mathcal{X}$. By the continuity of $\widehat{x}(\bullet)$ (cf. Lemma 2.5.3) it holds that $\widehat{x}(\bar{x}) = \bar{x}$. By Proposition 2.1.1 $\bar{x}$ is also a stationary solution of Problem (2.1). □

3. DECENTRALIZED FIRST-ORDER ALGORITHMS FOR NON-CONVEX OPTIMIZATION OVER NETWORKS AND SECOND-ORDER GUARANTEES

In this chapter we extend the network setting to ad-hoc (directed) topologies and we consider the minimization of a smooth unconstrained nonconvex function, in the following form:

$$\min_{\theta\in\mathbb{R}^d} F(\theta) \triangleq \sum_{\mathrm{i}=1}^{m} f_{\mathrm{i}}(\theta), \tag{3.1}$$

where m is the number of agents in the network; and $f_{\mathrm{i}} : \mathbb{R}^d \to \mathbb{R}$ is the cost function of agent i, assumed to be smooth and known only to agent i. Agents are connected through a communication network, modeled as a (possibly directed, strongly) connected graph. No specific topology is assumed for the graph (such as star or hierarchical structure). In this setting, agents seek to cooperatively solve Problem (3.1) by exchanging information with their immediate neighbors in the network.

Main objective: We call θ a critical point of F if $\nabla F(\theta) = 0$; a critical point θ is a *strict saddle* of F if $\nabla^2 F(\theta)$ has at least one negative eigenvalue; and it is a *Second-order Stationary* (SoS) solution if $\nabla^2 F(\theta)$ is positive semidefinite. Critical points that are not minimizers are of little interest in the nonconvex setting. It is thus desirable to consider methods for (3.1) that are not attracted to such points. When F has a favorable structure, stronger guarantees can be claimed. For instance, a wide range of salient objective functions arising from applications in machine learning and signal processing have been shown to enjoy the so-called *strict saddle* property: all the critical points of F are either strict saddles or local minimizers. Examples include principal component analysis and fourth order tensor factorization [9], low-rank matrix completion [67], and some instances of neural networks [68], just to name a few. In all these cases, converging to SoS solutions–and thus circumventing strict saddles–guarantees finding a local minimizer.

In this chapter, we study the second-order guarantees of two renowned distributed gradient-based algorithms for Problem (3.1), namely: the Distributed Gradient Descent (DGD) [69], [70] and the family of distributed algorithms based on gradient-tracking [71]–[73]. The former

is implementable on undirected graphs while the latter is suitable also for directed graphs. Convergence of these schemes applied to *convex* instances of (3.1) is well understood; however, less is known in the *nonconvex* case, let alone second-order guarantees; the relevant works are discussed next.

3.1 Literature review

Recent years have witnessed many studies proving asymptotic solution- and convergence rate-guarantees of a variety of algorithms for specific classes of nonconvex optimization problems (e.g., satisfying suitable regularity conditions); a good overview can be found in [74]. Since these analyses are heavily tailored to specific applications and it is unclear how to generalize them to a wider class of nonconvex functions, we omit further details and discuss next only results of centralized and distributed algorithms for *general* nonconvex instances of (3.1).

3.1.1 Second-order guarantees of centralized optimization algorithms

Second-order guarantees of centralized solution methods for general nonconvex optimization (3.1) have been extensively studied in the literature.

Hessian-based methods: Algorithms based on *second-order* information have long been known to converge to SoS solutions of (3.1); they rely on computing the Hessian to distinguish between first- and second-order stationary points. The classical cubic-regularization [75]–[79] and trust region (e.g. [80]–[83]) methods can provably find *approximate* SoS solutions in polynomial time (by approximate SoS we mean θ such that $||\nabla F(\theta)|| \leq \epsilon_g$ and $\lambda_{\min}(\nabla^2 F(\theta)) \geq -\epsilon_h$, for small $\epsilon_g, \epsilon_h > 0$); they however require access to the full Hessian matrix. A recent line of works [84]–[86] show that the requirement of full Hessian access can be relaxed to Hessian-vector products in each iteration, hence solving simpler sub-problems per iteration, but at the cost of requiring more iterations to reach approximate SoS solutions.
First-order methods: For general nonconvex problems, Gradient Descent (GD) is known to find a stationary point in polynomial time [87]. In [88], it was proved that randomly initialized GD with a fixed step-size converges to SoS solutions almost surely. The elegant

analysis of [88], leveraging tools from the theory of dynamical systems (e.g., the Stable Manifold Theorem), has been later extended in a number of follow-up works establishing same kind of second-order guarantees of a variety of first-order methods, including the proximal point algorithm, block coordinate descent, mirror descent [89]; the heavy-ball method and the Nesterov's accelerated method [90]; block coordinate descent and alternating minimization [91]; and a primal-dual optimization procedure for solving linear equality constrained nonconvex optimization problems [92]. These results are all asymptotic in nature and it is unclear whether polynomial convergence rates can be obtained for these methods. In [93] it was actually proven that, even with fairly natural random initialization schemes and for non-pathological functions, GD can be significantly slowed down by saddle points, taking exponential time to escape. Recent work has analyzed variations of GD that include stochastic perturbations. It has been shown that when perturbations are incorporated into GD at each step the resulting algorithm can escape strict saddle points in polynomial time [9]; the same conclusion was earlier established in [94] for stochastic gradient methods, although without escape time guarantees. It has also been shown that episodic perturbations suffice; in particular, [95] introduced an algorithm that occasionally adds a perturbation to GD, and proved that the number of iterations to escape saddle points depends only poly-logarithmically on dimension (i.e., it is nearly dimension-independent). Fruitful follow-up results show that other first-order perturbed algorithms escape from strict saddle points efficiently [96], [97].

3.1.2 Distributed algorithms for (3.1) and guarantees

Distributed algorithms for *convex* instances of (3.1) have a long history; less results are available for nonconvex objectives. Since the focus on this work is on nonconvex problems, next, we mainly comment on distributed algorithms for minimizing nonconvex objectives.

• **DGD and its variants:** DGD (and its variants) is unquestionably among the first and most studied decentralizations of the gradient descent algorithm for (3.1) [69], [70]. The instance of DGD considered in this work reads: given $x_{\mathrm{i}}^0 \in \mathbb{R}^d$, $\mathrm{i} \in [m]$,

$$x_{\mathrm{i}}^{\nu+1} = \sum_{\mathrm{j}=1}^{m} D_{\mathrm{ij}}\, x_{\mathrm{j}}^{\nu} - \alpha \nabla f_{\mathrm{i}}(x_{\mathrm{i}}^{\nu}), \quad \mathrm{i} \in [m], \tag{3.2}$$

where x_{i}^{ν} is the agent i's estimate at iteration ν of the vector variable θ; $\{D_{\mathrm{ij}}\}_{\mathrm{i,j}}$ are suitably chosen set of nonnegative weights (cf. Assumption 3.4.1), matching the graph topology (i.e., $D_{\mathrm{ij}} > 0$ if there is a link between node i and j, and $D_{\mathrm{ij}} = 0$ otherwise); and $\alpha > 0$ is the step-size. Roughly speaking, the update of each agent i in (3.2) is the linear combination of two components: i) the gradient ∇f_{i} evaluated at the agent's latest iterate (recall that agents do not have access to the entire gradient ∇F); and ii) a convex combination of the current iterates of the neighbors of agent i (including agent i itself). The latter term (a.k.a. consensus step) is instrumental to asymptotically enforcing agreement among the agents' local variables.

When each f_{i} in (3.1) is (strongly) *convex*, convergence of DGD is well understood. With a diminishing step-size, agents' iterates converge to a consensual *exact* solution; if a constant step-size is used, convergence is generally faster but only to a neighborhood of the solution, and exact consensus is not achieved. When (3.1) is *nonconvex*, the available convergence guarantees are weaker. In [98] it was shown that if a constant step-size is employed, every limit point $(x_1^{\infty}, \dots, x_m^{\infty})$ of the sequence generated by (3.2) satisfies $\sum_{\mathrm{i}=1}^{m} \nabla_{x_{\mathrm{i}}} f_{\mathrm{i}}(x_{\mathrm{i}}^{\infty}) = \mathbf{0}$; the limit points of agents' iterates are not consensual; asymptotic consensus is achieved only using a diminishing step-size. Since in general f_{i} are all different, such limit points are *not* critical points of F. Nothing is known about the connection of the critical points of $\sum_{\mathrm{i}=1}^{m} f_{\mathrm{i}}(x_{\mathrm{i}})$ and those of F, *let alone its second-order guarantees*. A first contribution of this research is to establish second-order guarantees of DGD (3.2) applied to (3.1) over undirected graphs.

Several extensions/variants of the vanilla DGD followed the seminal works [69], [70]. The projected (stochastic) DGD for nonconvex constrained instances of (3.1) was proposed in [99]; with a diminishing step-size, the algorithm converges to a stationary solution of the problem (almost surely, if noisy instances of the local gradients are used). The extension of DGD to *digraphs* was studied in [100] for convex unconstrained optimization, and later extended in [101] to nonconvex objectives. The algorithm, termed push-sum DGD, combines a local gradient step with the push-sum algorithm [102]. When a diminishing step-size is employed, push-sum DGD converges to an exact stationary solution of (3.1); and its noisy perturbed version almost surely converges to local minimizers, provided that F does not have

any saddle point [101]. To our knowledge, no other guarantees are known for DGD-like algorithms in the nonconvex setting. In particular, it is unclear whether DGD (3.2) escapes strict saddles of F.

• **Gradient tracking-based methods:** To cope with the speed-accuracy dilemma of DGD, [71], [72] proposed a new class of distributed gradient-based methods that converge to an *exact* consensual solution of nonconvex (constrained) problems while using a *fixed step-size*. The algorithmic framework, termed NEXT, introduces the idea of *gradient tracking* to correct the DGD direction and cancel the steady state error in it while using a fixed step-size: each agent updates its own local variables along a surrogate direction that tracks the gradient ∇F of the entire objective (the same idea was proposed independently in [73] for convex unconstrained smooth problems). The generalization of NEXT to digraphs–the SONATA algorithm–was proposed in [6], [103]–[105], with [104], [105] proving convergence of the agents' iterates to consensual stationary solutions of nonconvex problems at a sublinear rate. *No second-order guarantees have been established for these methods.* Extensions of the SONATA family based on different choices of the weight matrices were later introduced in [106], [107] for *convex* smooth unconstrained problems. In this chapter, we consider the following family of distributed algorithms based on gradient tracking, which encompasses the majority of the above schemes (see, e.g., [104, Sec. 5]), and refer to it as Distributed Optimization with Gradient Tracking (DOGT):

$$x_{\mathrm{i}}^{\nu+1} = \sum_{\mathrm{j}=1}^{m} R_{\mathrm{ij}} x_{\mathrm{j}}^{\nu} \; - \alpha \, y_{\mathrm{i}}^{\nu}, \tag{3.3}$$

$$y_{\mathrm{i}}^{\nu+1} = \sum_{\mathrm{j}=1}^{m} C_{\mathrm{ij}} y_{\mathrm{j}}^{\nu} + \nabla f_{\mathrm{i}}(x_{\mathrm{i}}^{\nu+1}) - \nabla f_{\mathrm{i}}(x_{\mathrm{i}}^{\nu}), \quad \text{(Gradient Tracking)} \tag{3.4}$$

where $(R_{\mathrm{ij}})_{\mathrm{i,j}}$ and $(C_{\mathrm{ij}})_{\mathrm{i,j}}$ are suitably chosen nonnegative weights compliant to the graph structure (cf. Assumption 3.5.1); and $y_{\mathrm{i}} \in \mathbb{R}^d$ is an auxiliary variable, controlled by agent i via the update (3.4), which aims at tracking locally the gradient sum $\sum_{\mathrm{i}} \nabla f_{\mathrm{i}}(x_{\mathrm{i}}^{\nu})$. Overall, the update (3.4) in conjunction with the consensus step in (3.3) is meant to "correct" the local gradient direction $-\nabla f_{\mathrm{i}}(x_{\mathrm{i}}^{\nu})$ (as instead used in the DGD algorithm) and thus nulls asymptotically the steady error $\nabla f_{\mathrm{i}}(x_{\mathrm{i}}^{\nu}) - \nabla F(x_{\mathrm{i}}^{\nu})$. This permits the use of a constant

step-size α while still achieving *exact* consensus without penalizing the convergence rate. Another important difference between DOGT and DGD in (3.2) is that the former serves as a unified platform for distributed algorithms applicable over both undirected and directed graphs. Convergence of DOGT in the form (3.3)-(3.4) when F is nonconvex remains an open problem, let alone second-order guarantees. A second contribution of this work is to fill this gap and provide a first- and second-order convergence analysis of DOGT.

• **Primal-dual distributed algorithms:** We conclude this literature review by commenting on distributed algorithms for nonconvex (3.1) using a primal-dual form [108]–[110]. Because of their primal-dual nature, all these schemes are implementable only over *undirected* graphs. In [108] a distributed approximate dual (sub)gradient algorithm, coupled with a consensus step is introduced. Assuming zero-duality gap, the algorithm is proved to asymptotically find a pair of primal-dual solutions of an auxiliary problem, which however might not be critical points of F; also, consensus is not guaranteed. No rate analysis is provided. In [109], a proximal primal-dual algorithm is proposed; the algorithm, termed Prox-PDA, employs either a constant or increasing penalty parameter (which plays the role of the step-size); a sublinear convergence rate of a suitably defined primal-dual gap is proved. A perturbed version of Prox-PDA, P-Prox-PDA, was introduced in [110], which can also deal with nonsmooth convex, additive functions in the objective of (3.1). P-Prox-PDA converges to an ϵ-critical point (and thus also to *inexact* consensus), under a proper choice of the penalty parameters that depends on ϵ. A sublinear convergence rate is also proved. No second-order guarantees have been established for the above schemes. The only primal-dual algorithms we are aware of with provable convergence to SoS solutions is the one in [92], proposed for a linearly constrained nonconvex optimization problem. When linear constraints are used to enforce consensus, the primal-dual method [92] becomes distributed and applicable to Problem (3.1), but only for *undirected* graphs (DOGT is instead implementable also over digraphs). Second-order guarantees of such a scheme are established under slightly stronger assumptions than those required for DOGT (cf. Remark 3.5.2, Sec. 3.5.3.3). Finally, notice that, since [92] substantially differs from DGD and DOGT–the former is a primal-dual scheme while the latter are primal methods–the convergence analysis put forth in [92] is not applicable to DGD and DOGT. Since DGD and DOGT in their general form encompass

two classic algorithms for distributed optimization, the open problem of their second-order properties leaves a significant gap in the literature.

3.2 Summary of the technical results

We establish for the first time second-order guarantees of DGD (3.2) and DOGT (3.3)-(3.4). The main results are summarized next.

3.2.1 DGD algorithm (3.2).

We prove that:

(i) For a sufficiently small step-size α, agents' iterates $\{x^\nu\}$ generated by (3.2) converge to an $O(\alpha)$-critical point of F for all initializations–see Lemma 3.4.1; neighborhood convergence to critical points is also established (cf.Theorem 3.4.3). This complements the convergence results in [98];

(ii) The average sequence $\{\overline{x}^\nu \triangleq (1/m)\sum_{\mathrm{i}=1}^m x_{\mathrm{i}}^\nu\}$ converges almost surely to a neighborhood of a SoS solution of (3.1), where the probability is taken over the initializations–see Theorem 3.4.4.

To prove **(ii)**, we employ a novel analysis, which represents a major technical contribution of this work. In fact, existing techniques developed to established second-order guarantees of the centralized GD are not readily applicable to DGD–roughly speaking, this is due to the fact that DGD (3.2) converges only to a neighborhood of critical points of F [fixed points of (3.2) are not critical points of F]. We elaborate next on this challenge and outline our analysis.

The elegant roadmap developed in [88], [89] to establish second-order guarantees of the centralized GD builds on the Stable Manifold theorem: roughly speaking, fixed-points of the gradient map corresponding to strict saddles of the objective function are "unstable" (more formally, the stable set[1] of strict saddles has zero measure), implying almost sure convergence

[1]↑ Given $\mathcal{X} \subseteq \mathbb{R}^d$, $g : \mathbb{R}^d \to \mathbb{R}^d$, and the fixed-point iterate $x^{\nu+1} = g(x^\nu)$, the stable set of $\mathcal{X}$ is $\{x : \lim_\nu g^\nu(x) \in \mathcal{X}\}$, i.e., the set of initial points such that $\{x^\nu\}$ converges to a member of $\mathcal{X}$.

of GD iterates to SoS points [89, Corollary 2]. It is known that the DGD iterates (3.2) can be interpreted as instances of the GD applied to the following auxiliary function [98], [111]: denoting $x \triangleq [x_1^\top, \dots x_m^\top]^\top$,

$$L_\alpha(x) \triangleq \underbrace{\sum_{i=1}^m f_i(x_i)}_{\triangleq F_c(x)} + \frac{1}{2\alpha} \sum_{i=1}^m \sum_{i=j}^m (e_{ij} - D_{ij}) x_i^\top x_j, \tag{3.5}$$

where $e_{ij} = 1$ if there is an edge in the graph between agent i and agent j; and $e_{ij} = 0$ otherwise. Using (3.5), (3.2) can be rewritten as: denoting $x^\nu \triangleq [x_1^{\nu\top}, \dots x_m^{\nu\top}]^\top$,

$$x^{\nu+1} = x^\nu - \alpha \nabla L_\alpha(x^\nu). \tag{3.6}$$

One can then apply the above argument (cf. [89, Corollary 2]) to (3.6) and readily establish the following result (see Theorem 3.4.1 for the formal statement)

Fact 1 (informal): For sufficient small $\alpha > 0$, randomly initialized DGD (3.6) [and thus (3.2)] converges almost surely to a second-order critical point of L_α.

Unfortunately, this result alone is not satisfactory, as no connection is known between the critical points of L_α and those of F (note that $L_\alpha : \mathbb{R}^{m \cdot d} \to \mathbb{R}$ whereas $F : \mathbb{R}^d \to \mathbb{R}$). To cope with this issue we prove the following two facts.

Fact 2 (informal): Every limit point $\bar{x}^\infty$ of the average sequence $\bar{x}^\nu = 1/m \sum_{i=1}^m x_i^\nu$ can be made arbitrarily close to a critical point of F by using a sufficiently small $\alpha > 0$ (Theorem 3.4.3);

Fact 3 (informal): Whenever the limit point $\bar{x}^\infty = 1/m \sum_{i=1}^m x_i^\infty$ belongs to a sufficiently small neighborhood of a strict saddle of F, $x^\infty = [x_1^{\infty\top}, \dots, x_m^{\infty\top}]^\top$ must be a strict saddle of L_α (Proposition 3.4.1 and Corollary 3.4.1).

The above three facts will then ensure that, for sufficiently small $\alpha > 0$, with almost complete certainty, $\{\bar{x}^\nu\}$ will not get trapped in a neighborhood of a strict saddle of F–as x^∞ would be a strict saddle of L_α–thus landing in a neighborhood of a SoS solution of (3.1).

Facts 2 & 3 above are proved under a regularity condition on F which recalls (albeit slightly weaker than) [112]. Roughly speaking, the gradient flow over some *annulus* must be uniformly positive correlated with any outward (from the origin) direction (cf. Assumption 3.3.3). This condition is quite mild and is satisfied by functions arising, e.g., from several machine learning applications, including distributed PCA, matrix sensing, and binary classification problems; see Sec. 3.3 for more details. Furthermore, this condition is also sufficient to prove convergence of DGD without assuming the objective function to be globally L-smooth (but just locally L-smooth, LC^1 for short), a requirement that instead is common to existing (first-order) convergence conditions of DGD. Notice that the loss functions arising from many of the aforementioned machine learning problems are not globally L-smooth.

3.2.2 DOGT algorithm (3.3)-(3.4)

For DOGT, we establish the following three results.

(i) When F is nonconvex and the graph is either undirected or directed, it is proved that every limit point of the sequence generated by DOGT is a critical point of F. Furthermore, a merit function, measuring distance of the iterates from stationarity and consensus disagreement is introduced, and proved to vanish at a sublinear rate–see Theorem 3.5.1. This extends convergence results [106], [107], established only for convex functions. To deal with nonconvexity, our analysis builds on a novel Lyapunov-like function [cf. (3.44)], which properly combines optimization error dynamics, consensus and tracking disagreements. While these three terms alone do not "sufficiently" decrease along the iterates–as local optimization and consensus/tracking steps might act as competing forces–a suitable combination of them, as captured by the Lyapunov function, does monotonically decrease.

(ii) When F satisfies the Kurdyka-Łojasiewicz (KŁ) property [113], [114] at any of its critical points, convergence of the entire sequence to a critical point of F is proved (cf. Theorem 3.5.2), and a convergence rate is provided (cf. Theorem 3.5.3). Although inspired by [115], establishing similar convergence results (but no rate analysis) for centralized first-order methods, our proof follows a different path building on the descent of the

Lyapunov function introduced in (i), which does not satisfy [115, conditions H1-H2]); see Sec. 3.5.2 for details.

(iii) The sequence of iterates generated by DOGT is shown to converge to SoS solutions of (3.1) almost surely, when initial points are randomly drawn from a suitably chosen linear subspace–see Theorem 3.5.5. This result is proved for undirected and directed networks. The proofs build on the stable manifold theorem, based upon the interpretation of DOGT dynamics as fixed-point iterates of a suitably defined map. The challenge in finding such a map is ensuring that the stable set of its undesirable fixed-points–those associated with the strict saddles of F–has measure zero in the subspace where the initialization of DOGT takes place. Note that this subspace is not full dimensional.

The rest of the chapter is organized as follows. The main assumptions on the optimization problem and network are introduced in Sec. 3.3. Sec. 3.4 studies guarantees of DGD over undirected graphs, along the following steps: i) existing convergence results are discussed in Sec. 3.4.1; ii) Sec. 3.4.2 studies convergence to a neighborhood of a critical point of F; and iii) Sec. 3.4.3 establishes second-order guarantees. DOGT algorithms are studied in Sec. 3.5 along the following steps: i) Sub-sequence convergence is proved in Sec. 3.5.1; ii) Sec. 3.5.2 establishes global convergence under the KŁ property of F; and iii) Sec. 3.5.3 derives second-order guarantees over undirected and directed graphs. Finally, Sec. 3.6 presents some numerical results.

The sequence generated by DGD (and DOGT) depends on the step-size α and the initialization x^0. When necessary, we write $\{x^\nu(\alpha, x^0)\}$ for $\{x^\nu\}$.

Throughout the chapter, we assume that all the probability measures are absolutely continuous with respect to the Lebesgue measure.

3.3 Problem & network setting

In this section, we introduce the various assumptions on the functions f_i and the graph, under which our results are derived.

Assumption 3.3.1 (On Problem 3.1)**.** *Given Problem* (3.1)*,*

(i) Each f_{i} *is* $r+1$ *times continuously differentiable for some* $r \geq 1$*, and* ∇f_{i} *is* L_{i}*-Lipschitz continuous. Denote* $L_{\max} \triangleq \max_{\mathrm{i}} \ L_{\mathrm{i}}$*;*

(ii) F *is coercive.*

For some convergence results of DGD we need the following slightly stronger condition.

Assumption 2.1'. Assumption 3.3.1-(i) is satisfied and (ii) each f_{i} is coercive.

We also make the blanket assumption that each agent i knows only its own f_{i} but not the rest of the objective function.

Note that Assumption 3.3.1, particularly the global Lipschitz gradient continuity of f_{i}, is quite standard in the literature. Motivated by some applications of interest (see examples below), we will also prove convergence of DGD under LC^1 only and the mild condition (3.8) below (cf. Assumption 3.3.3). Although strictly not necessary, coercivity in Assumptions 3.3.1 & 2.1' simplifies some of our derivations; our results can be extended under the weaker assumption that (3.1) has a solution.

Some of the convergence results of DGD and DOGT are established under the assumption that F satisfies the Kurdyka-Łojasiewicz (KŁ) inequality [113], [114].

Definition 3.3.1 (KŁ property)**.** *Given a function* $U : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$*, we set* $[a < U < b] \triangleq \{z \in \mathbb{R}^d \ : \ a < U(z) < b\}$*, and*

(a) The function U *has KŁ property at* $\acute{z} \in \operatorname{dom} \partial U$ *if there exists* $\eta \in (0, +\infty]$*, a neighborhood* $\mathcal{V}_{\acute{z}}$*, and a continuous concave function* $\phi : [0, \eta) \to \mathbb{R}_+$ *such that:*

(i) $\phi(0) = 0$*,*

(ii) ϕ *is* $\mathcal{C}^1$ *on* $(0, \eta)$*,*

(iii) for all $s \in (0, \eta)$*,* $\phi(s) > 0$*,*

(iv) for all $z \in \mathcal{V}_{\acute{z}} \cap [U(\acute{z}) < U < U(\acute{z}) + \eta]$*, the KŁ inequality holds:*

$$\phi\left(U(z) - U(\acute{z})\right) \operatorname{dist}(0, \partial U(z)) \geq 1. \tag{3.7}$$

(b) A proper lower-semicontinuous function U is called KŁ if it satisfies the KŁ inequality at every point in dom ∂U.

Many problems involve functions satisfying the KŁ inequality; real semi-algebraic functions provide a very rich class of functions satisfying the KŁ, see [116] for a thorough discussion.

Second-order guarantees of DGD are obtained under the following two extra assumptions below; Assumption 3.3.2 is quite standard and widely used in the literature to establish second-order guarantees of centralized algorithms (e.g., [9], [76]–[79], [82], [95]) as well as of distributed algorithms [92], [117], [118]. Assumption 3.3.3 is introduced for the first time in this work and is commented below.

Assumption 3.3.2. *Each $f_{\mathrm{i}} : \mathbb{R}^d \to \mathbb{R}$ is twice differentiable and $\nabla^2 f_{\mathrm{i}}$ is $L_{\nabla^2_{\mathrm{i}}}$-Lipschitz continuous. The Lipschitz constant of $\nabla^2 F$ and $\nabla^2 F_c$ are $L_{\nabla^2} = \sum_{\mathrm{i}=1}^m L_{\nabla^2_{\mathrm{i}}}$ and $L_{\nabla^2_c} = \max_{\mathrm{i}} L_{\nabla^2_{\mathrm{i}}}$, respectively, where F_c is defined in eq. (3.5).*

Assumption 3.3.3. *(i) Each f_{i} is LC^1; and (ii) there exist $0 < \epsilon < R$ and $\delta > 0$ such that*

$$\inf_{\theta \in \mathcal{S}_{R,\epsilon}} \left\langle \nabla f_{\mathrm{i}}(\theta), \theta / \left\|\theta\right\| \right\rangle \geq \delta, \quad \forall \mathrm{i} \in [m]. \tag{3.8}$$

Roughly speaking, the condition above postulates that the gradient $\nabla f_{\mathrm{i}}(\theta)$ is positively correlated with any radial direction $\theta/\left\|\theta\right\|$, for all θ in the annulus $\mathcal{S}_{R,\epsilon}$. A slightly more restrictive form of the above assumption has appeared in [112, Assumption A3]. Many functions of practical interest satisfy this assumption; some examples arising from machine learning applications are listed below.

Distributed PCA [119]: Given matrices $M_{\mathrm{i}} \in \mathbb{R}^{d\times d}$, $\mathrm{i} \in [m]$, the distributed PCA problem is to find the leading eigenvector of $\sum_{\mathrm{i}=1}^m M_{\mathrm{i}}$ by solving

$$\min_{\theta\in\mathbb{R}^d} \quad \frac{1}{4}\left\|\theta\theta^\top - \sum_{\mathrm{i}=1}^m M_{\mathrm{i}}\right\|_F^2, \tag{3.9}$$

which can be rewritten in the form (3.1);

Phase retrieval [74]: Let $\{(a_i, y_i)\}_{i=1}^m$, with $a_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$ such that $y_i = a_i^\top M^* a_i = (a_i^\top \theta^*)^2$, and $M^* = \theta^* \theta^{*\top} \in \mathbb{R}^{d\times d}$. The phase retrieval problem reads

$$\min_{\theta\in\mathbb{R}^d} \quad \frac{1}{4}\sum_{i=1}^m \left(||a_i^\top \theta||^2 - y_i\right)^2 + \frac{\lambda}{2}||\theta||^2, \tag{3.10}$$

where $\lambda > 0$ is a given parameter.

Matrix sensing [74]**:** Let $\{(A_i, y_i)\}_{i=1}^m$, with $A_i \in \mathbb{R}^{d\times d}$ and $y_i \in \mathbb{R}$ such that $y_i = \langle A_i, M^*\rangle$, and $M^* = \Theta^*\Theta^{*\top} \in \mathbb{R}^{d\times d}$, $\Theta^* \in \mathbb{R}^{d\times r}$. The matrix sensing problem reads

$$\min_{\Theta\in\mathbb{R}^{d\times r}} \quad \frac{1}{4}\sum_{i=1}^m \left(\left\langle A_i, \Theta\Theta^\top\right\rangle - y_i\right)^2 + \frac{\lambda}{2}\left\|\Theta\right\|_F^2, \tag{3.11}$$

where $\lambda > 0$ is a given parameter.

Gaussian mixture model [120]: Let $\{z_i\}_{i=1}^m$ be m points drawn from a mixture of q Gaussian distributions, i.e., $z_i \sim \sum_{j=1}^q \mathcal{N}(\mu_j^*, \Sigma)$, where $\mathcal{N}(\mu_j^*, \Sigma)$ is the Gaussian distribution with mean $\mu_j^* \in \mathbb{R}^d$ and covariance $\Sigma \in \mathbb{R}^{d\times d}$. The goal is to estimate the mean values $\mu_1^*, \ldots, \mu_q^*$ by solving the maximum likelihood problem

$$\min_{\{\theta_j\in\mathbb{R}^d\}_{j=1}^q} \quad -\sum_{i=1}^m \log\left(\sum_{j=1}^q \phi_d(z_i - \theta_j)\right) + \frac{\lambda}{2}\left\|\theta_j\right\|^2, \tag{3.12}$$

where $\phi_d(\theta)$ is the multivariate normal distribution with 0 mean and covariance Σ;

Bilinear logistic regression [121]: The description of the problem along with some numerical results can be found in Sec. 3.6.2;

Artificial neuron [122], [123]: Let $\{(s_i, \xi_i)\}_{i=1}^m$ be m samples, with $s_i \in \mathbb{R}^d$, $\xi_i \in \mathbb{R}$, and measurement model $\xi_i = \sigma(s_i^\top \theta^*)$, where θ^* is the optimal weights and $\sigma(\cdot)$ is a *transfer* function; e.g., the logistic regression function $\sigma(\theta) = 1/(1+\exp(-\theta))$. The goal is to estimate θ^* by solving

$$\min_{\theta\in\mathbb{R}^d} \quad \sum_{i=1}^m \frac{1}{2m}\left[\left(\xi_i - \sigma(s_i^\top \theta)\right)^2 + \frac{\lambda}{2}\left\|\theta\right\|^2\right], \tag{3.13}$$

where $\lambda > 0$ is a given parameter. Further binary classification models satisfying Assumption 3.3.3 include f_{i} functions such as [123]

$$\begin{aligned} f_{\mathrm{i}}(\theta) =& 1 - \tanh \xi_{\mathrm{i}} s_{\mathrm{i}}^{\top} \theta + \frac{\lambda}{2}\left\|\theta\right\|^2, \\ f_{\mathrm{i}}(\theta) =& \left(1 - \sigma(\xi_{\mathrm{i}} s_{\mathrm{i}}^{\top} \theta)\right)^2 + \frac{\lambda}{2}\left\|\theta\right\|^2, \\ f_{\mathrm{i}}(\theta) =& -\ln \sigma(\xi_{\mathrm{i}} s_{\mathrm{i}}^{\top} \theta) + \ln \sigma(\xi_{\mathrm{i}} s_{\mathrm{i}}^{\top} \theta + \mu) + \frac{\lambda}{2}\left\|\theta\right\|^2, \end{aligned} \tag{3.14}$$

where $\lambda > 0$ and $\mu > 0$ are given parameters. In all these examples, Assumption 3.3.3 is satisfied for any sufficiently large R and $R - \epsilon$; the proof can be found in Appendix 3.8.1. Note that many of the functions listed above are not L-smooth on their entire domain, violating thus (part of) Assumption 3.3.1(i). Motivated by these examples, we will extend existing convergence results of DGD, replacing Assumption 3.3.1(i) with Assumption 3.3.3.

Network model: The network is modeled as a (possibly) directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where the set of vertices $\mathcal{V} = [m]$ coincides with the set of agents, and the set of edges $\mathcal{E}$ represents the agents' communication links: $(\mathrm{i}, \mathrm{j}) \in \mathcal{E}$ if and only if there is link directed from agent i to agent j. The in-neighborhood of agent i is defined as $\mathcal{N}_{\mathrm{i}}^{\mathrm{in}} = \{\mathrm{j} | (\mathrm{j}, \mathrm{i}) \in \mathcal{E}\} \cup \{\mathrm{i}\}$ and represents the set of agents that can send information to agent i (including agent i itself, for notational simplicity). The out-neighborhood of agent i is similarly defined $\mathcal{N}_{\mathrm{i}}^{\mathrm{out}} = \{\mathrm{j} | (\mathrm{i}, \mathrm{j}) \in \mathcal{E}\} \cup \{\mathrm{i}\}$. When the graph is undirected, these two sets coincide and we use $\mathcal{N}_{\mathrm{i}}$ to denote the neighborhood of agent i (with a slight abuse of notation, we use the same symbol $\mathcal{G}$ to denote either directed or undirected graphs). Given a nonnegative matrix $A \in \mathcal{M}_m(\mathbb{R})$, the directed graph induced by A is defined as $\mathcal{G}_A = (\mathcal{V}_A, \mathcal{E}_A)$, where $\mathcal{V}_A \triangleq [m]$ and $(\mathrm{j}, \mathrm{i}) \in \mathcal{E}_A$ if and only if $A_{\mathrm{ij}} > 0$. The set of roots of all the directed spanning trees in $\mathcal{G}_A$ is denoted by $\mathcal{R}_A$. We make the following blanket standard assumptions on $\mathcal{G}$.

Assumption 3.3.4 (On the network)**.** *The graph (resp. digraph) $\mathcal{G}$ is connected (resp. strongly connected).*

3.4 The DGD algorithm

Consider Problem (3.1) and assume that the network is modeled as an undirected graph $\mathcal{G}$. As described previously, the DGD algorithm is based on a decentralization of GD as described in (3.2). It is convenient to rewrite the update (3.2) in the matrix/vector form: Using the definition of *aggregate* function $F_c(x)$ [cf. (3.5)] and $x^\nu \triangleq [x_1^{\nu\top}, \dots x_m^{\nu\top}]^\top$, we have

$$x^{\nu+1} = W_D\, x^\nu - \alpha \nabla F_c(x^\nu), \tag{3.15}$$

given $x^0 \in \mathbb{R}^{md}$, where $W_D \triangleq D \otimes I_d$, and $D \in \mathcal{M}_m(\mathbb{R})$ satisfying the following assumption.

Assumption 3.4.1. *$D \in \mathcal{M}_m(\mathbb{R})$ is nonnegative, doubly-stochastic, and compliant to $\mathcal{G}$, i.e., $D_{\mathrm{ij}} > 0$ if and only if $(\mathrm{j}, \mathrm{i}) \in \mathcal{E}$, and $D_{\mathrm{ij}} = 0$ otherwise.*

3.4.1 Existing convergence results

Convergence of DGD applied to the nonconvex problem (3.1) has been established [98], [111], and summarized below.

Theorem 3.4.1 ([98], [111])**.** *Let Assumptions 2.1', 3.3.4 hold. Given arbitrary $x^0 \in \mathbb{R}^{md}$ and $0 < \alpha < \alpha_{\max} \triangleq \sigma_{\min}(I + D)/L_c$, let $\{x^\nu\}$ be the sequence generated by the DGD algorithm (3.15) under Assumption 3.4.1. Then $\{x^\nu\}$ is bounded and*

(i) [almost consensus]: *for all $\mathrm{i} \in [m]$ and $\nu \in \mathbb{N}_+$,*

$$\|x_{\mathrm{i}}^\nu - \bar{x}^\nu\| \leq (\sigma_2)^\nu \|x_{\mathrm{i}}^0\| + \frac{\alpha H}{1 - \sigma_2},$$

where $\sigma_2 < 1$ is the second largest singular value of D, and H is a universal upper-bound of $\{\|\nabla F_c(x^\nu)\|\}$;

(ii) [stationarity]: *every limit point x^∞ of $\{x^\nu\}$ is such that $x^\infty \in \mathrm{crit}\ L_\alpha$.*

In addition, if L_α is a KŁ function, then $\{x^\nu\}$ is globally convergent to some $x^\infty \in \mathrm{crit}\ L_\alpha$.

Although L-smoothness of f_{i}'s is a common assumption in the literature, above convergence results can also be established without this condition but under Assumption 3.3.3–see Remark 3.4.1 and Appendix 3.8.2 for details.

Since (3.15) is the gradient update applied to L_α (cf. (3.6)), non-convergence of the DGD algorithm to strict saddle points of L_α can be established by applying [89, Corollary 2] to (3.6); the statement is given in Theorem 3.4.2 below. The following extra assumption on the weight matrix D is needed.

Assumption 3.4.2. *The matrix $D \in \mathcal{M}_m(\mathbb{R})$ is nonsingular.*

Theorem 3.4.2. *Consider Problem (3.1), under Assumptions 2.1', 3.3.4, and further assume that each f_{i} is a KŁ function. Let $\{x^\nu\}$ be the sequence generated by the DGD algorithm with step-size $0 < \alpha < \frac{\sigma_{\min}(D)}{L_c}$ and weight matrix D satisfying Assumptions 3.4.1 and 3.4.2. Then, the stable set of strict saddles has measure zero. Therefore, $\{x^\nu\}$ convergences almost surely to a SoS solution of L_α, where the probability is taken over the random initialization $x^0 \in \mathbb{R}^{md}$.*

As anticipated in Sec. 3.2.1, the above second-order guarantees are not satisfactory as they do not provide any information on the behavior of DGD near critical points of F, including the strict saddles of F. In the following, we fill this gap. We first show that the DGD algorithm convergences to neighborhood of the critical points of F, whose size is controlled by the step-size $\alpha > 0$ (cf. Section 3.4.2). Then, we prove that, for sufficiently small $\alpha > 0$, such critical points are almost surely SoS solutions of (3.1), where the randomization is taken on the initial point (cf. Section 3.4.3).

3.4.2 DGD converges to a neighborhood of critical points of F

Let us begin with introducing the definition of ϵ-critical points of F.

Definition 3.4.1. *A point $\theta \in \mathbb{R}^d$ such that $||\nabla F(\theta)|| \leq \varepsilon$, with $\varepsilon > 0$, is called ε-critical point of F. The set of ε-critical points of F is denoted by* $\mathrm{crit}_\varepsilon F$.

In this section, we prove that when the step-size is sufficiently small and DGD is initialized in a compact set, the iterates $\{x_{\mathrm{i}}^\nu\}$, $\mathrm{i} \in [m]$, converge to an arbitrarily small neighborhood of

critical points of F–the result is formally stated in Theorem 3.4.3. Roughly speaking, this is proved chaining the following intermediate results:

i) Lemma 3.4.1: Every limit point of DGD is an $\mathcal{O}(\alpha)$-critical point of F;

ii) Lemma 3.4.2: Every sequence generated by DGD for given $\alpha > 0$ and initialization in a compact set, is enclosed in some compact set, for all $\alpha \downarrow 0$; and

iii) Lemma 3.4.3: Any ϵ-critical point of F achievable by DGD is arbitrarily close to a critical point of F, when ϵ is sufficiently small. Lemma 3.4.1 implies that, for any given $\epsilon > 0$, one can find arbitrarily small $\alpha > 0$ so that every limit point of each $\{x_{\mathrm{i}}^{\nu}\}$ (whose existence is guaranteed by Lemma 3.4.2) is an ϵ-critical point of F. Finally, Lemma 3.4.3 guarantees that every such ϵ-critical point can be made arbitrarily close to a critical point of F as $\epsilon \downarrow 0$. The proof of the above three lemmata follows.

Lemma 3.4.1. *Let Assumptions 2.1' and 3.3.4 hold. Given arbitrary $x^0 \in \mathbb{R}^{md}$ and $0 < \alpha < \sigma_{\min}(I+D)/L_c$, every limit point $x^\infty = [x_1^{\infty\top}, \dots, x_m^{\infty\top}]^\top$ of $\{x^\nu\}$ generated by the DGD algorithm satisfies $\bar{x}^\infty \in \mathrm{crit}_{K\alpha} F$, with $\bar{x}^\infty \triangleq (1/m)\sum_{\mathrm{i}=1}^m x_{\mathrm{i}}^\infty$ and $K = m\sqrt{m}L_c H/(1-\sigma_2)$, where H and σ_2 are defined in Theorem 3.4.1.*

Proof. By Theorem 3.4.1(ii), $(1\otimes I)^\top \nabla L_\alpha(x^\infty) = 0$, which using (3.5) and the column stochasticity of D yields $(1\otimes I)^\top \nabla F_c(x^\infty) = 0$. Hence,

$$\begin{aligned}\left\|\nabla F(\bar{x}^\infty)\right\| =& \left\|(1\otimes I)^\top \left(\nabla F_c(1\otimes \bar{x}^\infty) - \nabla F_c(x^\infty)\right)\right\| \\ \leq& L_c\sqrt{m}\left\|x^\infty - 1\otimes \bar{x}^\infty\right\| \overset{(a)}{\leq} \alpha\cdot\frac{m\sqrt{m}L_c H}{1-\sigma_2},\end{aligned} \tag{3.16}$$

where in (a) we used Theorem 3.4.1(i). □

To proceed, we limit DGD initialization to $x_{\mathrm{i}}^0 \in \mathcal{X}_{\mathrm{i}}$, $\mathrm{i}\in[m]$, where $\mathcal{X}_{\mathrm{i}}^0 \subseteq \mathbb{R}^d$ is some compact set with positive Lebesgue measure.

Lemma 3.4.2. *Consider Problem (3.1), under Assumptions 2.1', 3.3.3 and 3.3.4. Let $\{x^\nu(\alpha, x^0)\}$ be any sequence generated by DGD under Assumption 3.4.1, with step-size α and initialization x^0. Then, there exists a bounded set $\mathcal{Y}$ such that $\{x^\nu(\alpha,x^0)\}\subseteq\mathcal{Y}$, for all $0<\alpha\leq\alpha_{\max} = \sigma_{\min}(I+D)/L_c$ and $x_{\mathrm{i}}^0 \in \mathcal{X}_{\mathrm{i}} \subseteq \mathcal{B}_R^d, \mathrm{i}\in[m]$, where R is defined in Assumption 3.3.3.*

Proof. We proceed by induction. For the sake of notation, throughout the proof, we will use for $x^\nu(\alpha, x^0)$ the shorthand x^ν. Define $h \triangleq \max_{\mathrm{i}\in[m],\theta\in\mathcal{B}_R^d} ||\nabla f_\mathrm{i}(\theta)||$. By assumption, there holds $x_\mathrm{i}^0 \in \mathcal{B}_R^d$, for all i. Suppose $x_\mathrm{i}^\nu \in \mathcal{B}_R^d$, for all i. If $x_\mathrm{i}^\nu \in \mathcal{B}_{R-\epsilon}^d$ and $\alpha \leq \epsilon D_\mathrm{ii}/h$, then $x_\mathrm{i}^\nu - \frac{\alpha}{D_\mathrm{ii}}\nabla f_\mathrm{i}(x_\mathrm{i}^\nu) \in \mathcal{B}_R^d$, since

$$\left\|x_\mathrm{i}^\nu - \frac{\alpha}{D_\mathrm{ii}}\nabla f_\mathrm{i}(x_\mathrm{i}^\nu)\right\| \leq \left\|x_\mathrm{i}^\nu\right\| + \frac{\alpha}{D_\mathrm{ii}}\left\|\nabla f_\mathrm{i}(x_\mathrm{i}^\nu)\right\| \leq R - \epsilon + \frac{\alpha h}{D_\mathrm{ii}}. \tag{3.17}$$

If $x_\mathrm{i}^\nu \in \mathcal{S}_{R,\epsilon}$ and $\alpha \leq 2D_\mathrm{ii}\delta(R-\epsilon)/h^2$, then $x_\mathrm{i}^\nu - \frac{\alpha}{D_\mathrm{ii}}\nabla f_\mathrm{i}(x_\mathrm{i}^\nu) \in \mathcal{B}_R^d$, since

$$\begin{aligned}\left\|x_\mathrm{i}^\nu - \frac{\alpha}{D_\mathrm{ii}}\nabla f_\mathrm{i}(x_\mathrm{i}^\nu)\right\|^2 &= \left\|x_\mathrm{i}^\nu\right\|^2 - \frac{2\alpha||x_\mathrm{i}^\nu||}{D_\mathrm{ii}}\left\langle \frac{x_\mathrm{i}^\nu}{||x_\mathrm{i}^\nu||}, \nabla f_\mathrm{i}(x_\mathrm{i}^\nu)\right\rangle + \frac{\alpha^2}{D_\mathrm{ii}^2}\left\|\nabla f_\mathrm{i}(x_\mathrm{i}^\nu)\right\|^2 \\ &\leq R^2 - \frac{2\alpha\delta(R-\epsilon)}{D_\mathrm{ii}} + \frac{\alpha^2 h^2}{D_\mathrm{ii}^2}.\end{aligned} \tag{3.18}$$

By agents' updates $x_\mathrm{i}^{\nu+1} = \sum_{\mathrm{j}\neq\mathrm{i}} D_\mathrm{ij}x_\mathrm{j}^\nu + D_\mathrm{ii}(x_\mathrm{i}^\nu - \frac{\alpha}{D_\mathrm{ii}}\nabla f_\mathrm{i}(x_\mathrm{i}^\nu))$ and convexity of the norm, we conclude that if $x_\mathrm{i}^\nu \in \mathcal{B}_R^d$, for all i, and $0 < \alpha \leq \alpha_b \triangleq \min_\mathrm{i} \min\{\epsilon D_\mathrm{ii}/h, 2D_\mathrm{ii}\delta(R-\epsilon)/h^2\}$, then $x_\mathrm{i}^{\nu+1} \in \mathcal{B}_R^d$. This proves that, for $\alpha \in (0, \alpha_b]$, any sequence $\{x_\mathrm{i}^\nu\}$ initialized in $\mathcal{B}_R^d$ lies in $\mathcal{B}_R^d$, for all i.

We prove now the same result for $\alpha \in [\alpha_b, \sigma_{\min}(I+D)/L_c]$. Note that since each f_i is coercive (cf. Assumption 2.1'(ii)), any sublevel set of L_α is compact. Also, since $\{L_\alpha(x^\nu)\}$ is non-increasing for all $\alpha \in (0, \sigma_{\min}(I+D)/L_c]$ (cf. [111, lemma 2]), then $\{x^\nu\} \subseteq \mathcal{L}_{L_\alpha}(F_c(x^0) + \frac{1}{2\alpha}||x^0||_{I-W}^2)$, and furthermore,

$$\begin{aligned}&\mathcal{L}_{L_\alpha}\left(F_c(x^0) + \frac{1}{2\alpha}||x^0||_{I-W}^2\right) \subseteq \mathcal{L}_{L_\alpha}\left(F_c(x^0) + \frac{1}{2\alpha_b}||x^0||_{I-W}^2\right) \\ \subseteq &\mathcal{L}_{F_c}\left(F_c(x^0) + \frac{1}{2\alpha_b}||x^0||_{I-W}^2\right) \subseteq \mathcal{L}_{F_c}\left(\max_{x_\mathrm{i}^0\in\mathcal{B}_R^d,\mathrm{i}\in[m]}\left\{F_c(x^0) + \frac{1}{2\alpha_b}||x^0||_{I-W}^2\right\}\right).\end{aligned} \tag{3.19}$$

Since $||x^0||_{I-W}^2 \leq 2||x^0||^2$, it follows

$$\mathcal{L}_{L_\alpha}\left(F_c(x^0) + \frac{1}{2\alpha}||x^0||_{I-W}^2\right) \subseteq \underbrace{\mathcal{L}_{F_c}\left(\max_{x_\mathrm{i}^0\in\mathcal{B}_R^d,\mathrm{i}\in[m]}\left\{\sum_{\mathrm{i}=1}^m f_\mathrm{i}(x_\mathrm{i}^0)\right\} + \frac{R^2}{\alpha_b}\right)}_{\triangleq\bar{\mathcal{L}}}. \tag{3.20}$$

The statement of the lemma hods with $\mathcal{Y} = \bar{\mathcal{L}} \cup \prod_{\mathrm{i}=1}^m \mathcal{B}_R^d$. □

The following lemma shows that any ϵ-critical point of F achievable by DGD (i.e., any point in $\text{crit}_\varepsilon F \cap \bar{\mathcal{Y}}$) can be made arbitrarily close to a critical point of F, when $\epsilon > 0$ (and thus $\alpha > 0$) is sufficiently small.

Lemma 3.4.3. *Suppose $F : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable. For any given compact set $\bar{\mathcal{Y}} \subseteq \mathbb{R}^d$, there holds*

$$\lim_{\varepsilon \to 0} \max_{q \in \text{crit}_\varepsilon F \cap \bar{\mathcal{Y}}} \text{dist}(q, \text{crit } F) = 0. \tag{3.21}$$

Proof. We prove the lemma by contradiction. Suppose

$$\limsup_{\varepsilon \to 0} \max_{q \in \text{crit}_\varepsilon F \cap \bar{\mathcal{Y}}} \text{dist}(q, \text{crit } F) = \gamma > 0. \tag{3.22}$$

Then, one can construct $\{q^\nu\}$ with $q^\nu \in \text{crit}_{1/\nu} F \cap \bar{\mathcal{Y}}$ such that $\text{dist}(q^\nu, \text{crit} F) \geq \gamma$ for all $\nu \in \mathbb{N}$. Since ∇F is continuous, $\text{crit}_1 F$ is closed and $\text{crit}_1 F \cap \bar{\mathcal{Y}}$ is compact. Note that $\{q^\nu\} \subseteq \text{crit}_1 F \cap \bar{\mathcal{Y}}$, which ensures $\{q^\nu\}$ is bounded. Let $\{q^{t_\nu}\}$ be a convergent subsequence of $\{q^\nu\}$; its limit point q^∞ satisfies $\text{dist}(q^\infty, \text{crit } F) \geq \gamma$. By construction, for any $\acute{\nu} \in \mathbb{N}$, $\{q^{t_\nu}\}$ eventually settles in $\text{crit}_{1/\acute{\nu}} F \cap \bar{\mathcal{Y}}$, thus $q^\infty \in \text{crit}_{1/\acute{\nu}} F \cap \bar{\mathcal{Y}}$. This means that $||\nabla F(q^{\acute{\nu}})|| \leq 1/\acute{\nu}$, for all $\acute{\nu} \in \mathbb{N}$, implying $||\nabla F(q^\infty)|| = 0$. Hence $\text{dist}(q^\infty, \text{crit } F) = 0$, which contradicts (3.22). □

We can now combine Lemmas 3.4.1-3.4.3 with Theorem 3.4.1(i) and state the main result of this section.

Theorem 3.4.3. *Let Assumptions 2.1', 3.3.3 and 3.3.4 hold. Let $\epsilon > 0$. There exists $\bar{\alpha} > 0$ (which depends on ϵ) such that with any initialization $x_{\text{i}}^0 \in \mathcal{X}_{\text{i}}^0 \subseteq \mathcal{B}_R^d$ ($R > 0$ is defined in Assumption 3.3.3), $\text{i} \in [m]$, and any step-size $0 < \alpha \leq \bar{\alpha}$, all the limit points $x^\infty(\alpha, x^0) = [x_1^\infty(\alpha, x^0)^\top, \ldots, x_m^\infty(\alpha, x^0)^\top]^\top$ of the sequence $\{x^\nu(\alpha, x^0)\}$, generated by DGD satisfies*

$$\text{dist}\Big(\bar{x}^\infty(\alpha, x^0), \text{crit F}\Big) < \epsilon \quad \textit{and} \quad \Big\|x^\infty(\alpha, x^0) - 1 \otimes \bar{x}^\infty(\alpha, x^0)\Big\| < \epsilon, \tag{3.23}$$

where $\bar{x}^\infty(\alpha, x^0) \triangleq (1/m) \sum_{\text{i}=1}^m x_{\text{i}}^\infty(\alpha, x^0)$.

Proof. Combining Lemmata 3.4.1-3.4.3 proves that there exists some $\alpha_1 > 0$ such that $\text{dist}(\bar{x}^\infty(\alpha, x^0), \text{crit F}) < \epsilon$, for all $\alpha \leq \alpha_1$. In addition, Theorem 3.4.1(i), with $H = \sup_{x \in \mathcal{Y}} F_c(x)$, implies that there exists some $\alpha_2 > 0$ such that $||x^\infty(\alpha, x^0) - 1 \otimes \bar{x}^\infty(\alpha, x^0)|| < \epsilon$, for all $\alpha \leq \alpha_2$. Hence, choosing $\bar{\alpha} = \min\{\alpha_1, \alpha_2\}$ proves (3.23). □

3.4.3 DGD likely converges to a neighborhood of SoS solutions of F

We study now second-order guarantees of DGD. Our path to prove almost sure convergence to a neighborhood of SoS solutions of (3.1) will pass through the non-convergence of DGD to strict saddles of L_α (cf. Theorem 3.4.1). Roughly speaking, our idea is to show that whenever $\bar{x}^\infty = 1/m \sum_{\text{i}=1}^m x_\text{i}^\infty$ belongs to a sufficiently small neighborhood of a strict saddle of F inside the region (3.23), $x^\infty = [x_1^{\infty\top}, \dots, x_m^{\infty\top}]^\top$ must be a *strict saddle of* L_α. The escaping properties of DGD from strict saddles of L_α will then ensure that it is unlikely that $\{\bar{x}^\nu = 1/m \sum_{\text{i}=1}^m x_\text{i}^\nu\}$ gets trapped in a neighborhood of a strict saddle of F, thus ending in a neighborhood of a SoS solution of (3.1). Proposition 3.4.1 makes this argument formal; in particular, conditions (i)-(iii) identify the neighborhood of a strict saddle of F with the mentioned escaping properties.

Proposition 3.4.1. *Consider the setting of Lemma 3.4.2 and further assume that Assumption 3.3.2 hold. Let $\bar{\mathcal{Y}}$ be the image of the compact set $\mathcal{Y}$ (defined in Lemma 3.4.2) through the linear operator $(1_m \otimes I_d)^\top$. Suppose that the limit point $x^\infty = [x_1^{\infty\top}, \dots, x_m^{\infty\top}]^\top$ of $\{x^\nu\}$, along with $\bar{x}^\infty = 1/m \sum_{\text{i}=1}^m x_\text{i}^\infty$, satisfy*

(i) $\text{dist}(\bar{x}^\infty, \text{crit F}) < \dfrac{\delta}{2L_{\nabla^2}}$,

(ii) $\left\|x^\infty - 1 \otimes \bar{x}^\infty\right\| < \dfrac{\delta}{2mL_{\nabla_c^2}}$,

(iii) There exists $\theta^* \in \text{proj}_{\text{crit F}}(\bar{x}^\infty) \cap \Theta_{ss}^*$,

for some δ such that $\delta \leq -\lambda_{\min}\left(\nabla^2 F(\theta^*)\right), \forall \theta^* \in \Theta_{ss}^* \cap \bar{\mathcal{Y}}$. Then, x^∞ is a strict saddle point of L_α.

Proof. Given $\theta \in \mathbb{R}^d$, let $\boldsymbol{v}(\theta)$ denote the unitary eigenvector of $\nabla^2 F(\theta)$ associated with the smallest eigenvalue, and define $\tilde{\boldsymbol{v}}(\theta) \triangleq 1 \otimes \boldsymbol{v}(\theta)$. Then, we have

$$\begin{aligned}
&\tilde{\boldsymbol{v}}(\theta)^\top \nabla^2 L_\alpha(x^\infty)\tilde{\boldsymbol{v}}(\theta) \overset{(a)}{=} \tilde{\boldsymbol{v}}(\theta)^\top \nabla^2 F_c(x^\infty)\tilde{\boldsymbol{v}}(\theta) \\
&\quad \leq \boldsymbol{v}(\theta)^\top \nabla^2 F(\theta)\boldsymbol{v}(\theta) \\
&\qquad + ||\nabla^2 F(\bar{x}^\infty) - \nabla^2 F(\theta)|| \left\|\boldsymbol{v}(\theta)\right\|^2 + ||\nabla^2 F_c(x^\infty) - \nabla^2 F_c(1 \otimes \bar{x}^\infty)|| \left\|\tilde{\boldsymbol{v}}(\theta)\right\|^2 \\
&\quad \overset{(b)}{\leq} \boldsymbol{v}(\theta)^\top \nabla^2 F(\theta)\boldsymbol{v}(\theta) + L_{\nabla^2} \left\|\bar{x}^\infty - \theta\right\| + m\, L_{\nabla^2_c} \left\|x^\infty - 1 \otimes \bar{x}^\infty\right\|
\end{aligned} \tag{3.24}$$

where (a) follows from $\tilde{\boldsymbol{v}}(\theta) \in \text{null}(W_D - I)$; and (b) is due to Assumption 3.3.2. Let us now evaluate (3.24) at some θ^* as defined in condition (iii) of the proposition; using $\boldsymbol{v}(\theta^*)^\top \nabla^2 F(\theta^*)\boldsymbol{v}(\theta^*) \leq -\delta$ and conditions (i) and (ii), yields $\tilde{\boldsymbol{v}}(\theta^*)^\top \nabla^2 L_\alpha(x^\infty)\ \tilde{\boldsymbol{v}}(\theta^*) < 0$. By the Rayleigh-Ritz theorem, it must be $\lambda_{\min}(\nabla^2 L_\alpha(x^\infty)) < 0$. This, together with $x^\infty \in$ crit L_α (cf. Theorem 3.4.1(ii)), proves the proposition. □

Invoking now Theorem 3.4.3, we infer that there exists a sufficiently small $\alpha > 0$ such that conditions (i) and (ii) of Proposition 3.4.1 are always satisfied, implying that x^∞ is a strict saddle of L_α if there exists a strict saddle of F "close" to $\bar{x}^\infty$ [in the sense of (iii)]. This is formally stated next.

Corollary 3.4.1. *Consider the setting of Theorem 3.4.3 and Proposition 3.4.1. There exists a sufficiently small $\alpha > 0$ such that, if $\text{proj}_{\text{crit F}}(\bar{x}^\infty) \cap \Theta^*_{ss} \neq \emptyset$, then x^∞ is a strict saddle of L_α.*

To state our final result, let us introduce the following merit function: given $x = [x_1^\top, \dots, x_m^\top]^\top$ let

$$M(x) \triangleq \max\left(\text{dist}(\bar{x}, \mathcal{X}_{SOS}), \left\|x - 1 \otimes \bar{x}\right\|\right),$$

where $\mathcal{X}_{SOS}$ denotes the set of SoS solutions of (3.1), and $\bar{x} = 1/m \sum_{\text{i}=1}^m x_{\text{i}}$. $M(x)$ capture the distance of the average $\bar{x}$ from the set of SoS solutions of (3.1) and well as the consensus disagreement of the agents' local variables $\bar{x}_{\text{i}}$.

Theorem 3.4.4. *Consider Problem (3.1) under Assumptions 2.1', 3.3.2, 3.3.3, and 3.3.4; further assume that each f_{i} is a KŁ function. For every $\epsilon > 0$, there exists sufficiently small $0 < \bar{\alpha} < \frac{\sigma_{\min}(D)}{L_c}$ such that*

$$\mathbb{P}_{x^0}\Big(M(x^\infty) \leq \epsilon\Big) = 1,$$

where $x^\infty = [x_1^{\infty\top}, \ldots, x_m^{\infty\top}]^\top$ is the limit point of the sequence $\{x^\nu\}$ generated by the DGD algorithm (3.15) with $\alpha \in (0, \bar{\alpha}]$, the weight matrix D satisfying Assumptions 3.4.1 and 3.4.2, and initialization $x^0 \in \prod_{\mathrm{i}=1}^m \mathcal{X}_{\mathrm{i}}^0 \subseteq \prod_{\mathrm{i}=1}^m \mathcal{B}_R^d$; R is defined in Assumption 3.3.3 and each $\mathcal{X}_{\mathrm{i}}^0$ has positive Lebesgue measure; and the probability is taken over the initialization $x^0 \in \prod_{\mathrm{i}=1}^m \mathcal{X}_{\mathrm{i}}^0$. Furthermore, any $\theta^* \in \mathrm{proj}_{\mathrm{crit\ F}}(\bar{x}^\infty)$ is almost surely a SoS solution of F where $\bar{x}^\infty = (1/m)\sum_{\mathrm{i}=1}^m x_{\mathrm{i}}^\infty$.

Proof. For sufficiently small $\alpha < \bar{\alpha}_1$, if $\mathrm{proj}_{\mathrm{crit\ F}}(\bar{x}^\infty)$ contains a strict saddle point of F, then x^∞ is also a strict saddle point of L_α (by Corollary 3.4.1). Let also $\bar{\alpha}_2$ be a sufficiently small step-size such that every limit point x^∞ satisfies $\mathrm{dist}(\bar{x}^\infty, \mathrm{crit}\ F) \leq \epsilon$ and $\left\|x^\infty - 1 \otimes \bar{x}^\infty\right\| \leq \epsilon$ (by Theorem 3.4.3). Now consider DGD update (3.15) with $\alpha < \min\{\bar{\alpha}_1, \bar{\alpha}_2\}$ and x^0 being drawn randomly from the set of probability one measure $\prod_{\mathrm{i}=1}^m \mathcal{X}_{\mathrm{i}}^0$ for which the algorithm converges to a SoS solution of L_α (by Theorem 3.4.2[2]). Finally, by the above properties of α, it holds that $M(x^\infty) \leq \epsilon$ and $\mathrm{proj}_{\mathrm{crit\ F}}(\bar{x}^\infty)$ must contain only SoS solutions of F. Therefore, there exists a $\theta^* \in \mathrm{crit}\ F$ such that $\theta^* \in \mathcal{X}_{SoS}$ and $\|\bar{x}^\infty - \theta^*\| \leq \epsilon$. □

Remark 3.4.1. *All (first- and second-order) convergence results of DGD established in this section remain valid when ∇f_{i}'s are not globally Lipschitz continuous [Assumption 3.3.1(i)] but Assumption 3.3.3 holds. Specifically, Theorems 3.4.1, 3.4.2, 3.4.3 and Lemmata 3.4.1-3.4.2 hold if one replaces Assumption 3.3.1(i) with Assumption 3.3.3 and the global Lipschitz constant L_c with the Lipschitz constant of ∇F_c restricted to the compact set $\tilde{\mathcal{Y}}$, defined in Appendix 3.8.2, where we refer to for the technical details.*

[2]↑Note that the conclusion of Theorem 3.4.2 is valid also when the set of initial points is restricted to $\prod_{\mathrm{i}=1}^m \mathcal{X}_{\mathrm{i}}^0$, as $\prod_{\mathrm{i}=1}^m \mathcal{X}_{\mathrm{i}}^0$ has positive measure (the Cartesian product of sets with positive measure has positive measure– cf. [124, Sec. 35]).

3.5 DOGT Algorithms

The family of DOGT algorithms is introduced in Sec. 3.1.2. We begin here rewriting (3.3)-(3.4) in matrix/vector form. Denoting $x^\nu \triangleq [x_1^{\nu\top}, \dots x_m^{\nu\top}]^\top$ and $y^\nu \triangleq [y_1^{\nu\top}, \dots y_m^{\nu\top}]^\top$, we have

$$\begin{cases} x^{\nu+1} = W_R\, x^\nu - \alpha\, y^\nu, \\ y^{\nu+1} = W_C\, y^\nu + \nabla F_c\big(x^{\nu+1}\big) - \nabla F_c\big(x^\nu\big), \end{cases} \tag{3.25}$$

where $W_R \triangleq R \otimes I_d$ and $W_C \triangleq C \otimes I_d$ with $R \triangleq (R_{\mathrm{ij}})_{\mathrm{i,j}=1}^m$ and $C \triangleq (C_{\mathrm{ij}})_{\mathrm{i,j}=1}^m$ being some *column-stochastic* and *row-stochastic* matrices (respectively) compliant to the graph $\mathcal{G}$ (cf. Assumption 3.5.1 below). The initialization of (3.25) is set to $x^0 \in \mathbb{R}^{md}$ and $y^0 \in \nabla F_c(x^0) + \operatorname{span}(W_C - I)$. Note that the latter condition is instrumental to preserve the *total-sum* of the y-variables, namely $\sum_{\mathrm{i}} y_{\mathrm{i}}^\nu = \sum_{\mathrm{i}} f_{\mathrm{i}}(x_{\mathrm{i}}^\nu)$ (which holds due to the column-stochasticity of matrix C–cf. Assumption 3.5.1). This property is imperative for the y-variables to track the sum-gradient. Notice that the condition used in the literature [72], [104], [106], [125], [126]–$y^0 = \nabla F_c(x^0)$–is a special case of the proposed initialization. On the practical side, this initialization can be enforced in a distributed way, with minimal coordination. For instance, agents first choose independently a vector $y_{\mathrm{i}}^{-1} \in \mathbb{R}^d$; then they run one step of consensus on the y-variables using the values y_{i}^{-1}'s and weights matrix C, and set $y_{\mathrm{i}}^0 = \nabla f_{\mathrm{i}}(x_{\mathrm{i}}^0) + \sum_{\mathrm{j}\in\mathcal{N}_{\mathrm{i}}^{\mathrm{in}}} C_{\mathrm{ij}} y_{\mathrm{j}}^{-1} - y_{\mathrm{i}}^{-1}$, resulting in $y^0 \in \nabla F_c(x^0) + \operatorname{span}(W_C - I)$.

Different choices for R and C are possible, resulting in different existing algorithms. For instance, if $R = C \in \mathcal{M}_m(\mathbb{R})$ are doubly-stochastic matrices compliant to the graph $\mathcal{G}$, (3.25) reduces to the NEXT algorithm [71], [72] (or the one in [73], when (3.1) is convex). If R and C are allowed to be time-varying (suitably chosen) (3.25) reduces to the SONATA algorithm applicable to (possibly time-varying) digraphs [6], [103]–[105] [or the one later proposed in [126] for strongly convex instances of (3.1)]. Finally, if R and C are chosen according to Assumption 3.5.1 below, the scheme (3.25) becomes the algorithm proposed independently in [107] and [106], for strongly convex objectives in (3.1), and implementable over fixed digraphs.

Assumption 3.5.1. *(On the matrices R and C) The weight matrices $R, C \in \mathcal{M}_m(\mathbb{R})$ satisfy the following:*

(i) R is nonnegative row-stochastic and $R_{\mathrm{ii}} > 0$, for all $\mathrm{i} \in [m]$;

(ii) C is nonnegative column-stochastic and $C_{\mathrm{ii}} > 0$, for all $\mathrm{i} \in [m]$;

(iii) The graphs $\mathcal{G}_R$ and $\mathcal{G}_{C^\top}$ each contain at least one spanning tree; and $\mathcal{R}_R \cap \mathcal{R}_{C^\top} \neq \emptyset$.

It is not difficult to check that matrices R and C above exist if and only if the digraph $\mathcal{G}$ is strongly connected; however, $\mathcal{G}_R$ and $\mathcal{G}_{C^\top}$ need not be so. Several choices for such matrices are discussed in [106], [107]. Here, we only point out the following property of R and C, as a consequence of Assumption 3.5.1, which will be used in our analysis. The result is a consequence of [127, Lemma 1].

Lemma 3.5.1. *Given R and C satisfying Assumption 3.5.1 with stochastic left eigenvector r (resp. right eigenvector c) of R (resp. C) associated with the eigenvalue one, then there exist matrix norms*

$$||x||_R \triangleq ||\operatorname{diag}(\sqrt{r})x \operatorname{diag}(\sqrt{r})^{-1}||_2, \tag{3.26}$$

$$||x||_C \triangleq ||\operatorname{diag}(\sqrt{c})^{-1}x \operatorname{diag}(\sqrt{c})||_2, \tag{3.27}$$

such that $\rho_R \triangleq \|R - 1r^\top\|_R < 1$ and $\rho_C \triangleq \|C - c1^\top\|_C < 1$. Furthermore, $r^\top c > 0$.

Using Lemma 3.5.1, it is not difficult to check that the following properties hold:

$$\rho_R = \sigma_2\left(\operatorname{diag}(\sqrt{r})R \operatorname{diag}(\sqrt{r})^{-1}\right), \tag{3.28}$$

$$\rho_C = \sigma_2\left(\operatorname{diag}(\sqrt{c})^{-1}C \operatorname{diag}(\sqrt{c})\right), \tag{3.29}$$

$$||R||_R = ||1r^T||_R = ||I - 1r^T||_R = 1, \tag{3.30}$$

$$||C||_C = ||c1^T||_C = ||I - c1^T||_R = 1. \tag{3.31}$$

The vector norms associated with above matrix norms are

$$||x||_R = ||\operatorname{diag}(\sqrt{r})x||_2, \tag{3.32}$$

$$||x||_C = ||\operatorname{diag}(\sqrt{c})^{-1}x||_2; \tag{3.33}$$

and $||\cdot||_a \leq K_{a,b}||\cdot||_b$ holds for $a, b \in \{R, C, 2\}$ with

$$\begin{aligned} K_{R,2} &= \sqrt{r_{\max}}, \quad K_{2,R} = 1/\sqrt{r_{\min}}, \\ K_{C,2} &= 1/\sqrt{c_{\min}}, \quad K_{2,C} = \sqrt{c_{\max}}, \\ K_{R,C} &= \sqrt{r_{\max} c_{\max}}, \quad K_{C,R} = 1/\sqrt{c_{\min} r_{\min}}, \end{aligned} \tag{3.34}$$

where $r_{\min}$ (resp. $c_{\min}$) and $r_{\max}$ (resp. $c_{\max}$) are minimum and maximum elements of r (resp. c).

Convergence of DOGT algorithms in the form (3.25) (with R and C satisfying Assumption 3.5.1) has not been studied in the literature when F is nonconvex. In next subsection we fill this gap and provide a full characterization of the convergence behavior of DOGT including its second-order guarantees.

3.5.1 First-order convergence & rate analysis

In this section, we study asymptotic convergence to first-order stationary solutions; we assume $d = 1$ (scalar optimization variables); while this simplifies the notation, all the conclusions hold for the general case $d > 1$. As in [107], define the weighted sums

$$\bar{x}^\nu \triangleq r^\top x^\nu, \quad \bar{y}^\nu \triangleq 1^\top y^\nu, \quad \text{and} \quad \bar{g}^\nu \triangleq 1^\top \nabla F_c(x^\nu), \tag{3.35}$$

where we recall that r is the Perron vector associated with R (cf. Lemma 3.5.1). Note that ∇F_c is L_c-Lipschitz continuous with $L_c \triangleq L_{\max}$.

Using (3.25), it is not difficult to check that the following holds

$$\bar{x}^{\nu+1} = \bar{x}^\nu - \zeta\alpha\bar{y}^\nu - \alpha r^\top (y^\nu - c\bar{y}^\nu) \quad \text{and} \quad \bar{y}^\nu = \bar{g}^\nu, \tag{3.36}$$

where c is the Perron vector associated with C, and $\zeta \triangleq r^\top c > 0$ (cf. Lemma 3.5.1).

3.5.1.1 Descent on F

Using the descent lemma along with (3.36) yields

$$\begin{aligned}F(\bar{x}^{\nu+1}) &= F\left(\bar{x}^{\nu}-\zeta\alpha\bar{y}^{\nu}-\alpha r^{\top}\left(y^{\nu}-c\bar{y}^{\nu}\right)\right)\\ &\leq F(\bar{x}^{\nu})-\zeta\alpha\left\langle\nabla F(\bar{x}^{\nu}),\bar{y}^{\nu}\right\rangle-\alpha\left\langle\nabla F(\bar{x}^{\nu}),r^{\top}\left(y^{\nu}-c\bar{y}^{\nu}\right)\right\rangle\\ &\quad+\frac{L}{2}\left\|\zeta\alpha\bar{y}^{\nu}+\alpha r^{\top}\left(y^{\nu}-c\bar{y}^{\nu}\right)\right\|^{2}.\end{aligned}$$

Adding/subtracting suitably chosen terms we obtain

$$\begin{aligned}F(\bar{x}^{\nu+1}) \leq & F(\bar{x}^{\nu})-\zeta\alpha\left\langle\nabla F(\bar{x}^{\nu})-\bar{y}^{\nu},\bar{y}^{\nu}\right\rangle-\zeta\alpha|\bar{y}^{\nu}|^{2}\\ &-\alpha\left\langle\nabla F(\bar{x}^{\nu})-\bar{y}^{\nu},r^{\top}\left(y^{\nu}-c\bar{y}^{\nu}\right)\right\rangle-\alpha\left\langle\bar{y}^{\nu},r^{\top}\left(y^{\nu}-c\bar{y}^{\nu}\right)\right\rangle\\ &+L\zeta^{2}\alpha^{2}|\bar{y}^{\nu}|^{2}+L\alpha^{2}\left\|y^{\nu}-c\bar{y}^{\nu}\right\|^{2}\\ \leq & F(\bar{x}^{\nu})+\frac{\zeta\alpha}{2\epsilon_{1}}|\nabla F(\bar{x}^{\nu})-\bar{y}^{\nu}|^{2}+\frac{\zeta\alpha\epsilon_{1}}{2}|\bar{y}^{\nu}|^{2}-\zeta\alpha|\bar{y}^{\nu}|^{2}\\ &+\frac{\alpha}{2}|\nabla F(\bar{x}^{\nu})-\bar{y}^{\nu}|^{2}+\frac{\alpha}{2}\left\|y^{\nu}-c\bar{y}^{\nu}\right\|^{2}+\frac{\alpha\epsilon_{2}}{2}|\bar{y}^{\nu}|^{2}+\frac{\alpha}{2\epsilon_{2}}\left\|y^{\nu}-c\bar{y}^{\nu}\right\|^{2}\\ &+L\zeta^{2}\alpha^{2}|\bar{y}^{\nu}|^{2}+L\alpha^{2}\left\|y^{\nu}-c\bar{y}^{\nu}\right\|^{2}\\ = & F(\bar{x}^{\nu})+\left(\frac{\zeta\alpha\epsilon_{1}}{2}-\zeta\alpha+\frac{\alpha\epsilon_{2}}{2}+L\zeta^{2}\alpha^{2}\right)|\bar{y}^{\nu}|^{2}\\ &+\left(\frac{\zeta\alpha}{2\epsilon_{1}}+\frac{\alpha}{2}\right)|\nabla F(\bar{x}^{\nu})-\bar{y}^{\nu}|^{2}+\left(\frac{\alpha}{2}+\frac{\alpha}{2\epsilon_{2}}+L\alpha^{2}\right)\left\|y^{\nu}-c\bar{y}^{\nu}\right\|^{2},\end{aligned}\tag{3.37}$$

where ϵ_1 and ϵ_2 are some arbitrary positive quantities (to be chosen). By $\bar{y}^{\nu}=\bar{g}^{\nu}$ [cf. (3.36)], it holds that

$$|\nabla F(\bar{x}^{\nu})-\bar{y}^{\nu}|=\left|\sum_{\mathrm{i}=1}^{m}\nabla f_{\mathrm{i}}(\bar{x}^{\nu})-\sum_{\mathrm{i}=1}^{m}\nabla f_{\mathrm{i}}(x_{\mathrm{i}}^{\nu})\right|\leq L_{c}\sqrt{m}\left\|x^{\nu}-1\bar{x}^{\nu}\right\|.\tag{3.38}$$

Combining (3.37) and (3.38) yields

$$\begin{aligned}
&F(\bar{x}^{\nu+1}) \\
&\leq F(\bar{x}^{\nu}) + \left(\frac{\zeta\alpha\epsilon_1}{2} - \zeta\alpha + \frac{\alpha\epsilon_2}{2} + L\zeta^2\alpha^2\right)|\bar{y}^{\nu}|^2 \\
&\quad + mL_c^2K_{2,R}^2\left(\frac{\zeta\alpha}{2\epsilon_1} + \frac{\alpha}{2}\right)\left\|x^{\nu} - 1\bar{x}^{\nu}\right\|_R^2 + K_{2,C}^2\left(\frac{\alpha}{2} + \frac{\alpha}{2\epsilon_2} + L\alpha^2\right)\left\|y^{\nu} - c\bar{y}^{\nu}\right\|_C^2,
\end{aligned} \tag{3.39}$$

where $K_{2,R} = 1/\sqrt{r_{\min}}$ and $K_{2,C} = \sqrt{c_{\max}}$ [cf. (3.60)].

3.5.1.2 Bounding the consensus and gradient tracking errors

Let us bound the consensus error $\|x^{\nu} - 1\bar{x}^{\nu}\|_R$. Using $\|z+w\|_R^2 \leq (1+\epsilon)\|x\|_R^2 + (1+1/\epsilon)\|y\|_R^2$, for arbitrary $z, w \in \mathbb{R}^d$ and $\epsilon > 0$, along with Lemma 3.5.1, yields

$$\begin{aligned}
&\left\|x^{\nu+1} - 1\bar{x}^{\nu+1}\right\|_R^2 = \left\|\left(R - 1r^\top\right)(x^{\nu} - 1\bar{x}^{\nu}) - \alpha\left(I - 1r^\top\right)(y^{\nu} - 1\bar{y}^{\nu})\right\|_R^2 \\
&\leq (1+\epsilon_x)\left\|\left(R - 1r^\top\right)(x^{\nu} - 1\bar{x}^{\nu})\right\|_R^2 + \alpha^2\left(1 + \frac{1}{\epsilon_x}\right)\left\|\left(I - 1r^\top\right)(y^{\nu} - 1\bar{y}^{\nu})\right\|_R^2 \\
&\leq \rho_R^2(1+\epsilon_x)\left\|x^{\nu} - 1\bar{x}^{\nu}\right\|_R^2 + \alpha^2\left(1 + \frac{1}{\epsilon_x}\right)\|I - 1r^\top\|_R^2\left\|y^{\nu} - 1\bar{y}^{\nu}\right\|_R^2 \\
&\overset{(3.30)}{\leq} \rho_R^2(1+\epsilon_x)\left\|x^{\nu} - 1\bar{x}^{\nu}\right\|_R^2 + 2\alpha^2\left(1 + \frac{1}{\epsilon_x}\right)\left\|y^{\nu} - c\bar{y}^{\nu}\right\|_R^2 \\
&\qquad + 2\alpha^2(1 + \frac{1}{\epsilon_x})\left\|(1-c)\bar{y}^{\nu}\right\|_R^2 \\
&\leq \rho_R^2(1+\epsilon_x)\left\|x^{\nu} - 1\bar{x}^{\nu}\right\|_R^2 + \alpha^2K_2\left\|y^{\nu} - c\bar{y}^{\nu}\right\|_C^2 + \alpha^2K_3|\bar{y}^{\nu}|_2^2,
\end{aligned} \tag{3.40}$$

where $\epsilon_x > 0$ is arbitrary and we defined

$$K_2 \triangleq 2K_{R,C}^2\left(1 + \frac{1}{\epsilon_x}\right), \quad K_3 \triangleq 2m(1 + \frac{1}{\epsilon_x}). \tag{3.41}$$

Similarly, the tracking error can be bounded as

$$
\begin{aligned}
&\left\|y^{\nu+1}-c\bar{y}^{\nu+1}\right\|_C^2=\left\|\left(C-c1^\top\right)y^\nu+\left(I-c1^\top\right)\left(\nabla F_c(x^{\nu+1})-\nabla F_c(x^\nu)\right)\right\|_C^2\\
&\le(1+\epsilon_y)\left\|\left(C-c1^\top\right)(y^\nu-c\bar{y}^\nu)\right\|_C^2\\
&\quad+(1+\frac{1}{\epsilon_y})\left\|\left(I-c1^\top\right)\left(\nabla F_c(x^{\nu+1})-\nabla F_c(x^\nu)\right)\right\|_C^2\\
&\overset{(3.31)}{\le}\rho_C^2(1+\epsilon_y)\left\|y^\nu-c\bar{y}^\nu\right\|_C^2+K_{C,2}^2L_c^2\left(1+\frac{1}{\epsilon_y}\right)\left\|x^{\nu+1}-x^\nu\right\|^2\\
&\overset{(a)}{=}\rho_C^2(1+\epsilon_y)\left\|y^\nu-c\bar{y}^\nu\right\|_C^2\\
&+3K_{C,2}^2L_c^2\left(1+\frac{1}{\epsilon_y}\right)\left[\left\|(R-I)(x^\nu-1\bar{x}^\nu)\right\|^2+\alpha^2\left\|y^\nu-c\bar{y}^\nu\right\|^2+\alpha^2|\bar{y}^\nu|^2\left\|c\right\|^2\right]\\
&\overset{(3.30)}{\le}\rho_C^2(1+\epsilon_y)\left\|y^\nu-c\bar{y}^\nu\right\|_C^2\\
&+3K_{C,2}^2L_c^2\left(1+\frac{1}{\epsilon_y}\right)\left[K_{2,R}^2\left\|x^\nu-1\bar{x}^\nu\right\|^2+K_{2,C}^2\alpha^2\left\|y^\nu-c\bar{y}^\nu\right\|_C^2+\alpha^2|\bar{y}^\nu|^2\right]\\
&=\rho_C^2(1+\epsilon_y)\left\|y^\nu-c\bar{y}^\nu\right\|_C^2\\
&+3K_{C,2}^2L_c^2\left(1+\frac{1}{\epsilon_y}\right)\left[K_{2,R}^2\left\|x^\nu-1\bar{x}^\nu\right\|^2+K_{2,C}^2\alpha^2\left\|y^\nu-c\bar{y}^\nu\right\|_C^2+\alpha^2|\bar{y}^\nu|^2\right]\\
&\le\left(\rho_C^2+\frac{\alpha^2K_4}{\epsilon_y}\right)(1+\epsilon_y)\left\|y^\nu-c\bar{y}^\nu\right\|_C^2+\alpha^2K_5|\bar{y}^\nu|_2^2+K_6\left(1+\frac{1}{\epsilon_y}\right)\left\|x^\nu-1\bar{x}^\nu\right\|_R^2,
\end{aligned}
\tag{3.42}
$$

where in (a) we used $x^{\nu+1}-x^\nu=(R-I)(x^\nu-1\bar{x}^\nu)-\alpha(y^\nu-c\bar{y}^\nu)-\alpha c\bar{y}^\nu$ and the Jensen's inequality; and in the last inequality we defined

$$
K_4=3K_{C,2}^2K_{2,C}^2L_c^2,\quad K_5=3K_{C,2}^2L_c^2,\quad K_6=3K_{C,2}^2K_{2,R}^2L_c^2. \tag{3.43}
$$

3.5.1.3 Lyapunov function

Let us introduce now the candidate Lyapunov function: denoting $J_R\triangleq 1r^\top$ and $J_C\triangleq c1^\top$, define

$$
L(x,y)\triangleq F_c(J_Rx)+\left\|(I-J_R)x\right\|_R^2+\varkappa\left\|(I-J_C)y\right\|_C^2, \tag{3.44}
$$

where $\varkappa > 0$ is a positive constant (to be properly chosen). Combining (3.39)-(3.42) and using $\bar{y}^\nu = \bar{g}^\nu = \sum_{\mathrm{i}=1}^m \nabla f_{\mathrm{i}}(x_{\mathrm{i}}^\nu)$ [cf. (3.36)] leads to the following descent property for L:

$$L(x^{\nu+1}, y^{\nu+1}) \leq L(x^\nu, y^\nu) - d(x^\nu, y^\nu)^2, \tag{3.45}$$

where

$$d(x,y) \triangleq \sqrt{(1-\tilde{\rho}_R)\left\|(I-J_R)x\right\|_R^2 + \varkappa(1-\tilde{\rho}_C)\left\|(I-J_C)y\right\|_C^2 + \Gamma\left|\sum_{\mathrm{i}=1}^m \nabla f_{\mathrm{i}}(x_{\mathrm{i}})\right|^2} \tag{3.46}$$

and

$$\begin{aligned}
\tilde{\rho}_R &\triangleq \rho_R^2(1+\epsilon_x) + \frac{\alpha m L_c^2 K_{2,R}^2}{2}\left(1+\frac{\zeta}{\epsilon_1}\right) + \varkappa K_6\left(1+\frac{1}{\epsilon_y}\right),\\
\tilde{\rho}_C &\triangleq \rho_C^2(1+\epsilon_y) + \frac{\alpha K_{2,C}^2}{2\varkappa}\left(1+\frac{1}{\epsilon_2}\right) + \alpha^2\left(\frac{LK_{2,C}^2+K_2}{\varkappa} + K_4\left(1+\frac{1}{\epsilon_y}\right)\right),\\
\Gamma &\triangleq \left(\zeta - \frac{\epsilon_1\zeta}{2} - \frac{\epsilon_2}{2}\right)\alpha - \left(L\zeta^2 + K_3 + K_5\varkappa\right)\alpha^2.
\end{aligned} \tag{3.47}$$

Note that the function $d(\bullet,\bullet)$ is a valid measure of optimality/consensus for DOGT: i) it is continuous; and ii) $d(x,y) = 0$ implies $x_{\mathrm{i}} = x_{\mathrm{j}} = x^*$, for all $\mathrm{i},\mathrm{j} \in [m]$ and some x^* such that $\sum_{\mathrm{i}=1}^m \nabla f_{\mathrm{i}}(x^*) = 0$, meaning that all x_{i} are consensual and equal to a critical point of F.

To ensure $\tilde{\rho}_R < 1$, $\tilde{\rho}_C < 1$, and $\Gamma > 0$ in $d(x,y)$, we choose the free parameters ϵ_x, ϵ_y, ϵ_1, ϵ_2, and $\varkappa$ as follows:

$$\begin{aligned}
&0 < \epsilon_x < \frac{1-\rho_R^2}{2\rho_R^2}, \qquad 0 < \epsilon_y < \frac{1-\rho_C^2}{\rho_C^2},\\
&\epsilon_1 = \epsilon_2 = \epsilon, \qquad 0 < \epsilon < \frac{2\zeta}{1+\zeta}, \qquad 0 < \varkappa \leq \frac{\rho_R^2\epsilon_x}{K_6(1+1/\epsilon_y)},
\end{aligned} \tag{3.48}$$

and finally, $\alpha > 0$ must satisfy

$$\begin{aligned}
\alpha &< \frac{2}{mL_c^2 K_{2,R}^2 \left(1 + \frac{\zeta}{\epsilon}\right)} \left(1 - \rho_R^2(1 + 2\epsilon_x)\right), \\
\alpha &< \frac{1 - \rho_C^2(1 + \epsilon_y)}{\frac{1}{2\varkappa} K_{2,C}^2 \left(1 + \frac{1}{\epsilon} + 2L\right) + \frac{K_2}{\varkappa} + K_4 \left(1 + \frac{1}{\epsilon_y}\right)}, \\
\alpha &< \frac{\zeta - \frac{\epsilon}{2}(\zeta + 1)}{L\zeta^2 + K_3 + K_5 \varkappa}.
\end{aligned} \tag{3.49}$$

Substituting (3.34), (3.41), and (3.43) in (3.49) and setting for simplicity

$$\epsilon_x = \frac{1 - \rho_R^2}{4\rho_R^2}, \quad \epsilon_y = \frac{1 - \rho_C^2}{2\rho_C^2}, \quad \epsilon = \frac{\xi}{1 + \xi}, \quad \varkappa = \frac{c_{\min} r_{\min}}{24 L_c^2}(1 - \rho_R^2)(1 - \rho_C^2), \tag{3.50}$$

we obtain the following sufficient conditions for (3.49):

$$\begin{aligned}
\alpha &\leq \tilde{\alpha}_1 \triangleq \frac{r_{\min}(1 - \rho_R^2)}{3mL_c^2}, \\
\alpha &\leq \tilde{\alpha}_2 \triangleq \frac{(1 - \rho_R^2)^2 (1 - \rho_C^2)^2 r_{\min}^2 c_{\min}^2}{1152 L_c^2 (2 + L)}, \\
\alpha &\leq \tilde{\alpha}_3 \triangleq \frac{r_{\min} c_{\min}(1 - \rho_R^2)}{2(L + 16m)}.
\end{aligned} \tag{3.51}$$

A further simplification, leads to the following final more restrictive condition on α:

$$0 < \alpha \leq \frac{(1 - \rho_R^2)^2 (1 - \rho_C^2)^2 r_{\min}^2 c_{\min}^2}{1152 L_c^2 (L + 16m)}. \tag{3.52}$$

The descent property (3.45) readily implies the following convergence result for $\{L(x^\nu, y^\nu)\}$ and $\{d(x^\nu, y^\nu)\}$.

Lemma 3.5.2. *Under Assumptions 3.3.1, 3.3.4, and 3.5.1, and the above choice of parameter, there hold:*

(i) The sequence $\{L(x^\nu, y^\nu)\}$ converges;

(ii) $\sum_{\nu=0}^{\infty} d(x^\nu, y^\nu)^2 < \infty$, and thus $\lim_{\nu \to \infty} d(x^\nu, y^\nu) = 0$.

We conclude this subsection by lower bounding $d(x^\nu, y^\nu)$ by the magnitude of the gradient of the Lyapunov function L. This will allow us to transfer the convergence properties of

$\{d(x^\nu, y^\nu)\}$ to $\{||\nabla L(x^\nu, y^\nu)||\}$. The lemma below will also be useful to establish global convergence of DOGT under the KŁ property (cf. Sec. 3.5.2.1).

Lemma 3.5.3. *Let* $\nabla L(x^\nu, y^\nu) \triangleq (\nabla_x L(x^\nu, y^\nu), \nabla_y L(x^\nu, y^\nu))$, *where* $\nabla_x L$ *(resp.* $\nabla_y L$*) are the gradient of* L *with respect to the first (resp. second) argument. In the setting above, there holds*

$$\left\|\nabla L(x^\nu, y^\nu)\right\| \leq M d(x^\nu, y^\nu), \quad \nu \geq 0, \tag{3.53}$$

with

$$M = \sqrt{2}\max\left(\frac{(2r_{\max} + L_c\sqrt{m})^2}{r_{\min}(1-\tilde{\rho}_R)}, \frac{2\varkappa c_{\max}}{c_{\min}^2(1-\tilde{\rho}_C)}, \frac{1}{\Gamma}\right)^{\frac{1}{2}}. \tag{3.54}$$

Proof. Recall that $J_R = 1r^\top$ and $J_C = c1^\top$. By definition (3.44) and Lemma 3.5.1, we can write

$$\begin{aligned}
\nabla_x L(x^\nu, y^\nu) &= J_R^\top \nabla F_c(J_R x^\nu) + 2(I - J_R)^\top \operatorname{diag}(r)(I - J_R)x^\nu \\
&\overset{(a)}{=} r\,\bar{y}^\nu + J_R^\top\left(\nabla F_c(J_R x^\nu) - \nabla F_c(x^\nu)\right) \\
&\quad + 2(I - J_R)^\top \operatorname{diag}(r)(x^\nu - 1\bar{x}^\nu), \\
\nabla_y L(x^\nu, y^\nu) &= 2\varkappa(I - J_C)^\top \operatorname{diag}(c)^{-1}(I - J_C)y^\nu \\
&= 2\varkappa(I - J_C)^\top \operatorname{diag}(c)^{-1}(y^\nu - c\bar{y}^\nu),
\end{aligned} \tag{3.55}$$

where (a) is due to $\bar{y}^\nu = \bar{g}^\nu$ (cf. (3.36)). Thus there holds

$$\begin{aligned}
||\nabla_x L(x^\nu, y^\nu)|| &\leq ||r||\;|\bar{y}^\nu| + ||J_R^\top\left(\nabla F_c(J_R x^\nu) - \nabla F_c(x^\nu)\right)|| \\
&\quad + 2||(I - J_R)^\top \operatorname{diag}(r)(x^\nu - 1\bar{x}^\nu)|| \\
&\overset{(b)}{\leq} |\bar{y}^\nu| + K_{2,R}\left(2r_{\max} + L_c\sqrt{m}\right)||x^\nu - 1\bar{x}^\nu||_R, \\
||\nabla_y L(x^\nu, y^\nu)|| &\overset{(c)}{\leq} 2\varkappa K_{2,C} c_{\min}^{-1}||y^\nu - c\bar{y}^\nu||_C,
\end{aligned} \tag{3.56}$$

where (b) holds due to $||\operatorname{diag}(r)||_R = ||\operatorname{diag}(r)||_2 = r_{\max}$, $||r|| \leq 1$, $||J_R||_2 \leq \sqrt{m}$ and (3.30); (c) is due to $||\operatorname{diag}(c)^{-1}||_C = ||\operatorname{diag}(c)^{-1}||_2 = c_{\min}^{-1}$ and (3.31). Eq. (3.53) follows readily from (3.56). □

3.5.1.4 Main result

We can now state the main convergence result of DOGT to critical points of F.

Theorem 3.5.1. *Consider Problem (3.1), and suppose that Assumptions 3.3.1 and 3.3.4 are satisfied. Let $\{(x^\nu, y^\nu)\}$ be the sequence generated by the DOGT Algorithm (3.25), with R and C satisfying Assumption 3.5.1, and α chosen according to (3.52) [or (3.50)]; let $\{\bar{x}^\nu\}$ and $\{\bar{y}^\nu\}$ be defined in (3.35); and let $\{d(x^\nu, y^\nu)\}$ be defined in (3.46). Given $\epsilon > 0$, let $T_\epsilon = \min\{\nu \in \mathbb{N}_+ : d(x^\nu, y^\nu) \leq \epsilon\}$. Then, there hold*

(i) [consensus]: $\lim_{\nu\to\infty} \left\|x^\nu - 1\bar{x}^\nu\right\| = 0$ *and* $\lim_{\nu\to\infty} \bar{y}^\nu = 0$;

(ii) [stationarity]: *let x^∞ be a limit point of $\{x^\nu\}$; then, $x^\infty = \theta^\infty 1$, for some $\theta^\infty \in$ crit F;*

(iii) [sublinear rate]: $T_\epsilon = o(1/\epsilon^2)$.

Proof. (i) follows readily from Lemma 3.5.2(ii).

We prove (ii). Let (x^∞, y^∞) be a limit point of $\{(x^\nu, y^\nu)\}$. By (i), it must be $(I - J_R)x^\infty = 0$, implying $x^\infty = 1\theta^\infty$, for some $\theta^\infty \in \mathbb{R}$. Also, $\lim_{\nu\to\infty} 1^\top \nabla F_c(x^\nu) = \lim_{\nu\to\infty} \bar{g}^\nu = \lim_{\nu\to\infty} \bar{y}^\nu = 0$, which together with the continuity of ∇F_c, yields $0 = 1^\top \nabla F_c(1\theta^\infty) = \nabla F(\theta^\infty)$. Therefore, $\theta^\infty \in$ crit F.

We prove now (iii). Using (3.45) and the definition of T_ϵ, we can write

$$\frac{T_\epsilon}{2}\epsilon^2 \leq \sum_{t=\lfloor \frac{T_\epsilon}{2} \rfloor + 1}^{T_\epsilon} d(x^t, y^t)^2 \leq l^{\lfloor \frac{T_\epsilon}{2} \rfloor + 1} - l^{T_\epsilon + 1}, \tag{3.57}$$

where we used the shorthand $l^\nu \triangleq L(x^\nu, y^\nu)$. Consider the following two cases: (1) $T_\epsilon \to \infty$ as $\epsilon \to 0$, then $l^{\lfloor \frac{T_\epsilon}{2} \rfloor + 1} - l^{T_\epsilon + 1} \to 0$ (recall that $\{l^\nu\}$ converges, cf. Lemma 3.5.2(i)); and (2) $T_\epsilon < \infty$ as $\epsilon \to 0$, then $\{l^\nu\}$ converges in a finite number of iterations. Therefore, by (3.57), we have $T_\epsilon = o(1/\epsilon^2)$. □

Note that, as a direct consequence of Lemma 3.5.3, one can infer the following further property of the limit points (x^∞, y^∞) of the sequence $\{(x^\nu, y^\nu)\}$: any such a (x^∞, y^∞) is a critical point of L [defined in (3.44)].

3.5.2 Convergence under the KŁ property

We now strengthen the subsequence convergence result in Theorem 3.5.1, under the additional assumption that F is a KŁ function [113], [114]: We prove that the entire sequence

$\{x^\nu\}$ converges to a critical point of F (cf. Theorem 3.5.2), and establish asymptotic convergence rates (cf. Theorem 3.5.3). We extend the analysis developed in [115], [128] for centralized first-order methods to our distributed setting and complement it with a rate analysis. The major difference with [115] is that the *sufficient decent* condition postulated in [115] is neither satisfied by the objective value sequence $\{F(x^\nu)\}$ (as requested in [115]), due to consensus and gradient tracking errors, nor by the Lyapunov function sequence $\{L(x^\nu, y^\nu)\}$, which instead satisfies (3.45). A key step to cope with this issue is to establish necessary connections between $\nabla L(x, y)$ and $d(x, y)$ (defined in (3.44) and (3.46), respectively)–see Lemma 3.53 and Proposition 3.5.1.

3.5.2.1 Convergence analysis

We begin proving the following abstract intermediate results similar to [115] but extended to our distributed setting, which is at the core of the subsequent analysis; we still assume $d = 1$ without loss of generality.

Proposition 3.5.1. *In the setting of Theorem 3.5.1, let L defined in (3.44) is KŁ at some $\acute{z} \triangleq (\acute{x}, \acute{y})$. Denote by $\mathcal{V}_{\acute{z}}$, η, and $\phi : [0, \eta) \to \mathbb{R}_+$ the objects appearing in Definition 3.3.1. Let $\rho > 0$ be such that $\mathcal{B}(\acute{z}, \rho)^{2md} \subseteq \mathcal{V}_{\acute{z}}$. Consider the sequence $\{z^\nu \triangleq (x^\nu, y^\nu)\}$ generated by the DOGT Algorithm (3.25), with initialization $z^0 \triangleq (x^0, y^0)$; and define $\acute{l} \triangleq L(\acute{z})$ and $l^\nu \triangleq L(z^\nu)$. Suppose that*

$$\acute{l} < l^\nu < \acute{l} + \eta, \quad \forall \nu \geq 0, \tag{3.58}$$

and

$$K\, M\, \phi(l^0 - \acute{l}) + \left\| z^0 - \acute{z} \right\| < \rho, \tag{3.59}$$

where

$$K = \sqrt{3}(1 + L_c) \max\left(\frac{4nK_{||}^2}{1 - \tilde{\rho}_R}, \frac{K_{||}^2}{\varkappa(1 - \tilde{\rho}_C)} \left(\alpha + \frac{2\sqrt{m}}{1 + L_c} \right)^2, \alpha^2/\Gamma \right)^{1/2}, \tag{3.60}$$

and $M > 0$ is defined in (3.53) (cf. Lemma 3.5.3).

Then, $\{z^\nu\}$ satisfies:

(i) $z^\nu \in \mathcal{B}(\acute{z}, \rho)^{2md}$, *for all* $\nu \geq 0$;

(ii) $\sum_{t=k}^{\nu} \left\| z^{t+1} - z^t \right\| \leq KM \left(\phi(l^k - \acute{l}) - \phi(l^{\nu+1} - \acute{l}) \right)$ *for all* $\nu, k \geq 0$ *and* $\nu \geq k$;

(iii) $l^\nu \to \acute{l}, \quad$ as $\nu \to \infty$.

Proof. Throughout the proof, we will use the following shorthand $d^\nu \triangleq d(x^\nu, y^\nu)$. Let $d^\nu > 0$, for all integer $\nu \geq 0$; otherwise, $\{x^\nu\}$ converges in a finite number of steps, and its limit point is $x^\infty = 1\theta^\infty$, for some $\theta^\infty \in \text{crit } F$.

We first bound the "length" $\sum_{t=k}^{\nu} \left\| z^{t+1} - z^t \right\|$. By (3.25), there holds

$$\begin{aligned} x^{\nu+1} - x^\nu &= (R - I)\,(x^\nu - 1\bar{x}^\nu) - \alpha\,(y^\nu - c\bar{y}^\nu) - \alpha c\bar{y}^\nu, \\ y^{\nu+1} - y^\nu &= (C - I)\,(y^\nu - c\bar{y}^\nu) + \nabla F_c(x^{\nu+1}) - \nabla F_c(x^\nu). \end{aligned}$$

Using $||A||_2 \leq \sqrt{m}||A||_\infty$ and $||A||_2 \leq \sqrt{m}||A||_1$, with $A \in \mathcal{M}_m(\mathbb{R})$; and $||R - I||_\infty \leq 2$ and $||C - I||_1 \leq 2$, we get

$$\begin{aligned} \sum_{t=k}^{\nu} \left\| x^{t+1} - x^t \right\| &\leq \sum_{t=k}^{\nu} 2\sqrt{m} \left\| x^t - 1\bar{x}^t \right\| + \alpha \left\| y^t - c\bar{y}^t \right\| + \alpha |\bar{y}^t|, \\ \sum_{t=k}^{\nu} \left\| y^{t+1} - y^t \right\| &\leq \sum_{t=k}^{\nu} 2\sqrt{m} \left\| y^t - c\bar{y}^t \right\| + L_c \sum_{t=k}^{\nu} \left\| x^{t+1} - x^t \right\|, \end{aligned}$$

where L_c is the Lipschitz constant of ∇F_c. The above inequalities imply

$$\begin{aligned} &\sum_{t=k}^{\nu} \left\| z^{t+1} - z^t \right\| \\ &\leq \sum_{t=k}^{\nu} 2(1 + L_c)\sqrt{m} K_{||} \left\| x^t - 1\bar{x}^t \right\|_R + K_{||} \left(\alpha(1 + L_c) + 2\sqrt{m} \right) \left\| y^t - c\bar{y}^\nu \right\|_C \\ &\quad + \alpha(1 + L_c)|\bar{y}^t| \leq K \sum_{t=k}^{\nu} d^t, \end{aligned} \tag{3.61}$$

where K is defined in (3.60).

We prove now the proposition, starting from statement (ii). Multiplying both sides of (3.45) by $\phi(l^\nu - \acute{l})$ and using $\phi(l^\nu - \acute{l}) > 0$ [due to property (iii) in Definition 3.3.1 and (3.58)] and the concavity of ϕ, yield

$$(d^\nu)^2 \ \phi(l^\nu - \acute{l}) \leq \phi(l^\nu - \acute{l}) \ \left(l^\nu - l^{\nu+1}\right) \leq \phi(l^\nu - \acute{l}) - \phi(l^{\nu+1} - \acute{l}). \tag{3.62}$$

For all $z \in \mathcal{V}_{\acute{z}} \cap [\acute{l} < L < \acute{l} + \eta]$, the KŁ inequality (3.7) holds; hence, assuming $z^t \in \mathcal{B}(\acute{z}, \rho)^{2md}$ for all $t = 0, \ldots, \nu$, yields

$$\phi(l^t - \acute{l})||\nabla L(z^t)|| \geq 1, \quad t = 0, \ldots, \nu, \tag{3.63}$$

which together with (3.62) and (3.53) (cf. Lemma 3.5.3), gives

$$M\left(\phi(l^t - \acute{l}) - \phi(l^{t+1} - \acute{l})\right) \geq d^t, \quad t = 0, \ldots, \nu,$$

and thus

$$M\left(\phi(l^k - \acute{l}) - \phi(l^{\nu+1} - \acute{l})\right) \geq \sum_{t=k}^{\nu} d^t. \tag{3.64}$$

Combining (3.64) with (3.61), we obtain

$$\sum_{t=k}^{\nu} \left\|z^{t+1} - z^t\right\| \leq KM\left(\phi(l^k - \acute{l}) - \phi(l^{\nu+1} - \acute{l})\right). \tag{3.65}$$

Ineq. (3.65) proves (ii) if $z^\nu \in \mathcal{B}(\acute{z}, \rho)^{2md}$ for all $\nu \geq 0$, which is shown next.

Now let us prove statement (i). Letting $k = 0$ in (3.65), by (3.59), we obtain

$$\left\|z^{\nu+1} - \acute{z}\right\| \leq KM\left(\phi(l^0 - \acute{l}) - \phi(l^{\nu+1} - \acute{l})\right) + \left\|z^0 - \acute{z}\right\| < \rho.$$

Therefore, $z^\nu \in \mathcal{B}(\acute{z}, \rho)^{2md}$, for all $\nu \geq 0$.

We finally prove statement (iii). Inequalities (3.53) (cf. Lemma 3.5.3) and (3.63) imply

$$\phi(l^\nu - \acute{l}) \ d^\nu \geq 1/M, \quad \nu \geq 0. \tag{3.66}$$

On the other hand, by Lemma 3.5.2-(i), as $\nu \to \infty$, we have $l^\nu \to p$, for some $p \geq \acute{l}$. In fact, $p = \acute{l}$, otherwise $p - \acute{l} > 0$, which would contradict (3.66) (because $d^\nu \to 0$ as $\nu \to \infty$ and $\phi(p - \acute{l}) < \infty$). □

Roughly speaking, Proposition 3.5.1 states that, if the algorithm is initialized in a suitably chosen neighborhood of a point at which L satisfies the KŁ property, then it will converge to that point. Combining this property with the subsequence convergence proved in Theorem 3.5.2 we can obtain global convergence of the sequence to critical points of F, as stated next.

Theorem 3.5.2. *Consider the setting of Theorem 3.5.1, and furthermore assume that F is real-analytic. Any sequence $\{(x^\nu, y^\nu)\}$ generated by the DOGT Algorithm (3.25) converges to some $(x^\infty, y^\infty) \in$ crit L. Furthermore, $x^\infty = 1 \otimes \theta^\infty$, for some $\theta^\infty \in$ crit F.*

Proof. Let $z^\infty \triangleq (x^\infty, y^\infty)$ be a limit point of $\{z^\nu \triangleq (x^\nu, y^\nu)\}$. Since $\{l^\nu \triangleq L(z^\nu)\}$ is convergent (cf. Lemma 3.5.2) and L is continuous, we deduce $l^\nu \to l^\infty \triangleq L(z^\infty)$. Since F is real-analytic, L is real analytic (due to Lemma 3.5.1 and the fact that summation/composition of functions preserve real-analytic property [129, Prop. 2.2.8]) and thus KŁ at at z^∞ [113]. Set $\acute{z} = z^\infty$ and $\acute{l} = l^\infty$; denote by $\mathcal{V}_{\acute{z}}$, η, and $\phi : [0, \eta) \to \mathbb{R}_+$ the objects appearing in Definition 3.3.1; and let $\rho > 0$ be such that $\mathcal{B}(\acute{z}, \rho)^{2md} \subseteq \mathcal{V}_{\acute{z}}$. By the continuity of ϕ and the properties above, we deduce that there exists an integer ν_0 such that i) $l^\nu \in (\acute{l}, \acute{l} + \eta)$, for all $\nu \geq \nu_0$; and ii) $K\ M\ \phi(l^{\nu_0} - \acute{l}) + \left\| z^{\nu_0} - \acute{z} \right\| < \rho$, with K and M defined in (3.60) and (3.53), respectively. Global convergence of the sequence $\{z^\nu\}$ follows by applying Proposition 3.5.1 to the sequence $\{z^{\nu+\nu_0}\}$.

Finally, by Lemma 3.5.2(ii), $d(x^\nu, y^\nu) \to 0$ as $\nu \to \infty$. Invoking the continuity of ∇L and Lemma 3.5.3, we have $\nabla L(x^\infty, y^\infty) = 0$, thus $(x^\infty, y^\infty) \in$ crit L. By Theorem 3.5.1(ii), $x^\infty = 1 \otimes \theta^\infty$, with $\theta^\infty \in$ crit F. □

In the following theorem, we provide some convergence rate estimates.

Theorem 3.5.3. *In the setting of Theorem 3.5.2, let L be a KŁ function with $\phi(s) = cs^{1-\theta}$, for some constant $c > 0$ and $\theta \in [0, 1)$. Let $\{z^\nu \triangleq (x^\nu, y^\nu)\}$ be a sequence generated by DOGT Algorithm (3.25). Then, there hold:*

(i) If $\theta = 0$, $\{z^\nu\}$ converges to z^∞ in a finite number of iterations;

(ii) If $\theta \in (0, 1/2]$, then $||z^\nu - z^\infty|| \leq C\tau^\nu$, for all $\nu \geq \bar{\nu}$ for some $\tau \in [0, 1)$, $\bar{\nu} \in \mathbb{N}_+$, $C > 0$;

(iii) If $\theta \in (1/2, 1)$, then $||z^\nu - z^\infty|| \leq C\nu^{-\frac{1-\theta}{2\theta-1}}$, for all $\nu \geq \bar{\nu}$ for some $\bar{\nu} \in \mathbb{N}_+$, $C > 0$.

Proof. For sake of simplicity of notation, denote $d^\nu \triangleq d(x^\nu, y^\nu)$ and define $D^\nu \triangleq \sum_{t=\nu}^{\infty} d^t$. By (3.61), we have

$$\left\|z^{\nu+1} - z^\infty\right\| \leq \sum_{t=\nu}^{\infty} \left\|z^{t+1} - z^t\right\| \leq K D^\nu. \tag{3.67}$$

It is then sufficient to establish the convergence rates for the sequence $\{D^\nu\}$.

By KŁ inequality (3.7) and (3.53), we have

$$Md^\nu \phi(l^\nu - l^\infty) \geq 1 \implies \tilde{M}(d^\nu)^{(1-\theta)/\theta} \geq (l^\nu - l^\infty)^{1-\theta}, \quad \forall \nu \geq \bar{\nu} \tag{3.68}$$

for sufficiently large $\bar{\nu}$, where $\tilde{M} = (Mc(1-\theta))^{(1-\theta)/\theta}$, $l^\nu \triangleq L(z^\nu)$, and $l^\infty \triangleq L(z^\infty)$. In addition, by (3.64) (setting $\acute{l} = l^\infty$), we have $D^\nu \leq M\phi(l^\nu - l^\infty) = Mc(l^\nu - l^\infty)^{1-\theta}$, which together with (3.68), yields

$$D^\nu \leq \tilde{M}Mc(d^\nu)^{(1-\theta)/\theta} = \tilde{M}Mc(D^\nu - D^{\nu+1})^{(1-\theta)/\theta}, \quad \forall \nu \geq \bar{\nu}. \tag{3.69}$$

The convergence rate estimates as stated in the theorem can be derived from (3.69), using the same line of analysis introduced in [128]. The remaining part of the proof is provided in Appendix 3.8.3 for completeness. □

3.5.3 Second-order guarantees

We prove that the DOGT algorithm almost surely converges to SoS solutions of (3.1), under a suitably chosen initialization and some additional conditions on the weight matrices R and C. Following a path first established in [88] and further developed in [89], the key to our argument for the non-convergence to strict saddle points of F lies in formulating the DOGT algorithm as a dynamical system while leveraging an instantiation of the stable manifold theorem, as given in [89, Theorem 2]. The nontrivial task is finding a self-map representing DOGT so that the stable set of the strict saddles of F is zero measure with

respect to the domain of the mapping; note that the domain of the map–which is the set of initialization points–is not full dimensional and is the same as the support of the probability measure.

Our analysis is organized in the following three steps: 1) Sec. 3.5.3.1 introduces the preparatory background; 2) Sec. 3.5.3.2 tailors the results of Step 1 to the DOGT algorithm; and 3) finally, Sec. 3.5.3.3 states our main results about convergence of the DOGT algorithm to SoS solutions of (3.1).

3.5.3.1 The stable manifold theorem and unstable fixed-points

Let $g : \mathcal{S} \to \mathcal{S}$ be a mapping from $\mathcal{S}$ to itself, where $\mathcal{S}$ is a manifold without boundary. Consider the dynamical system $u^{\nu+1} = g(u^\nu)$, with $u^0 \in \mathcal{S}$; we denote by g^ν the ν-fold composition of g. Our focus is on the analysis of the trajectories of the dynamical system around the fixed points of g; in particular we are interested in the set of unstable fixed points of g. We begin introducing the following definition.

Definition 3.5.1 (Chapter 3 of [130]). *The differential of the mapping $g : \mathcal{S} \to \mathcal{S}$, denoted as $\mathrm{D}g(u)$, is a linear operator from $\mathcal{T}(u) \to \mathcal{T}(g(u))$, where $\mathcal{T}(u)$ is the tangent space of $\mathcal{S}$ at $u \in \mathcal{S}$. Given a curve γ in $\mathcal{S}$ with $\gamma(0) = u$ and $\frac{d\gamma}{dt}(0) = v \in \mathcal{T}(u)$, the linear operator is defined as $\mathrm{D}g(u)v = \frac{d(g\circ\gamma)}{dt}(0) \in \mathcal{T}(g(u))$. The determinant of the linear operator $\det(\mathrm{D}g(u))$ is the determinant of the matrix representing $\mathrm{D}g(u)$ with respect to a standard basis.*[3]

We can now introduce the definition of the set of unstable fixed points of g.

Definition 3.5.2 (Unstable fixed points). *The set of unstable fixed points of g is defined as*

$$\mathcal{A}_g = \Big\{ u : g(u) = u,\ \mathrm{spradii}\Big(\mathrm{D}g(u)\Big) > 1 \Big\}. \tag{3.70}$$

The theorem below, which is based on the stable manifold theorem [131, Theorem III.7], provides tools to let us connect $\mathcal{A}_g$ with the set of limit points which $\{u^\nu\}$ can escape from.

[3]↑This determinant may not be uniquely defined, in the sense of being completely invariant to the basis used for the geometry. In this work, we are interested in properties of the determinant that are independent of scaling, and thus the potentially arbitrary choice of a standard basis does not affect our conclusions.

Theorem 3.5.4 ([89, Theorem 2]). *Let $g : \mathcal{S} \to \mathcal{S}$ be a $\mathcal{C}^1$ mapping and*

$$\det\left(\mathrm{D}g(u)\right) \neq 0, \quad \forall u \in \mathcal{S}.$$

Then, the set of initial points that converge to an unstable fixed point (termed stable set of $\mathcal{A}_g$) is zero measure in $\mathcal{S}$. Therefore,

$$\mathbb{P}_{u^0}\left(\lim_{\nu\to\infty} g^\nu(u^0) \in \mathcal{A}_g\right) = 0,$$

where the probability is taken over the starting point $u^0 \in \mathcal{S}$.

3.5.3.2 DOGT as a dynamical system

Theorem 3.5.4 sets the path to the analysis of the convergence of the DOGT algorithm to SoS solutions of F: it is sufficient to describe the DOGT algorithm by a proper mapping $g : \mathcal{S} \to \mathcal{S}$ satisfying the assumptions in the theorem and such that the non-convergence of $g^\nu(u^0)$, $u^0 \in \mathcal{S}$, to $\mathcal{A}_g$ implies the non-convergence of the DOGT algorithm to strict saddles of F.

We begin rewriting the DOGT in an equivalent and more convenient form. Define $h^\nu \triangleq y^\nu - \nabla F_c(x^\nu)$; (3.25) can be rewritten as

$$\begin{cases} x^{\nu+1} = W_R x^\nu - \alpha\left(h^\nu + \nabla F_c(x^\nu)\right); \\ h^{\nu+1} = W_C h^\nu + \left(W_C - I\right)\nabla F_c(x^\nu), \end{cases} \tag{3.71}$$

with arbitrary $x^0 \in \mathbb{R}^{md}$ and $h^0 \in \operatorname{span}(W_C - I)$. By Theorem 3.5.1, every limit point (x^∞, h^∞) of $\{(x^\nu, h^\nu)\}$ has the form $x^\infty = 1_m \otimes \theta^\infty$ and $h^\infty = -\nabla F_c(1_m \otimes \theta^\infty)$, for some $\theta^\infty \in \operatorname{crit} F$. We are interested in the non-convergence of (3.71) to such points whenever $\theta^\infty \in \operatorname{crit} F$ is a strict saddle of F. This motivates the following definition.

Definition 3.5.3 (Consensual strict saddle points)**.** *Let* $\Theta_{ss}^* = \{\theta^* \in \text{crit } F : \lambda_{min}(\nabla^2 F(\theta^*)) < 0\}$ *denote the set of strict saddles of* F*. The set of* consensual strict saddle points *is defined as*

$$\mathcal{U}^* \triangleq \left\{ \begin{bmatrix} 1_m \otimes \theta^* \\ -\nabla F_c(1_m \otimes \theta^*) \end{bmatrix} : \theta^\star \in \Theta_{ss}^* \right\}. \tag{3.72}$$

Roughly speaking, $\mathcal{U}^*$ represents the candidate set of "adversarial" limit points which any sequence generated by (3.71) should escape from. The next step is then to write (3.71) as a proper dynamical system whose mapping satisfies conditions in Theorem 3.5.4 and its set of unstable fixed points $\mathcal{A}_g$ is such that $\mathcal{U}^* \subseteq \mathcal{A}_g$.

Identification of g and $\mathcal{S}$. Define $u \triangleq (x, h)$, where $x \triangleq [x_1^\top, \ldots, x_m^\top]^\top$, $h \triangleq [h_1^\top, \ldots, h_m^\top]^\top$, and each $x_\mathrm{i}, h_\mathrm{i} \in \mathbb{R}^d$; its value at iteration ν is denoted by $u^\nu \triangleq (x^\nu, h^\nu)$. Consider the dynamical system

$$u^{\nu+1} = g(u^\nu), \quad \text{with} \quad g\left(u\right) \triangleq \begin{bmatrix} W_R x - \alpha \nabla F_c\left(x\right) - \alpha h \\ W_C h + \left(W_C - I\right) \nabla F_c\left(x\right) \end{bmatrix}, \tag{3.73}$$

and $u^0 \in \mathbb{R}^{md} \times \text{span}(W_C - I)$. The fixed-point iterate (3.73) describes the trajectory generated by the DOGT algorithm (3.71). However, the initialization imposed by DOGT leads to a g that maps $\mathbb{R}^{md} \times \text{span}(W_C - I)$ into $\mathbb{R}^{md} \times \mathbb{R}^{md}$. We show next how to unify the domain and codomain of g to a subspace $\mathcal{S} \subseteq \mathbb{R}^{md} \times \mathbb{R}^{md}$ as in form of the mapping in Theorem 3.5.4.

Applying (3.71) telescopically to the update of the h-variables yields: $h^\nu = W_C^\nu h^0 + \left(W_C - I\right) g_{\text{acc}}^\nu$, for all $\nu \geq 1$, where $g_{\text{acc}}^\nu \triangleq \sum_{t=0}^{\nu-1} W_C^t \nabla F_c\left(x^{\nu-t-1}\right)$. Denoting $\bar{h}^\nu \triangleq (1_m^\top \otimes I_d) h^\nu$, we have

$$\bar{h}^\nu = \cdots = \bar{h}^0, \quad \text{and} \quad h^\nu \in W_C^\nu h^0 + \text{span}\left(W_C - I\right) \quad \forall \nu \geq 1. \tag{3.74}$$

The initialization $h^0 \in \text{span}\left(W_C - I\right)$ in (3.71) naturally suggests the following $(2m-1)m$-dimensional linear subspace as candidate set $\mathcal{S}$:

$$\mathcal{S} \triangleq \mathbb{R}^{md} \times \text{span}\left(W_C - I\right). \tag{3.75}$$

Such an $\mathcal{S}$ also ensures that $g : \mathcal{S} \to \mathcal{S}$. In fact, by (3.74), $h^\nu \in \text{span}(W_C - I)$, for all $\nu \geq 1$, provided that $h^0 \in \text{span}\,(W_C - I)$. Therefore, $\{g^\nu(u^0)\} \subseteq \mathcal{S}$, for all $u^0 \in \mathcal{S}$.

Equipped with the mapping g in (3.73) and $\mathcal{S}$ defined in (3.75), we check next that the condition in Theorem 3.5.4 is satisfied; we then prove that $\mathcal{U}^* \subseteq \mathcal{A}_g$.

1) g is a diffeomorphism: To establish this property, we add the following extra assumption on the weight matrices R and C, which is similar to Assumption 3.4.2 for the DGD scheme.

Assumption 3.5.2. *Matrices $R \in \mathcal{M}_m(\mathbb{R})$ and $C \in \mathcal{M}_m(\mathbb{R})$ are nonsingular.*

The above condition is not particularly restrictive and it is compatible with Assumption 3.5.1. A rule of thumb is to choose $R = (\tilde{R} + I)/2$ and $C = (\tilde{C} + I)/2$, with $\tilde{R}$ and $\tilde{C}$ satisfying Assumption 3.5.1. The new matrices still satisfy Assumption 3.5.1 due to the following fact:

given two nonnegative matrices $A, B \in \mathcal{M}_m(\mathbb{R})$, if the directed graph associated with matrix A has a spanning tree and $B \geq \rho A$, for some $\rho > 0$, then the directed graph associated with matrix B has a spanning tree as well.

We build now the differential of g. Let $\tilde{g}$ be a smooth extension of (3.73) to $\mathbb{R}^{md} \times \mathbb{R}^{md}$, that is $g = \tilde{g}|_{\mathcal{S}}$. The differential $\mathrm{D}\tilde{g}(u)$ of $\tilde{g}$ at $u \in \mathcal{S}$ reads

$$\mathrm{D}\tilde{g}(u) = \begin{bmatrix} W_R - \alpha\nabla^2 F_c(x) & -\alpha I \\ (W_C - I)\,\nabla^2 F_c(x) & W_C \end{bmatrix}; \tag{3.76}$$

$\mathrm{D}\tilde{g}(u)$ is related to the differential of g by $\mathrm{D}g(u) = \mathrm{D}\tilde{g}(u)P_{\mathcal{T}(u)}$ [132], where $P_{\mathcal{T}(u)}$ is the orthogonal projector onto $\mathcal{T}(u)$. Using $\mathcal{T}(u) = \mathcal{S}$, for all $u \in \mathcal{S}$ (recall that $\mathcal{S}$ is a linear subspace) and denoting by $U_h \in \mathbb{R}^{md\times(m-1)d}$ an orthonormal basis of $\text{span}(W_C - I)$, $\mathrm{D}g(u)$ reads

$$\mathrm{D}g(u) = \begin{bmatrix} W_R - \alpha\nabla^2 F_c(x) & -\alpha I \\ (W_C - I)\,\nabla^2 F_c(x) & W_C \end{bmatrix} UU^\top, \quad \text{with} \quad U \triangleq \begin{bmatrix} I & 0 \\ 0 & U_h \end{bmatrix}. \tag{3.77}$$

Note that $P_{\mathcal{S}} = UU^\top$. We establish next the conditions for g to be a $\mathcal{C}^1$ diffeomorphism, as stated in Theorem 3.5.4.

Proposition 3.5.2. *Consider the mapping $g : \mathcal{S} \to \mathcal{S}$ defined in (3.73), under Assumptions 3.3.1-(i), 3.5.1, and 3.5.2, with $\mathcal{S}$ defined in (3.75). If the step-size is chosen according to*

$$0 < \alpha < \frac{\sigma_{\min}(CR)}{L_c}, \tag{3.78}$$

where $L_c = L_{\max}$, then $\det(\mathrm{D}g(u)) \neq 0$, for all $u \in \mathcal{S}$.

Proof. Since $\mathrm{D}g(u) : \mathcal{S} \to \mathcal{S}$, it is sufficient to verify that $\mathrm{D}g(u)$ is an invertible linear transformation for every $u \in \mathcal{S}$. Using the definition of U, this is equivalent to show that $U^T \mathrm{D}g(u) U$ is invertible, for all $u \in \mathcal{S}$. Invoking (3.77), $U^\top \mathrm{D}g(u) U$ reads

$$U^\top \mathrm{D}g(u) U = U^\top \mathrm{D}\tilde{g}(u) U = \begin{bmatrix} W_R - \alpha \nabla^2 F_c(x) & -\alpha U_h \\ U_h^\top (W_C - I) \nabla^2 F_c(x) & U_h^T W_C U_h \end{bmatrix}. \tag{3.79}$$

Since $U_h^\top W_C U_h$ is non-singular, we can use the Schur complement of $U^\top \mathrm{D}g(u) U$ with respect to $U_h^\top W_C U_h$ and write

$$U^\top \mathrm{D}g(u) U = S_1 \begin{bmatrix} W_R - \alpha \nabla^2 F_c(x) + \alpha \Phi (W_C - I) \nabla^2 F_c(x) & 0 \\ 0 & U_h^\top W_C U_h \end{bmatrix} S_2, \tag{3.80}$$

where $\Phi \triangleq U_h \left(U_h^\top W_C U_h\right)^{-1} U_h^\top$, and S_1 and S_2 are some nonsingular matrices. By (3.80), it is sufficient to show that

$$\begin{aligned} S \triangleq & W_R - \alpha \nabla^2 F_c(x) + \alpha \Phi (W_C - I) \nabla^2 F_c(x) \\ = & W_R - \alpha W_C^{-1} \nabla^2 F_c(x) + \alpha \left(\Phi - W_C^{-1}\right) (W_C - I) \nabla^2 F_c(x) \end{aligned} \tag{3.81}$$

is non-singular. Using $W_C - I = U_h \Delta$, for some $\Delta \in \mathbb{R}^{(m-1)d \times md}$ (recall that U_h is an orthonormal basis of $\mathrm{span}(W_C - I)$), we can write

$$\begin{aligned} \Phi = & U_h \left(U_h^\top W_C U_h\right)^{-1} U_h^\top = U_h (I + \Delta U_h)^{-1} U_h^\top \\ \overset{(a)}{=} & U_h U_h^\top - U_h \Delta (I + U_h \Delta)^{-1} U_h U_h^\top \\ = & U_h U_h^\top - (W_C - I) W_C^{-1} U_h U_h^\top = W_C^{-1} U_h U_h^\top, \end{aligned} \tag{3.82}$$

where (a) we used the Woodbury identity of inverse matrices. Using (3.82) in (3.81), we obtain

$$S = W_R - \alpha W_C^{-1}\nabla^2 F_c(x) - \alpha W_C^{-1}\underbrace{\left(I - U_h U_h^\top\right)(W_C - I)}_{=0}\nabla^2 F_c(x).$$

Therefore, if $\alpha < \frac{\sigma_{\min}(CR)}{L_c}$, S is invertible, and consequently, so is $U^\top \mathrm{D}g(u)U$. □

2) The consensual strict saddle points are unstable fixed points of g ($\mathcal{U}^* \subseteq \mathcal{A}_g$): First of all, note that every limit point of the sequence generated by (3.71) is a fixed point of g on $\mathcal{S}$; the converse might not be true. The next result establishes the desired connection between the set $\mathcal{A}_g$ of unstable fixed points of g (cf. Definition 3.5.2) and the set $\mathcal{U}^*$ of consensual strict saddle points (cf. Definition 3.5.3). This will let us infer the instability of $\mathcal{U}^*$ from that of $\mathcal{A}_g$.

Proposition 3.5.3. *Suppose that Assumptions 3.3.1-(i) and 3.5.1 hold along with one of the following two conditions*

(i) The weight matrices R and C are symmetric;

(ii) $d = 1$.

Then, any consensual strict saddle point is an unstable fixed point of g, i.e.,

$$\mathcal{U}^* \subseteq \mathcal{A}_g, \tag{3.83}$$

with $\mathcal{A}_g$ and $\mathcal{U}^*$ defined in (3.70) and (3.72), respectively.

Proof. Let $u^* \in \mathcal{U}^*$; u^* is a fixed point of g defined in (3.73). It is thus sufficient to show that $\mathrm{D}g(u^*)$ has an eigenvalue with magnitude greater than one.

To do so, we begin showing that the differential $\mathrm{D}\tilde{g}(u^*)$ of $\tilde{g}$ at u^* has an eigenvalue greater than one. Using (3.76), $\mathrm{D}\tilde{g}(u^*)$ reads

$$\mathrm{D}\tilde{g}(u^*) = \begin{bmatrix} W_R - \alpha\nabla^2 F_c^* & -\alpha I \\ (W_C - I)\,\nabla^2 F_c^\star & W_C \end{bmatrix}, \tag{3.84}$$

where we defined the shorthand $\nabla^2 F_c^* \triangleq \nabla^2 F_c\,(1 \otimes \theta^*)$, and $\theta^* \in \Theta_{ss}^*$. We need to prove

$$\det\left(\mathrm{D}\tilde{g}(u^*) - \lambda_u I\right) = 0, \qquad \text{for some} \qquad |\lambda_u| > 1. \tag{3.85}$$

If $|\lambda_u| > 1$, $W_C - \lambda_u I$ is nonsingular (since spradii$(C) = 1$). Using the Schur complement of $\mathrm{D}\tilde{g}(u^*) - \lambda_u I$ with respect to $W_C - \lambda_u I$, we have

$$\mathrm{D}\tilde{g}(u^*) - \lambda_u I = \tilde{S}_1 \begin{bmatrix} \left(\mathrm{D}\tilde{g}(u^*) - \lambda_u I\right) / \left(W_C - \lambda_u I\right) & 0 \\ 0 & W_C - \lambda_u I \end{bmatrix} \tilde{S}_2, \tag{3.86}$$

for some $\tilde{S}_1, \tilde{S}_2 \in \mathcal{M}_{2md}(\mathbb{R})$, with $\det(\tilde{S}_1) = \det(\tilde{S}_2) = 1$. Given (3.86), (3.85) holds if and only if

$$\det \begin{bmatrix} W_R - \lambda_u I - \alpha \nabla^2 F_c^\star + \alpha \left(W_C - \lambda_u I\right)^{-1} \left(W_C - I\right) \nabla^2 F_c^* & 0 \\ 0 & W_C - \lambda_u I \end{bmatrix} = 0,$$

or equivalently

$$\det\left(W_R - \lambda_u I - \alpha \nabla^2 F_c^* + \alpha \left(W_C - \lambda_u I\right)^{-1} \left(W_C - I\right) \nabla^2 F_c^*\right) = 0. \tag{3.87}$$

Multiplying both sides of (3.87) by $\det(W_C - \lambda_u I)$ yields

$$Q(\lambda_u) \triangleq \det\Bigg(\underbrace{\left(W_C - \lambda_u I\right)\left(W_R - \lambda_u I\right) + \alpha(\lambda_u - 1)\nabla^2 F_c^*}_{\triangleq T(\lambda_u)}\Bigg) = 0. \tag{3.88}$$

Trivially $Q(\lambda_u) > 0$, if $\lambda_u \gg 1$. Therefore, to show that (3.85) holds, it is sufficient to prove that there exists some $\lambda_u > 1$ such that $Q(\lambda_u) \leq 0$. Next, we prove this result under either condition (i) or (ii).

Suppose (i) holds; R and C are symmetric. Define $\tilde{\boldsymbol{v}} \triangleq 1 \otimes \boldsymbol{v}$, where $\boldsymbol{v}$ is the unitary eigenvector associated with a negative eigenvalue of $\nabla^2 F(\theta^*)$, and let $\lambda_{\min}(\nabla^2 F(\theta^*)) = -\delta$; we can write

$$\tilde{\boldsymbol{v}}^\top T(\lambda_u) \tilde{\boldsymbol{v}} = m(\lambda_u - 1)\left(\lambda_u - 1 - \alpha\delta/m\right) < 0, \tag{3.89}$$

for all $1 < \lambda_u < 1 + \alpha\delta/m$. By Rayleigh-Ritz theorem, $T(\lambda_u)$ has a negative eigenvalue, implying that there exists some real value $\bar{\lambda}_u > 1$ such that $Q(\bar{\lambda}_u) = 0$.

Suppose now that conditions (ii) holds; W_R and W_C reduce to R and C, respectively. Note that R and C are now not symmetric. Let $\lambda_u = 1+\epsilon$, and consider the Taylor expansion of

$$Q(1+\epsilon) = \det\Big((C-I)(R-I) + \epsilon\left(\alpha\nabla^2 F_c^* + 2I - C - R\right) + \epsilon^2 I\Big), \tag{3.90}$$

around $\epsilon = 0$. Define $M \triangleq (C-I)(R-I)$ and $N \triangleq \alpha\nabla^2 F_c^* + 2I - C - R$. It is clear that $Q(1) = 0$; then, by the Jacobi's formula, we have

$$Q(1+\epsilon) = \mathrm{tr}\Big(\mathrm{adj}\,(M)\,N\Big)\epsilon + O(\epsilon^2). \tag{3.91}$$

Expanding (3.91) yields

$$\begin{aligned} Q(1+\epsilon) =& \mathrm{tr}\Big(\mathrm{adj}\,(R-I)\,\mathrm{adj}\,(C-I)\,N\Big)\epsilon + O(\epsilon^2) \\ =& \mathrm{tr}\Big(1\tilde{r}^\top\tilde{c}1^\top N\Big)\epsilon + O(\epsilon^2) = (\tilde{r}^\top\tilde{c})1^\top N 1\epsilon + O(\epsilon^2), \end{aligned} \tag{3.92}$$

where $\tilde{r}$ and $\tilde{c}$ are the Perron vectors of R and C, respectively. The second equality in (3.92) is due to the following fact: a rank-$(m-1)$ matrix $A \in \mathcal{M}_m(\mathbb{R})$ has rank-1 adjugate matrix $\mathrm{adj}\,(A) = ab^\top$, where a and b are non-zero vectors belonging to the 1-dimensional null space of A and $A^\top$, respectively [133, Sec. 0.8.2]. We also have $\tilde{\zeta} \triangleq \tilde{r}^\top\tilde{c} > 0$, due to Lemma 3.5.1. Furthermore, since $\theta^* \in \Theta_{ss}^*$, $1^\top\nabla^2 F_c^* 1 \leq -\delta$, for some $\delta > 0$, and

$$Q(1+\epsilon) \leq -\delta\tilde{\zeta}\alpha\epsilon + O(\epsilon^2), \tag{3.93}$$

which implies the existence of a sufficiently small $\epsilon > 0$ such that $Q(1+\epsilon) < 0$. Consequently, there must exist some $\bar{\lambda}_u > 1$ such that (3.85) holds. Moreover, such $\bar{\lambda}_u$ is a real eigenvalue of $\mathrm{D}\tilde{g}(u^*)$.

To summarize, we proved that there exists an eigenpair $(\bar{\lambda}_u, v_u)$ of $\mathrm{D}\tilde{g}(u^*)$, with $\bar{\lambda}_u > 1$. Next we show that $(\bar{\lambda}_u, v_u)$ is also an eigenpair of $\mathrm{D}g(u^*)$. Let us partition $v_u \triangleq (v_u^x, v_u^h)$ such that

$$\begin{bmatrix} W_R - \alpha\nabla^2 F_c\left(x^*\right) & -\alpha I \\ \left(W_C - I\right)\nabla^2 F_c\left(x^*\right) & W_C \end{bmatrix} \begin{bmatrix} v_u^x \\ v_u^h \end{bmatrix} = \bar{\lambda}_u \begin{bmatrix} v_u^x \\ v_u^h \end{bmatrix}. \tag{3.94}$$

In particular, we have $\left(W_C - I\right)\left(\nabla^2 F_c\left(x^*\right) v_u^x + v_u^h\right) = (\bar{\lambda}_u - 1)v_u^h$, which implies $v_u^h \in \mathrm{span}(W_C - I)$, since $\bar{\lambda}_u - 1 \neq 0$. Therefore, $v_u \in \mathcal{S}$.

Now, let $P_{\mathcal{S}}$ be the orthogonal projection matrix onto $\mathcal{S}$. Since $v_u \in \mathcal{S}$, we have

$$\mathrm{D}\tilde{g}(u^*)v_u = \bar{\lambda}_u v_u \implies \mathrm{D}\tilde{g}(u^*)P_{\mathcal{S}}^\top v_u = \bar{\lambda}_u v_u \overset{(a)}{\implies} \mathrm{D}g(u^*)v_u = \bar{\lambda}_u v_u, \tag{3.95}$$

where (a) is due to $\mathrm{D}g(u^*) = \mathrm{D}\tilde{g}(u^*)P_{\mathcal{S}}^\top$ [cf. (3.77)]. Hence $(\bar{\lambda}_u, v_u)$ is also an eigenpair of $\mathrm{D}g(u^*)$, which completes the proof. □

Remark 3.5.1. *Note that condition (i) in Proposition 3.5.3 implies that $\mathcal{G}_C$ and $\mathcal{G}_R$ are undirected graphs. Condition (ii) extends the network model to directed topologies under assumption $d = 1$. For sake of completeness, we relax condition (ii) in Appendix 3.8.4 to arbitrary $d \in \mathbb{N}$, under extra (albeit mild) assumptions on the set of strict saddle points and the weight matrices R and C.*

3.5.3.3 DOGT likely converges to SoS solutions of (3.1)

Combining Theorem 3.5.4, Proposition 3.5.2, and Proposition 3.5.3, we can readily obtain the following second-order guarantees of the DOGT algorithms.

Theorem 3.5.5. *Consider Problem (3.1), under Assumptions 3.3.1 and 3.3.4; and let $\{u^\nu \triangleq (x^\nu, h^\nu)\}$ be the sequence generated by the DOGT Algorithm (3.71) under the following tuning: i) the step-size α satisfies (3.49) [or (3.52)] and (3.78); the weight matrices C and R are chosen according to Assumptions 3.5.1 and 3.5.2; and the initialization is set to $u^0 \in \mathcal{S}$, with $\mathcal{S}$ defined in (3.75). Furthermore, suppose that either (i) or (ii) in Proposition 3.5.3 holds. Then, we have*

$$\mathbb{P}_{u^0}\left(\lim_{\nu\to\infty} u^\nu \in \mathcal{U}^*\right) = 0, \tag{3.96}$$

where the probability is taken over $u^0 \in \mathcal{S}$.

In addition, if F *is a KŁ function, then* $\{x^\nu\}$ *converges almost surely to* $1 \otimes \theta^\infty$ *at a rate determined in Theorem 3.5.3, where* θ^∞ *is a SoS solution of* (3.1).

Note that (3.96) implies the desired second-order guarantees only when the sequence $\{u^\nu\}$ convergences [i.e., the limit in (3.96) exists]; otherwise (3.96) is trivially satisfied, and some limit point of $\{u^\nu\}$ can belong to $\mathcal{U}^*$ with non-zero probability. A sufficient condition for the required global convergence of $\{u^\nu\}$ is that F is a KŁ function, which is stated in the second part of the above theorem.

Remark 3.5.2 (Comparison with [92])**.** *As already discussed in Sec. 3.1.2, the primal-dual methods in [92] is applicable to* (3.1)*; it is proved to almost surely converge to SoS solutions. Convergence of [92] is proved under stricter conditions on the problem than DOGT, namely: i) the network must be undirected; and ii) the Hessian of each local* f_i *must be Lipschitz continuous. It does* not *seem possible to extend the analysis of [92] beyond this assumptions.*

3.6 Numerical Results

In this section we test the behavior of DGD and DOGT around strict saddles on three classes of nonconvex problems, namely: i) a quadratic function (cf. Sec. 3.6.1); ii) a classification problem based on the cross-entropy risk function using sigmoid activation functions (cf. Sec. 3.6.2); and iii) a two Gaussian mixture model (cf. Sec. 3.6.3).

3.6.1 Nonconvex quadratic optimization

Consider

$$\min_{\theta \in \mathbb{R}^d} F(\theta) = \frac{1}{2} \sum_{\mathrm{i}=1}^{m} (\theta - b_\mathrm{i})^\top Q_\mathrm{i} (\theta - b_\mathrm{i}), \tag{3.97}$$

where $d = 20$; $m = 10$; b_i's are i.i.d Gaussian zero mean random vectors with standard deviation 10^3; and Q_i's are $d \times d$ randomly generated symmetric matrices where $\sum_{\mathrm{i}=1}^m Q_\mathrm{i}$ has $d-1$ eigenvalues $\{\lambda_\mathrm{i}\}_{\mathrm{i}=1}^{d-1}$ uniformly distributed over $(0, m]$, and one negative eigenvalue $\lambda_d = -m\delta$, with $\delta = 0.01$. Clearly (3.97) is an instance of Problem (3.1), with F having a unique strict saddle point $\theta^* = \left(\sum_\mathrm{i} Q_\mathrm{i}\right)^{-1} \sum_\mathrm{i} Q_\mathrm{i} b_\mathrm{i}$. The network of m agents is modeled as a

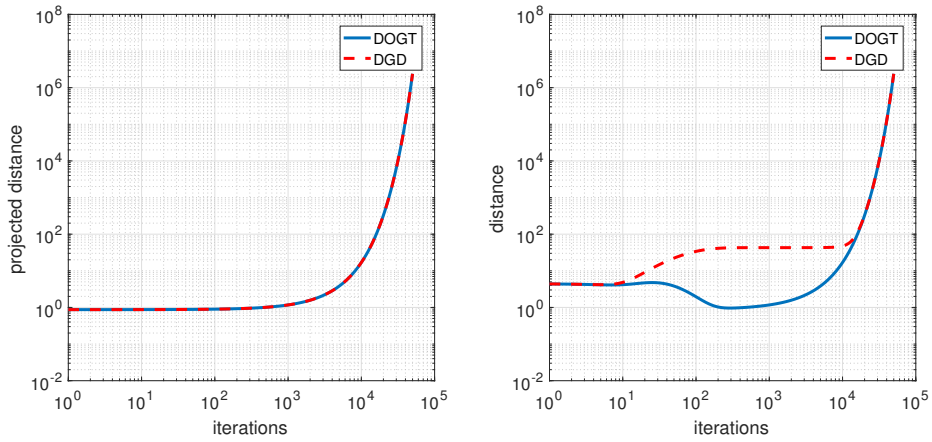


Figure 3.1. Escaping properties of DGD and DOGT, applied to Problem (3.97). Left plot: distance of the average iterates from θ^* projected onto the unstable manifold E_u versus the number of iterations. Right plot: distance of the average iterates from θ^* versus the number of iterations.

ring; the weight matrix $W \triangleq \{w_{\mathrm{ij}}\}_{\mathrm{i,j}=1}^{m}$, compliant to the graph topology, is generated to be doubly stochastic.

To test the escaping properties of DGD and DOGT from the strict saddle of F, we initialize the algorithms in a randomly generated neighborhood of θ^*. More specifically, every agent's initial point is $x_{\mathrm{i}}^0 = \theta^* + \delta_{x,\mathrm{i}}$, $\mathrm{i} \in [m]$. In addition, for the DOGT algorithm, we set $y_{\mathrm{i}}^0 = \nabla f_{\mathrm{i}}(x_{\mathrm{i}}^0) + (w_{\mathrm{ii}} - 1)\delta_{y,\mathrm{i}} + \sum_{\mathrm{j}\neq\mathrm{i}} w_{\mathrm{ij}}\delta_{y,\mathrm{j}}$, where $\delta_{x,\mathrm{i}}$'s and $\delta_{y,\mathrm{i}}$'s are realizations of i.i.d. Gaussian random vectors with standard deviation equal to 1. Both algorithms use the same step-size $\alpha = 0.99\,\sigma_{\min}(I+W)/L_c$, with $L_c = \max_{\mathrm{i}}\{|\lambda_{\mathrm{i}}|\}$; this is the largest theoretical step-size guaranteeing convergence of the DGD algorithm (cf. Theorem 3.4.1).

In the left panel of Fig. 3.1, we plot the distance of the average iterates $\bar{x}^\nu = (1/m)\sum_{\mathrm{i}=1}^{m} x_{\mathrm{i}}^\nu$ from the critical point θ^* projected on the unstable manifold $E_u = \mathrm{span}(u^u)$, where u^u is the eigenvector associated with the negative eigenvalue $\lambda_d = -m\delta$. In the right panel, we plot $\|\bar{x}^\nu - \theta^*\|$ versus the number of iterations. All the curves are averaged over 50 independent initializations. Figure in the left panel shows that, as predicted by our theory, both algorithms almost surely escapes from the unstable subspace E_u, at an indistinguishable practical rate. The right panel shows that DOGT gets closer to the strict saddle; this can

be justified by the fact that, differently from DGD, DOGT exhibits *exact* convergence to critical points.

3.6.2 Bilinear logistic regression

Consider a classification problem with distributed training data set $\{s_{\rm i}, \xi_{\rm i}\}_{{\rm i}=1}^m$, where $s_{\rm i} \in \mathbb{R}^d$ is the feature vector associated with the binary class label $\xi_{\rm i} \in \{0, 1\}$. The bilinear logistic regression problem [121] aims at finding the bilinear classifier $\zeta_{\rm i}(Q, w; s_{\rm i}) = s_{\rm i}^\top Q w$, with $Q \in \mathbb{R}^{d\times p}$ and $w \in \mathbb{R}^p$ that best separates data with distinct labels. Let $(s_{\rm i}, \xi_{\rm i})$ be private information for agent i. Using the sigmoid activation function $\sigma(x) \triangleq 1/(1+{\rm e}^{-x})$ together with the *cross-entropy risk* function, the optimization problem reads

$$\min_{Q,w} \quad -\frac{1}{m}\sum_{{\rm i}=1}^m \left[\xi_{\rm i} \ln\left(\sigma(s_{\rm i}^\top Q w)\right) + (1-\xi_{\rm i}) \ln\left(1-\sigma(s_{\rm i}^\top Q w)\right)\right] + \frac{\tau}{2}\left(\left\|Q\right\|_F^2 + \left\|w\right\|^2\right). \quad (3.98)$$

It is not difficult to show that (3.98) is equivalent to the following instance of (3.1):

$$\min_{Q,w} \quad F(Q,w) = \sum_{{\rm i}=1}^m \underbrace{\frac{1}{m}\left[-\ln\left(\sigma(\tilde{\xi}_{\rm i} s_{\rm i}^\top Q w)\right) + \frac{\tau}{2}\left(\left\|Q\right\|_F^2 + \left\|w\right\|^2\right)\right]}_{=f_{\rm i}(Q,w)}, \quad (3.99)$$

with

$$\tilde{\xi}_{\rm i} \triangleq \begin{cases} -1, & \text{if} \quad \xi_{\rm i} = 0; \\ 1, & \text{if} \quad \xi_{\rm i} = 1. \end{cases}$$

To visualize the landscape of $F(Q, w)$ (2D plot), we consider the following setting for the free parameters. We set $d = p = 1$; $\tau = 0.2$; $m = 5$; and we generate uniformly random $\tilde{\xi}_{\rm i} \in \{0, 1\}$, and we draw $s_{\rm i}$ from a normal distribution with mean $\xi_{\rm i}$ and variance 1. The gradient of the local loss $f_{\rm i}$ reads

$$\begin{bmatrix} \nabla_Q f_{\rm i}(Q,w) \\ \nabla_w f_{\rm i}(Q,w) \end{bmatrix} = \frac{1}{m}\begin{bmatrix} \tau Q - \tilde{\xi}_{\rm i} s_{\rm i} w \sigma(-\tilde{\xi}_{\rm i} s_{\rm i} Q w) \\ \tau w - \tilde{\xi}_{\rm i} s_{\rm i} Q \sigma(-\tilde{\xi}_{\rm i} s_{\rm i} Q w) \end{bmatrix}.$$

A surface plot of $F(Q, w)$ in the above setting is plotted in the right panel of Fig. 3.2. Note that such F has three critical points, two of which are local minima (see the location of

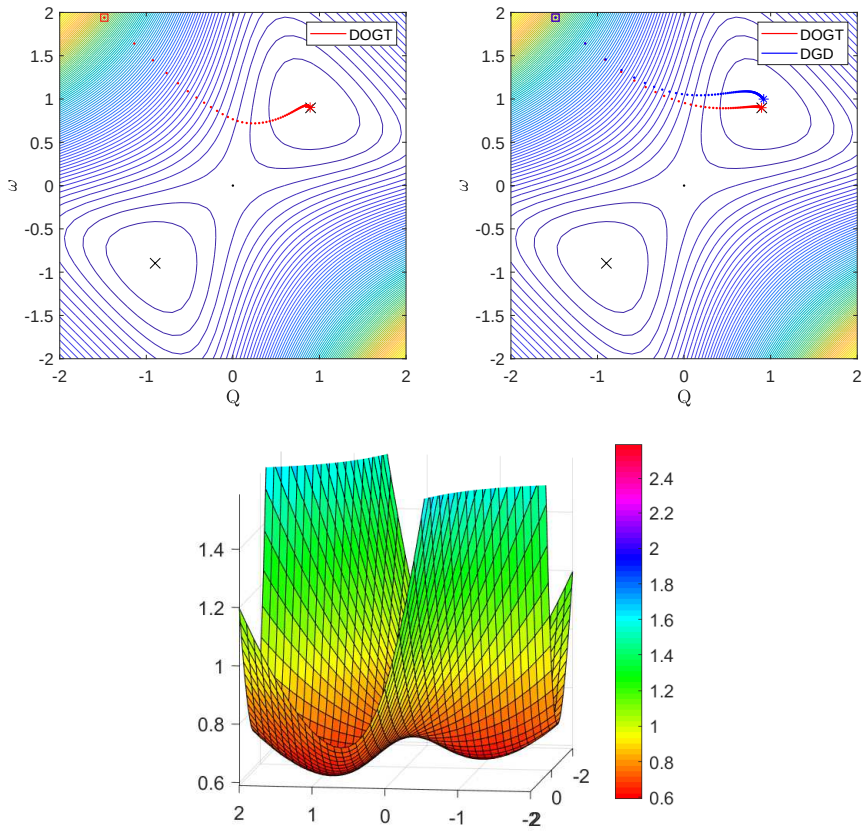


Figure 3.2. Escaping properties of the DGD and DOGT, applied to the bilinear logistic regression problem (3.98). Top left (resp. top right) plot: directed (resp. undirected) network; trajectory of the average iterates on the contour of F ($(0,0)$ is the strict saddle point and $\times$ are the local minima); DGD and DOGT are initialized at $\square$ and terminated after 100 iterations at $*$. Bottom plot: plot of F.

minima in the left or middle panel of Fig. 3.2 marked by $\times$) and one strict saddle point at $(0,0)$–the Hessian at $(0,0)$,

$$\nabla^2 F(0,0) = \begin{bmatrix} \tau & -\frac{1}{2m}\sum_{\mathrm{i}} \tilde{\xi}_{\mathrm{i}} s_{\mathrm{i}} \\ -\frac{1}{2m}\sum_{\mathrm{i}} \tilde{\xi}_{\mathrm{i}} s_{\mathrm{i}} & \tau \end{bmatrix},$$

has an eigenvalue at $\tau - \frac{1}{2m}\sum_{\mathrm{i}} \tilde{\xi}_{\mathrm{i}} s_{\mathrm{i}} = -0.26$.

We test DGD and DOGT over a network of $m = 5$ agents; for DGD we considered undirected graphs whereas we run DOGT on both undirected and directed graphs. Both

algorithms are initialized at the same random point and terminated after 100 iterations; the step-size is set to $\alpha = 0.9$. We denote by Q_{i}^{ν} and w_{i}^{ν} the agent i's ν-the iterate of the local copies of Q and w, respectively. The trajectories of the average iterates $(\bar{Q}^{\nu}, \bar{w}^{\nu}) \triangleq \frac{1}{m}(\sum_{\mathrm{i}} Q_{\mathrm{i}}^{\nu}, \sum_{\mathrm{i}} w_{\mathrm{i}}^{\nu})$ are plotted in Fig. 3.2; the left panel refers to the directed graph while the middle panel reports the same results for the undirected network. As expected, the DOGT algorithm converges to an exact critical point (local minimum) avoiding the strict saddle $(0, 0)$ while DGD converges to a neighborhood of the local minimum. The consensus error is $1/m\sqrt{\sum_{\mathrm{i}=1}^{m} ||(Q_{\mathrm{i}}^{\nu}, w_{\mathrm{i}}^{\nu}) - (\bar{Q}^{\nu}, \bar{w}^{\nu})||^2}$; at the termination, it reads 2.33×10^{-4} for DOGT over the directed network, and 2.18×10^{-4} and 9.74×10^{-2} for DOGT and DGD, respectively, over undirected networks.

3.6.3 Gaussian mixture model

Consider the Gaussian mixture model defined in Sec. 3.3. The data $\{z_{\mathrm{i}}\}_{\mathrm{i}=1}^{m}$ where $z_{\mathrm{i}} \in \mathbb{R}^d$ are realizations of the mixture model $z_{\mathrm{i}} \sim \frac{1}{2}\mathcal{N}(\mu_1, \Sigma_1) + \frac{1}{2}\mathcal{N}(\mu_2, \Sigma_2)$. Let each agent i own z_{i}. Both parameters (μ_1, μ_2) and (Σ_1, Σ_2) are unknown. The goal is to approximate (μ_1, μ_2) while (Σ_1, Σ_2) is set to an estimate $(\tilde{\Sigma}, \tilde{\Sigma})$. The problem reads

$$\min_{\theta_1, \theta_2 \in \mathbb{R}^d} \; -\sum_{\mathrm{i}=1}^{m} \log\left(\phi_d(z_{\mathrm{i}} - \theta_1) + \phi_d(z_{\mathrm{i}} - \theta_2)\right), \tag{3.100}$$

where $\phi_d(\theta)$ is the d-dimensional normal distribution with mean 0 and covariance $\tilde{\Sigma}$. Consider the case of mixture of two scalar Gaussians, i.e., $d = 1$. We draw $\{z_{\mathrm{i}}\}_{\mathrm{i}=1}^{5}$ from the the this model, with means $\mu_1 = 0$, $\mu_2 = -5$ and variance $\sigma_1 = \sigma_2 = 25$. The estimate of variance in problem (3.100) is pessimistically set to $\tilde{\sigma} = 1$. A surface plot of a random instance of above problem is depicted in right panel of Fig. 3.3. Note that this instance of problem has 2 global minima (marked by $\times$) and multiple local minima. We test DGD and DOGT on the above problem over the same networks as described in Sec. 3.6.2. Both algorithms are initialized at the same random point and terminated after 250 iterations; the step-size is set to $\alpha = 0.1$. In Fig. 3.3, we plot the trajectories of the average iterates $(\bar{\theta}_1^{\nu}, \bar{\theta}_2^{\nu}) \triangleq \frac{1}{m}(\sum_{\mathrm{i}} \theta_{1,\mathrm{i}}^{\nu}, \sum_{\mathrm{i}} \theta_{2,\mathrm{i}}^{\nu})$, where $\theta_{1,\mathrm{i}}^{\nu}$ and $\theta_{2,\mathrm{i}}^{\nu}$ are the agent i's ν-the iterate of the local copies of θ_1 and θ_2, respective; the left

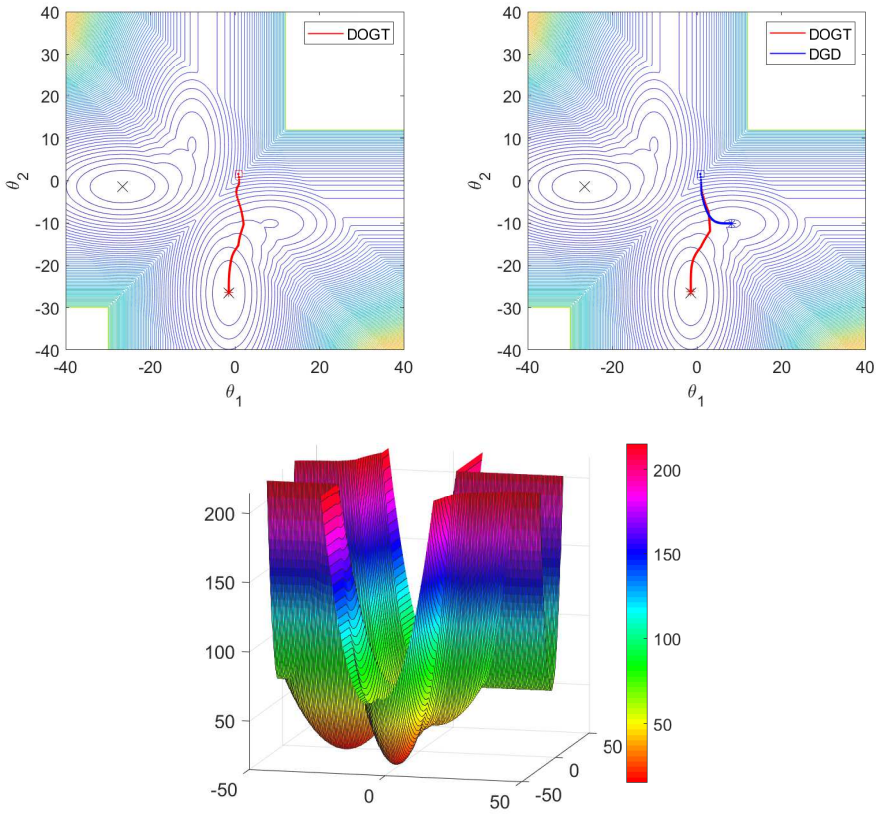


Figure 3.3. Escaping properties of the DGD and DOGT applied to the Gaussian mixture problem (3.100). Top left (resp. top right) plot: directed (resp. undirected) network; trajectory of the average iterates on the contour of F (the global minima are marked by $\times$); DGD and DOGT are initialized at $\square$ and terminated after 250 iterations at $*$. Bottom plot: plot of F.

(resp. middle) panel refers to the undirected (resp. directed) network. DOGT converges to the global minimum while DGD happens to converge to neighborhood of a local minima. The consensus error is measured by $(1/m)\sqrt{\sum_{\mathrm{i}=1}^{m} ||(\theta_{1,\mathrm{i}}^{\nu}, \theta_{2,\mathrm{i}}^{\nu}) - (\bar{\theta}_1^{\nu}, \bar{\theta}_2^{\nu})||^2}$ and at the termination it is equal to 1.9×10^{-3} for DOGT on the directed graph; and 2.8×10^{-3} and 1.135 for DOGT and DGD, respectively over the undirected graph.

3.7 Conclusions

We proposed the first second-order distributed algorithm for convex and strongly convex problems over meshed networks with *global* communication complexity bounds which, up to the network dependent factor $\widetilde{\mathcal{O}}(1/\sqrt{1-\rho})$, (almost) match the iteration complexity of

centralized second-order method [134] in the regime when the desired accuracy is moderate. We showed that this regime is reasonable when one considers ERM problems for which there is no need to optimize beyond the statistical error. Importantly, our method avoids expensive communications of Hessians over the network and keeps the amount of information sent in each communication round similar to first-order methods.

This work is just a starting point towards a theory of second-order methods with performance guarantees on meshed networks under statistical similarity; many questions remain open. An obvious one is incorporating acceleration to improve communication complexity bounds under statistical similarity. A first attempt towards this goal is the follow-up work [135], where an accelerated second-order method exploiting statistical similarity has been analyzed for master/workers architectures. The extension to arbitrary graphs remains an open problem. Second, our main goal here has been decreasing communications, which does not guarantee optimal oracle (computational) complexity–this is because we did not take advantage of the finite-sum structure of the *local* optimization problems. Stochastic optimization algorithms equipped with Variance Reduction (VR) techniques have been proved to be quite effective to obtain cheaper iterations while preserving fast convergence [136], [137]. However, these methods do not exploit any statistical similarity, resulting in less favorable communication complexity whenever $\beta/\mu \ll Q/\mu$. It would be then interesting to investigate whether VR techniques can improve both communication and oracle complexity when statistical similarity is explicitly employed in the algorithmic design.

3.8 Appendix

3.8.1 On the problems satisfying Assumption 3.3.3

We prove that all the functions arising from the examples in Sec. 3.3 satisfy Assumption 3.3.3, for sufficiently large R and $R-\epsilon$. To do so, for each function, we establish lowerbounds implying $\langle \nabla f_i(\theta), \theta/\left\|\theta\right\| \rangle \to \infty$ as $||\theta|| \to \infty$.

a) Distributed PCA: Let us expand the objective function in (3.9) as

$$\begin{aligned} F(\theta) =& \frac{1}{4}\mathrm{tr}\left(\theta\theta^\top\theta\theta^\top\right) - \frac{1}{2}\mathrm{tr}\left(\theta^\top\sum_{\mathrm{i}=1}^{m} M_\mathrm{i}\theta\right) + \frac{1}{4}\mathrm{tr}\left(\sum_{\mathrm{i}=1}^{m} M_\mathrm{i}^\top \sum_{\mathrm{i}=1}^{m} M_\mathrm{i}\right) \\ =& \sum_{\mathrm{i}=1}^{m} \underbrace{\frac{1}{4}\left\{\frac{1}{m}\left\|\theta\theta^\top\right\|_F^2 - 2\mathrm{tr}\left(\theta^\top M_\mathrm{i}\theta\right)\right\}}_{\triangleq f_\mathrm{i}(\theta)} + \frac{1}{4}\mathrm{tr}\left(\sum_{\mathrm{i}=1}^{m} M_\mathrm{i}^\top \sum_{\mathrm{i}=1}^{m} M_\mathrm{i}\right). \end{aligned}$$

We have

$$\begin{aligned} \left\langle \nabla f_\mathrm{i}(\theta), \theta/\left\|\theta\right\| \right\rangle =& \left\langle \frac{1}{m}\theta\theta^\top\theta - M_\mathrm{i}\theta, \theta \right\rangle / \left\|\theta\right\| \\ =& \frac{1}{m}\left\|\theta\theta^\top\right\|_F^2/\left\|\theta\right\| - \theta^\top M_\mathrm{i}\theta/\left\|\theta\right\| \\ \geq& \frac{1}{mK_{2,4}^4}\left\|\theta\right\|^3 - \sigma_{\max}(M_\mathrm{i})\left\|\theta\right\|, \end{aligned}$$

for some $K_{2,4} > 0$, where in the last inequality we used the equivalence of ℓ_4 and ℓ_2 norms, i.e. $||\theta||_2 \leq K_{2,4}||\theta||_4, \forall\theta \in \mathbb{R}^d$.

b) Phase retrieval: It is not difficult to show that for the objective function in (3.10), it holds

$$\begin{aligned} \left\langle \nabla f_\mathrm{i}(\theta), \theta/\left\|\theta\right\| \right\rangle =& \left(||a_\mathrm{i}^\top\theta||^2 - y_\mathrm{i}\right)||a_\mathrm{i}^\top\theta||^2/\left\|\theta\right\| + \lambda\left\|\theta\right\| \\ =& \left(||a_\mathrm{i}^\top\theta||^2 - y_\mathrm{i}/2\right)^2/\left\|\theta\right\| - \frac{y_\mathrm{i}^2}{4\left\|\theta\right\|} + \lambda\left\|\theta\right\|. \end{aligned}$$

c) Matrix sensing: Consider the objective function in (3.10). It is not difficult to show that

$$\begin{aligned} \left\langle \nabla f_\mathrm{i}(\Theta), \Theta/\left\|\Theta\right\|_F \right\rangle =& \left(\mathrm{tr}\left(\Theta^\top A_\mathrm{i}\Theta\right) - y_\mathrm{i}\right)\mathrm{tr}\left(\Theta^\top A_\mathrm{i}\Theta\right)/\left\|\Theta\right\|_F + \lambda\left\|\Theta\right\|_F \\ =& \mathrm{tr}\left(\Theta^\top A_\mathrm{i}\Theta\right)^2/\left\|\Theta\right\|_F - y_\mathrm{i}\mathrm{tr}\left(\Theta^\top A_\mathrm{i}\Theta\right)/\left\|\Theta\right\|_F + \lambda\left\|\Theta\right\|_F \\ =& \left(\mathrm{tr}\left(\Theta^\top A_\mathrm{i}\Theta\right) - y_\mathrm{i}/2\right)^2/\left\|\Theta\right\|_F - \frac{y_\mathrm{i}^2}{4\left\|\Theta\right\|_F} + \lambda\left\|\Theta\right\|_F. \end{aligned}$$

d-f) We prove the property only for the Gaussian mixture model; similar proof applies also to the other classes of problems. Denote $\theta = (\theta_d)_{d=1}^q$. Since ϕ_d is a bounded function, we have

$$\langle \nabla_{\theta_d} f_\mathrm{i}(\theta_d), \theta_d \rangle \geq -C_d + \lambda\left\|\theta_d\right\|^2,$$

for some $C_d > 0$. Hence, $\left\langle \nabla_\theta f_\mathrm{i}(\theta), \theta/\left\|\theta\right\| \right\rangle \geq -C/\left\|\theta\right\| + \lambda\left\|\theta\right\|$, with $C = \sum_d C_d$.

3.8.2 Convergence of DGD without $L-$smoothness of f_{i}'s

We sketch here how to extend the convergence results of DGD stated in Sec. 3.4 to the case when the gradient of the agents' loss functions is not globally Lipschitz continuous (i.e. removing Assumption 3.3.1 (i)). Due to the space limitation, we prove only the counterpart of Theorem 3.4.1; the other results in Sec. 3.4 can be extended following similar arguments.

We begin introducing some definitions. Under Assumptions 3.3.3 and 3.4.1, define the set $\tilde{\mathcal{Y}} \triangleq \mathcal{Y} + \mathcal{B}_b^{md}$, with $\mathcal{Y} = \bar{\mathcal{L}} \cup \prod_{\mathrm{i}=1}^m \mathcal{B}_R^d$ and

$$\bar{\mathcal{L}} = \mathcal{L}_{F_c}\left(\max_{x_{\mathrm{i}}^0 \in \mathcal{B}_R^d, \mathrm{i}\in[m]} \left\{\sum_{\mathrm{i}=1}^m f_{\mathrm{i}}(x_{\mathrm{i}}^0)\right\} + \frac{R^2}{\alpha_b}\right), \tag{3.101}$$

where

$$\begin{aligned} \alpha_b &= \min_{\mathrm{i}\in[m]} \min\{\epsilon D_{\mathrm{ii}}/h, 2D_{\mathrm{ii}}\delta(R-\epsilon)/h^2\} > 0, \\ h = \max_{\mathrm{i}\in[m], z\in\mathcal{B}_R^d} ||\nabla f_{\mathrm{i}}(z)||, &\quad \text{and} \quad b = \max_{\alpha\in[\alpha_b,1], \theta\in\mathcal{Y}} ||\nabla L_\alpha(\theta)||. \end{aligned} \tag{3.102}$$

Note that, under Assumption 2.1'(ii), $\mathcal{Y}$ and $\tilde{\mathcal{Y}}$ are compact. Hence, ∇F_c is globally Lipschitz on $\tilde{\mathcal{Y}}$, and so is ∇L_α; we denote such Lipschitz constants as $\tilde{L}_{\nabla F_c}$ and $\tilde{L}_{\nabla L_\alpha}$, respectively; it is not difficult to check that

$$\tilde{L}_{\nabla L_\alpha} = \tilde{L}_{\nabla F_c} + \frac{1-\sigma_{\min}(D)}{\alpha_b}. \tag{3.103}$$

The following result replaces Theorem 3.4.1 in the above setting.

Theorem 3.8.1. *Consider Problem (3.1), under Assumptions 2.1'(ii), 3.3.3 and 3.3.4. Let $\{x^\nu\}$ be the sequence generated by DGD in (3.15) under Assumption 3.4.1, with $x_{\mathrm{i}}^0 \in \mathcal{B}_R^d, \mathrm{i} \in [m]$ and $0 < \alpha < \bar{\alpha}_{\max} \triangleq \sigma_{\min}(I+D)/\tilde{L}_{\nabla F_c}$. Then, same conclusions of Theorem 3.4.1 hold.*

Proof. It is sufficient to show that $\{x^\nu\} \subseteq \mathcal{Y}$; the rest of the proof follows similar steps as those in [111, lemma 2] replacing L_c with $\tilde{L}_{\nabla F_c}$.

When $\alpha < \alpha_b$, $\{x^\nu\} \subseteq \mathcal{Y}$ can be proved leveraging the same arguments used in the proof of Lemma 3.4.2. Therefore, in the following, we consider only the case $\alpha_b \leq \alpha < \sigma_{\min}(I +$

$D)/\tilde{L}_{\nabla F_c}$, with $\alpha_b < \sigma_{\min}(I+D)/\tilde{L}_{\nabla F_c}$. We prove by induction. Clearly $x^0 \in \mathcal{L}_{L_\alpha}(L_\alpha(x^0))$ and, by (3.20) (cf. Lemma 3.4.2),

$$\mathcal{L}_{L_\alpha}(L_\alpha(x^0)) \subseteq \bar{\mathcal{L}} \subseteq \mathcal{Y}, \quad \forall \alpha \in [\alpha_b, 1].$$

Assume $\mathcal{L}_{L_\alpha}(L_\alpha(x^\nu)) \subseteq \mathcal{Y}$. Since $x^\nu \in \mathcal{Y}$, there hold $x^{\nu+1} = x^\nu - \alpha \nabla L_\alpha(x^\nu) \in \tilde{\mathcal{Y}}$ and $\theta x^\nu + (1-\theta)x^{\nu+1} \in \tilde{\mathcal{Y}}$, for all $\theta \in [0,1]$. Invoking the descent lemma on L_α at $x^{\nu+1}$ [recall that L_α is $\tilde{L}_{\nabla L_\alpha}$-smooth on $\tilde{\mathcal{Y}}$], we have:

$$L_\alpha(x^{\nu+1}) \leq L_\alpha(x^\nu) - \alpha \left(\frac{\sigma_{\min}(I+D) - \alpha \tilde{L}_{\nabla F_c}}{2} \right) \left\| \nabla L_\alpha(x^\nu) \right\|^2 \leq L_\alpha(x^\nu). \tag{3.104}$$

Therefore, $\mathcal{L}_{L_\alpha}(L_\alpha(x^{\nu+1})) \subseteq \mathcal{L}_{L_\alpha}(L_\alpha(x^\nu)) \subseteq \mathcal{Y}$, which completes the induction. □

3.8.3 Proof of Theorem 3.5.3: Supplement

We first show that, if there exists some ν_0 such that $d^{\nu_0} = 0$, $z^\nu = z^{\nu_0}$, for all $\nu \geq \nu_0$ [see updates in (3.25)]; this means that $\{z^\nu\}$ converges in finitely many iterations. Define $\mathcal{D} \triangleq \{\nu : d^\nu \neq 0\}$ and take ν in $\mathcal{D}$. Let $\theta = 0$, then the KŁ inequality yields $||\nabla L(x^\nu, y^\nu)|| \geq 1/c$, for all $\nu \in \mathcal{D}$. This together with (3.45) and Lemma 3.5.3, lead to $l^{\nu+1} \leq l^\nu - 1/(Mc)^2$, which by Assumption 3.3.1-(ii), implies that $\mathcal{D}$ must be finite and $\{z^\nu\}$ converges in a finite number of iterations.

Consider (3.69). Let $\theta \in (0, 1/2]$, then $(1-\theta)/\theta \geq 1$. Since $D^\nu \to 0$ as $\nu \to \infty$ [by Lemma 3.5.2-(ii)], there exists a sufficiently large ν_0 such that $(D^\nu - D^{\nu+1})^{(1-\theta)/\theta} \leq D^\nu - D^{\nu+1}$. By (3.69), we have

$$D^{\nu+1} \leq \frac{\tilde{M}Mc - 1}{\tilde{M}Mc} D^\nu,$$

which proves case (ii).

Finally, let us assume $\theta \in (1/2, 1)$, then $\theta/(1-\theta) > 1$. Eq. (3.69) implies

$$1 \leq \frac{\bar{M}(D^\nu - D^{\nu+1})}{(D^\nu)^{\theta/(1-\theta)}}$$

where $\bar{M} = (M\tilde{M}c)^{\theta/(1-\theta)}$. Define $h : (0, +\infty) \to \mathbb{R}$ by $h(s) \triangleq s^{-\frac{\theta}{1-\theta}}$. Since h is monotonically decreasing over $[D^{\nu+1}, D^{\nu}]$, we get

$$1 \leq \bar{M}(D^{\nu} - D^{\nu+1})h(D^{\nu}) \leq \bar{M} \int_{D^{\nu+1}}^{D^{\nu}} h(s)ds = \bar{M} \frac{1-\theta}{1-2\theta} \left((D^{\nu})^p - (D^{\nu+1})^p \right), \tag{3.105}$$

with $p = \frac{1-2\theta}{1-\theta} < 0$. By (3.105) one infers that there exists a constant $\mu > 0$ such that $(D^{\nu+1})^p - (D^{\nu})^p \geq \mu$. The following chain of implications then holds: $(D^{\nu+1})^p \geq \mu\nu + (D^1)^p \implies D^{\nu+1} \leq (\mu\nu + (D^1)^p)^{1/p} \implies D^{\nu+1} \leq C_0 \nu^{1/p}$, for some constant $C_0 > 0$. This proves case (iii).

3.8.4 Extension of Proposition 3.5.3

We relax conditions (i)-(ii) of Proposition 3.5.3 under the following additional mild assumptions on the set of strict saddle points and the weight matrices R and C.

Assumption 3.8.1. *There exists $\delta > 0$ such that $\lambda_{\min}(\nabla^2 F(\theta^*)) \leq -\delta$, for all $\theta^* \in \Theta^*_{ss}$ (Θ^*_{ss} is the set of strict saddle of F, cf. Definition 3.5.3).*

Assumption 3.8.2. *The matrices R and C are chosen according to*

$$R = \frac{\widetilde{R} + (t-1)I}{t}, \quad C = \frac{\widetilde{C} + (t-1)I}{t},$$

for some $t \geq 1$, and some matrices $\widetilde{R}$ and $\widetilde{C}$ satisfying Assumption 3.5.1.

Note that R and C satisfy Assumption 3.5.1 as well. The main result is given in Proposition 3.8.1. Before proceeding, we recall the following result on spectral variation of non-normal matrices.

Theorem 3.8.2. *[138, Theorem VIII.1.1] For arbitrary $d \times d$ matrices A and B, it holds that*

$$s\left(\sigma(A), \sigma(B)\right) \leq \left(\|A\| + \|B\|\right)^{1-1/d} \|A - B\|^{1/d}$$

with

$$s\left(\sigma(A), \sigma(B)\right) \triangleq \max_{\mathrm{j}} \min_{\mathrm{i}} |\alpha_{\mathrm{i}} - \beta_{\mathrm{j}}|,$$

where $\alpha_1, \ldots, \alpha_d$ and $\beta_1, \ldots, \beta_d$ are the eigenvalues of A and B, respectively.

Following the same reasoning as in the proof of proposition, it is sufficient to show that for any $u^* \in \mathcal{U}^*$, the Jacobian matrix (recall from eq. (3.84))

$$\mathrm{D}\tilde{g}(u^*) = \begin{bmatrix} W_R - \alpha\nabla^2 F_c^* & -\alpha I \\ (W_C - I)\,\nabla^2 F_c^\star & W_C \end{bmatrix},$$

has an eigenvalue with absolute value strictly greater than 1; proving that such eigenpair is also a member of $\sigma(\mathrm{D}g(u^*))$ follows equivalent steps as in the proof of proposition and thus is omitted. Decompose $\mathrm{D}\tilde{g}(u^*)$ as

$$\mathrm{D}\tilde{g}(u^*) = \underbrace{\begin{bmatrix} I - \alpha\nabla^2 F_c^* & -\alpha I \\ 0 & I \end{bmatrix}}_{\triangleq Q} + \frac{1}{t}\underbrace{\begin{bmatrix} \widetilde{W}_R - I & 0 \\ (\widetilde{W}_C - I)\nabla F_c^\star & \widetilde{W}_C - I \end{bmatrix}}_{\triangleq P_t}, \tag{3.106}$$

where $\widetilde{W}_R \triangleq \widetilde{R} \otimes I_d$ and $\widetilde{W}_C \triangleq \widetilde{C} \otimes I_d$. Eq. (3.106) reads the Jacobian matrix $\mathrm{D}\tilde{g}(u^*)$ as a variation of Q by perturbation P_t. For any $u^* \in \mathcal{U}^*$, the spectrum of Q consists of $m \cdot d$ counts of 1 along with the eigenvalues of $I - \alpha\nabla^2 F_c^*$, which contains a real eigenvalue $\lambda_1 \geq 1 + \alpha\delta/(md)$, since $\theta^* \in \Theta_{ss}^*$. Theorem 3.8.2 guarantees that the spectrum variation of any perturbed arbitrary non-normal matrices is bounded by the norm of the perturbation matrix. Thus it is sufficient to show that the perturbed λ_1, as a member of $\sigma(\mathrm{D}\tilde{g}(u^*))$, is strictly greater than 1.

Applying Theorem 3.8.2 gives the following sufficient conditions: denote $\tilde{d} \triangleq 2md$,

$$(\|Q + P_t\| + \|Q\|)^{1-1/\tilde{d}}\,\|P_t\|^{1/\tilde{d}} < 2\alpha\delta/\tilde{d}. \tag{3.107}$$

By sub-additivity of the matrix norm, it is sufficient for (3.107) that

$$(\|P_t\| + 2\|Q\|)^{1-1/\tilde{d}}\,\|P_t\|^{1/\tilde{d}} \leq \frac{\alpha\delta}{\tilde{d}}. \tag{3.108}$$

Since each ∇f_{i} is Lipschitz continuous (cf. Assumption 3.3.1), there exist constants $C_Q > 0$ and $C_P > 0$ such that $\max_{u^* \in \mathcal{U}^*} \|Q\| \leq C_Q$ and $\max_{u^* \in \mathcal{U}^*} \|P_t\| \leq C_P/t$. It is not difficult to show that a sufficient condition for (3.108) is

$$t \geq \frac{(C_P + 2C_Q)^{\tilde{d}-1} C_P}{(\alpha\delta/\tilde{d})^{\tilde{d}}}, \qquad \tilde{d} = 2md. \tag{3.109}$$

Proposition 3.8.1. *Let Assumptions 3.5.1 and 3.8.1 hold, and matrices R and C be chosen according to Assumption 3.8.2, with t satisfying (3.109). Then, any consensual strict saddle point is an unstable fixed point of g, i.e., $\mathcal{U}^* \subseteq \mathcal{A}_g$, with $\mathcal{A}_g$ and $\mathcal{U}^*$ defined in (3.70) and (3.72), respectively.*

Note that above proposition ensures $\mathcal{U}^* \subseteq \mathcal{A}_g$ under (3.109) and given step-size α. Convergence of the sequence is proved under (3.52) and (3.78) for the step-size α. However, (3.52) may not hold for some large t (there can be instances where the set of step-size satisfying conditions (3.109) and (3.52) is empty). Hence, when $d > 1$, the statement in Theorem 3.5.5 is conditioned to the convergence of the algorithm.

Part II

Distributed Convex Optimization

4. DECENTRALIZED FIRST-ORDER ALGORITHMS FOR (STRONGLY) CONVEX OPTIMIZATION OVER (TIME-VARYING) NETWORKS

In this chapter, we study a general form of constrained non-smooth optimization over networks:

$$\begin{aligned} \min_{x} \quad & U(x) \triangleq \underbrace{\frac{1}{m}\sum_{\mathrm{i}=1}^{m} f_{\mathrm{i}}(x)}_{F(x)} + G(x) \\ \text{s.t.} \quad & x \in \mathcal{K}, \end{aligned} \tag{4.1}$$

where $f_{\mathrm{i}} : \mathbb{R}^d \to \mathbb{R}$ 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. Agents are connected through a communication network, modeled as a graph, possibly directed and/or time-varying. The goal is to cooperatively solve (4.1) by exchanging information only with their immediate neighbors.

Our focus pertains to such a design in two possible settings (one being a special case of the other) [139]: **1)** The scenario where no significant relationship can be assumed among the local functions f_{i}–this is what the literature of distributed optimization has extensively studied, and will be refereed to as the *unrelated* setting—and **2)** the case where the f_{i}'s are *related*, e.g., because they reflect statistical similarity in the data residing at different nodes. For instance, in the distributed ERM problem above, when data are i.i.d. among machines, one can show that quantities such as the gradients and Hessian matrices of the local functions differ only by $\beta = \mathcal{O}(1/\sqrt{n})$, due to concentrations of measure effects [140], [141]–we will refer to this as β-*related* setting (cf. Sec. 4.1.1.2). If properly exploited in the algorithmic design, such similarity can speed up the optimization/learning process over general purpose optimization algorithms.

Centralized algorithms Problem (4.1) in the two settings above has been extensively studied in the centralized environment, including star-networks wherein there is a master

node connected to all the other workers. Our interest is in the following (non-accelerated) algorithms:

1) Unrelated setting: (4.1) can be solved on star-networks employing the standard proximal gradient method: to reach precision $\epsilon > 0$ on the objective value, one needs $\mathcal{O}\Big(\kappa_g \log(1/\epsilon)\Big)$ iterations (which is also the number of communication rounds between the master and the workers), where κ_g is the condition number of F.

2) β-related setting: When the agents' functions $f_{\rm i}$ are sufficiently similar, a linear rate proportional to κ_g may be highly suboptimal. For instance, in the extreme case where all $f_{\rm i}$'s are identical ($\beta = 0$), the number of iterations/communications to an $\epsilon > 0$ solution would remain the same as for $\beta = \mathcal{O}(L)$. In fact, when $1 + \beta/\mu < \kappa_g$, faster rates can be obtained exploiting the similarity of the $f_{\rm i}$'s. Specifically, [140] proposed DANE: a mirror-descent type algorithm over star-networks, where each worker i replaces the quadratic term in its local proximal-gradient update with the Bregman divergence of the reference function $f_{\rm i} + \beta/2\| \bullet \|^2$; and the master averages the solutions of the workers. DANE is applicable to (4.1) with $G = 0$: For *quadratic* losses, it achieves an ϵ-solution in $\mathcal{O}\Big((\beta/\mu)^2 \cdot \log(1/\epsilon)\Big)$ iterations/communications (it is assumed $\beta/\mu \geq 1$) while no improvement is proved over the proximal gradient if the $f_{\rm i}$'s are not quadratic. More recently, [142] proposed CEASE, which achieves DANE's rate for (4.1) with $G \neq 0$ and nonquadratic losses. Using recent results in [143], it is not difficult to check that the mirror-descent algorithm implemented at the master (thus without averaging workers' iterates) with the Bregman divergence of $f_1 + \beta/2\| \bullet \|^2$ (f_1 is the local function at the master) achieves an $\epsilon > 0$ solution in $\widetilde{\mathcal{O}}\Big(\beta/\mu \cdot \log(1/\epsilon)\Big)$ iterations/communications, improving thus on DANE/CEASE's rates.

A natural question is whether similar results–in particular the dependence of the rate on global optimization parameters as obtained on star-networks in the unrelated and β-related settings–are achievable over general network topologies, possibly time-varying and directed. The literature of distributed algorithms over general network topologies–albeit vast–do not provide a satisfactory answer, leaving a gap between rate results over star networks and what has been certified over general graphs–see Sec 4.0.2 for a review of the state of the art. In a nutshell, **(i)** there are no distributed schemes provably achieving linear rate for (4.1) with $G \neq 0$ and/or constraints (cf. Table 4.1). Furthermore, even considering the unconstrained

minimization of F (i.e., $G = 0$ and $\mathcal{K} = \mathbb{R}^d$), **(ii)** linear convergence is certified at a rate depending on much larger quantities than the global condition number κ_g–see Table 4.2; 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 achieved over star-topologies. The goal is to close exactly this gap.

4.0.1 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) [144], applicable to the *composite, constrained* formulation (4.1) over (time-varying, directed) graphs. It combines the use of surrogate functions in the agents' subproblems with a perturbed (push-sum) consensus mechanism that aims at locally tracking the gradient of F. Surrogate functions replace the more classical first order approximation of the local f_i's, which is the omnipresent choice in current distributed algorithms, offering the potential to better suit the geometry of the problem. For instance, (approximate) Newton-type subproblems or mirror descent-type updates naturally fit our surrogate models; they are the key enabler of provably faster rates in the β-related setting. We comment SONATA's rates below (cf. Table 4.3).

2. **Unrelated setting (Table 4.3):** 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 show a much more pessimistic dependence on the optimization parameters and are proved under more restrictive assumptions–contrast Table 4.2 with Table 4.3. Linear rates over time-varying digraphs are reported in Table 4.4 (cf. Sec. 4.3.2).

3. **β-related setting (Table 4.3):** 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 4.2). Notice that this is achieved without exchanging any Hessian matrix over the network but leveraging function homogeneity via surrogation. When customized over star-topologies, SONATA's rates improve on DANE/CEASE's ones too.

Table 4.1. Existing linearly convergent distributed algorithms. SONATA is the only scheme achieving linear rate in the presence of G in (4.1) or constraints. The explicit expression of the rates of the above nonaccelerated schemes (for which is available) is reported in Table 4.2.

Algorithms		[145]–[153]	[154]–[157]	[126], [158]–[161]	**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 [153]		only [126], [161]	✓
	digraph		✓	only [126], [161]	✓

4.0.2 Related works

Early works on distributed optimization aimed at decentralizing the (sub)gradient algorithm. The Distributed Gradient Descent (DGD) was introduced in [69] for unconstrained instances of (4.1) and in [165] for least squares, bot over undirected graphs. A refined convergence rate analysis of DGD [69] can be found in [166]. Subsequent variants of DGD include the projected (sub)gradient algorithm [167] and the push-sum gradient consensus algorithm [168], the latter implementable over digraphs. While different, the updates of the agents' variables in the above algorithms can be abstracted as a combination of one (or multiple) consensus step(s) (weighted average with neighbors variables) and a local (sub)gradient descent step, controlled by a step-size (in some schemes, followed by a proximal operation). A diminishing step-size is used to reach *exact* consensus on the solution, converging thus at a

Table 4.2. Linear rate of existing non-accelerated algorithms over undirected graphs: communications rounds to reach $\epsilon > 0$ accuracy; $L_{\rm i}$ and $\mu_{\rm i}$ are the smoothness and strong convexity constants of $f_{\rm i}$'s, respectively; $L_{\rm mx} \triangleq max_{\rm i} L_{\rm i}$, $\mu_{\rm mn} \triangleq \min_{\rm i} \mu_{\rm i}$; and $\rho \in [0, 1)$ is the second largest eigenvalue modulus of the mixing matrix [cf. (4.27)]. The rates above include the quantities κ_l, $\hat{\kappa}$, and $\breve{\kappa}$ rather than the much desirable global condition number $\kappa_g \triangleq L/\mu$ (L and μ are the smoothness and strong convexity constants of F, respectively). Furthermore, they are independent on β, implying that faster rates are not certified when $1 + \beta/\mu < \kappa_g$ (β-related setting).

Algorithm	**Problem**	**Linear rate:** $\mathcal{O}(\delta \log(1/\epsilon))$
EXTRA [146]	F	$\delta = \mathcal{O}(\frac{\kappa_\ell^2}{1-\rho}), \quad \kappa_\ell = \frac{L_{\rm mx}}{\mu_{\rm mn}}$
DIGing [126], [159]	F	$\delta = \frac{\hat{\kappa}^{1.5}}{(1-\rho)^2}, \quad \hat{\kappa} \triangleq \frac{L_{\rm mx}}{(1/m)\sum_{\rm i} \mu_{\rm i}}$
Harnessing [145]	F	$\delta = \frac{\kappa_\ell^2}{(1-\rho)^2}$
NIDS [151], ABC [162]	F	$\delta = \max\{\kappa_\ell, \frac{1}{1-\rho}\}$
Exact Diffusion [160]	F	$\delta = \frac{\bar{\kappa}^2}{1-\rho}, \quad \bar{\kappa} \triangleq \frac{L_{\rm mx}}{\mu_{\rm mx}}$
Augmented Lagrangian [150]	F	$\delta = \frac{\kappa_\ell}{1-\rho}$
ADMM [149]	F	$\frac{\kappa_\ell^4}{1-\rho}$

sublinear rate. With a fixed step-size α, linear rate of the iterates is achievable, but it can only converge to a $\mathcal{O}(\alpha)$-neighborhood of the solution [69], [166].

Several subsequent attempts have been proposed to cope with this speed-accuracy dilemma, leading to algorithms converging to the *exact* solution while employing a *constant* step-size. Based upon the mechanism put forth to cancel the steady state error in the individual gradient direction, existing proposals can be roughly organized in three groups, namely: i) primal-based distributed methods leveraging the idea of gradient tracking [73], [126], [145], [154]–[156], [169]–[174]; ii) distributed schemes using ad-hoc corrections of the local optimization direction [146], [157], [175]; and iii) primal-dual-based methods [147]–[150], [164]. We elaborate next on these works, focusing on schemes achieving linear rate– Table 4.1 organizes these schemes based upon the setting their convergence is established while Table 4.2 reports the explicit expression of the rates.

Table 4.3. Summary of convergence rates of SONATA over undirected graphs: number of communication rounds to reach ϵ-accuracy. In the table, β is the homogeneity parameter measuring the similarity of the loss functions f_{i}'s (cf. Definition 5.1.4); the other quantities are defined as in Table 4.2. The extra averaging steps are performed using Chebyshev acceleration [163], [164]. The $\widetilde{O}$ notation hides log dependence on κ_g and β/μ (see Sec. 4.2.4.2 for the exact expressions). Rates over time-varying directed graphs are summarized in Table 4.4 (cf. Sec. 4.3.2).

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

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 [169], [170] for Problem (4.1) and in AUG-DGM [73] for strongly convex, smooth, unconstrained optimization. The work [176] introduced SONATA, extending NEXT over (time-varying) digraphs. A convergence rate analysis of [73] was later developed in [126], [145], [177], with [126] considering also (time-varying) digraphs. Other algorithms based on the idea of gradient tracking and implementable over digraphs are ADD-OPT [172] and [154]. Subsequent schemes, [155], the Push-Pull [156], and the $\mathcal{AB}$ [161] algorithms, relaxed previous conditions on the mixing matrices used in

the consensus and gradient tracking steps over digraphs, which neither need to be row- nor column-stochastic. All the schemes above but NEXT and SONATA are applicable only to *smooth, unconstrained* instances of (4.1), 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 [146] and its variant over digraphs, EXTRA-PUSH [157], introduce two different weight matrices for any two consecutive iterations as well as leverage history of gradient information. They are applicable only to it smooth, unconstrained problems; when each f_i is strongly convex, they generate iterates that converge linearly to the minimizer of F. To deal with an additive convex nonsmooth term in the objective, [178] proposed PG-EXTRA, which is thus applicable to (4.1) over undirected graphs, possibly with different local nonsmooth functions. However, linear convergence is not certified. A different approach is to use a linearly increasing number of consensus steps rather than correcting directly the gradient direction; this has been studied in [175] 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 primal-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 [149], [179] and their inexact implementations [147], [158]; ii) distributed Augmented Lagrangian-based methods with randomized primal variable updates [150]; and iii) a distributed dual ascent method employing tracking of the average of the primal variable [153]. All these schemes are applicable only to *smooth, unconstrained* optimization over undirected graphs, with [153] handling time-varying graphs. The extension of these methods to digraphs seems not straightforward, because it is not clear how to enforce consensus via constraints over directed networks.

To summarize, the above literature review shows that currently there exists no distributed algorithm for the general formulation (4.1) that provably converges at linear rate to the exact solution, in the presence of a nonsmooth function G or constraints (cf. Table 4.1); let alone

mentioning digraphs. Furthermore, when it comes to the dependence of the rate on the optimization parameters, Table 4.3 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 this work was under peer-review process and publicly available in [180], a few other related technical reports appeared online [181]–[183], which we briefly discuss next. The authors in [181] studied a class of distributed proximal gradient-based methods to solve Problem (4.1) 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 [182] proposed an inexact distributed projected gradient descent method for the unconstraint minimization of F and proved a communication complexity of $\tilde{O}\Big(\kappa_g\,(1-\rho)^{-1}\log^2(1/\epsilon)\Big)$ ($\tilde{\mathcal{O}}$ hides a log-dependence on $L_{\rm max}^2/\mu^2$), which is determined by the global condition number κ_g; the algorithm runs over time-varying, undirected, graphs (as long as they are connected at each iteration). SONATA's rates compare favorably with those above. Furthermore, since both schemes [181] and [182] are gradient-type methods, unlike SONATA, their performance cannot benefit from function similarity, resulting in convergence rates independent on β. On the other hand, [183] 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 [183] is established only for the *unconstrained* minimization of F ($G=0$ and $\mathcal{K}=\mathbb{R}^d$) over *undirected* graphs, with *each* $f_{\rm i}$ assumed to be strongly convex. Second, convergence rates therein are more pessimistic than what predicted by our analysis. In fact, the best communication complexity of Network-DANE reads $\tilde{O}\Big((1+(\beta/\mu)^2)(1-\rho)^{-1/2}\log(1/\epsilon)\Big)$ for quadratic $f_{\rm i}$'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 4.3, shows that SONATA' rates exhibit a better dependence on the optimization parameters (κ_g vs. κ_ℓ) and β/μ in all scenarios. In particular, in the β-related setting, SONATA retains faster rates, even when $f_{\rm i}$'s are nonquadratic.

4.1 Problem & Network Setting

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

4.1.1 Assumptions on Problem (4.1)

Our algorithmic design and convergence results pertain to two problem settings, namely: i) the one where the local functions f_{i} are generic and unrelated (cf. Sec. 4.1.1.1), and ii) the case where they are related (cf. Sec. 4.1.1.2). These two settings are formally introduced below.

4.1.1.1 The unrelated setting

Consider the following standard assumption.

Assumption 4.1.1 (On Problem (4.1))**.** *4.1.1.1 The set $\emptyset \neq \mathcal{K} \subseteq \mathbb{R}^d$ is closed and convex;*

4.1.1.2 Each $f_{\mathrm{i}} : \mathcal{O} \to \mathbb{R}$ is twice differentiable on the open set $\mathcal{O} \supseteq \mathcal{K}$ and convex;

4.1.1.3 F satisfies

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

with $\mu > 0$ and $0 < L < \infty$;

4.1.1.4 $G : \mathcal{K} \to \mathbb{R}$ is convex possibly nonsmooth.

Note that 4.1.1.3 together with 4.1.1.2 imply

$$\mu_{\mathrm{i}} I \preceq \nabla^2 f_{\mathrm{i}}(x) \preceq L_{\mathrm{i}} I, \quad \forall x \in \mathcal{K}, \ \forall \mathrm{i} \in [m], \tag{4.2}$$

for some $\mu_{\mathrm{i}} \geq 0$ and $0 < L_{\mathrm{i}} < \infty$. Unlike existing works (cf. Table 4.1), we do not require each f_{i} to be strongly convex but just F (cf. 4.1.1.3). Also, twice differentiability of f_{i} is not really necessary, but assumed here to simplify our derivations.

Under Assumption 4.1.1, we define the global conditional number associated to (4.1):

$$\kappa_g \triangleq \frac{L}{\mu}. \tag{4.3}$$

Related quantities determining the (linear) convergence rate of existing distributed algorithms are (cf. Table 4.2):

$$\kappa_\ell \triangleq \frac{L_{\rm mx}}{\mu_{\rm mn}}, \quad \hat{\kappa} \triangleq \frac{L_{\rm mx}}{(1/m)\sum_{\rm i}\mu_{\rm i}}, \quad \breve{\kappa} \triangleq \frac{L_{\rm mx}}{\mu}, \quad \text{and} \quad \bar{\kappa} \triangleq \frac{L_{\rm mx}}{\mu_{\rm mx}}, \tag{4.4}$$

where

$$L_{\rm mx} \triangleq \max_{{\rm i}=1,\ldots,m} L_{\rm i}, \quad \mu_{\rm mn} \triangleq \min_{{\rm i}=1,\ldots,m} \mu_{\rm i}, \quad \text{and} \quad \mu_{\rm mx} \triangleq \max_{{\rm i}=1,\ldots,m} \mu_{\rm i}. \tag{4.5}$$

When $\mu_{\rm i} = 0$, we set $\kappa_\ell = \infty$. It is not difficult to check that κ_g can be much smaller than $\breve{\kappa}$, $\bar{\kappa}$, $\hat{\kappa}$ and κ_ℓ, as shown in the following example.

Example 1: Consider the following instance of Problem (4.1):

$$f_{\rm i}(x) = \frac{1}{2}x^\top\left(\mathtt{a}I + m\cdot\mathtt{b}\,\mathrm{diag}(\mathrm{e}_{\rm i})\right)x, \quad F(x) = \frac{1}{m}\sum_{\rm i=1}^{m} f_{\rm i}(x) = \frac{\mathtt{a}+\mathtt{b}}{2}\|x\|^2,$$

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

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

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

In the setting above, our goal is to design linearly convergent distributed algorithms whose iterations complexity is proportional to κ_g, instead of the larger quantities in (4.4).

4.1.1.2 The β-related setting

This setting considers explicitly the case where the functions $f_{\rm i}$ are similar, in the sense defined below [139].

Definition 4.1.1 (β-related $f_{\rm i}$'s)**.** *The local functions $f_{\rm i}$'s (satisfying Assumption 4.1.1) are called β-related if $\|\nabla^2 F(x) - \nabla^2 f_{\rm i}(x)\|_2 \leq \beta$, for all $x \in \mathcal{K}$ and some $\beta \geq 0$.*

The more similar the $f_{\rm i}$'s, the smaller β. For arbitrary $f_{\rm i}$'s, β is of the order of

$$\beta \leq \max_{{\rm i}=1,\ldots,m} \sup_{x\in\mathcal{K},\, \|u\|=1} \left| u^\top \left(\nabla^2 F(x) - \nabla^2 f_{\rm i}(x) u \right) \right| \leq \max_{{\rm i}=1,\ldots,m} \max \left\{ |L - \mu_{\rm i}|,\, |\mu - L_{\rm i}| \right\}. \tag{4.6}$$

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

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

$$x^\star \in \operatorname*{argmin}_{x\in\mathcal{H}} F(x), \quad \text{with} \quad F(x) \triangleq \mathbb{E}_{z\sim\mathcal{P}} \left[f(x; z) \right], \tag{4.7}$$

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 $z \in \mathcal{Z}$; $\mathcal{H} \subseteq \mathbb{R}^d$ 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 $z \in \mathcal{Z}$. Furthermore, we assume that any $x^\star \in \mathcal{B}_B \triangleq \{x : \|x\| \leq B\}$, for some $0 < B < \infty$. This setting includes, for example, supervised generalized linear models, where $z = (w, y)$ and $f(x; (w, y)) = \ell(\phi(w)^\top x; y)$, for some (strongly) convex loss $\ell(\bullet; y)$ and feature mapping ϕ. For instance, in linear regression, $f(x; (w, y)) = (y - \phi(w)^\top x)^2$, with $\phi(w) \in \mathbb{R}^d$ and $y \in \mathbb{R}$; for logistic regression, we have $f(x; (w, y)) = \log(1 + \exp(-y(\phi(w)^\top x)))$, with $w \in \mathbb{R}^d$ and $y \in \{-1, 1\}$.

To solve (4.7), 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 earlier, the ERM problem reads:

$$\widehat{x} \triangleq \operatorname*{argmin}_{x\in\mathcal{H}} \widehat{F}(x) \triangleq \frac{1}{m} \sum_{{\rm i}=1}^{m} f_{\rm i}(x; \mathcal{D}^{({\rm i})}), \qquad f_{\rm i}(x; \mathcal{D}^{({\rm i})}) = \frac{1}{n} \sum_{{\rm j}=1}^{n} f(x; z_{\rm j}^{({\rm i})}) + \frac{\lambda}{2} \|x\|^2, \tag{4.8}$$

where $f_{\rm i}$ is regularized empirical loss of agent i, λ-strongly convex. Clearly (4.8) is an instance of (4.1), satisfying Assumption 4.1.1.

For the ERM problems (4.8) we derive next the associated β/μ and contrasts with κ_g. $\widehat{F}$ is λ-strongly convex; therefore, we can set $\mu = \lambda$. The optimal choice of λ is the one minimizing the statistical error resulting in using $\widehat{x}$ as proxy for $x^\star$. We have [184, Th. 7], with high probability, $F(\widehat{x}) - F(x^\star) \leq \frac{\lambda}{2}\|\theta^\star\|^2 + \mathcal{O}\big(\frac{G_f^2}{\lambda N}\big) \leq \mathcal{O}\big(\lambda B^2 + \frac{G_f^2}{\lambda N}\big)$, where G_f is the Lipschitz constant of $f(\bullet; z)$ on $\mathcal{H} \bigcap \mathcal{B}_B$, for all $z \in \mathcal{Z}$. The optimal choice of λ and resulting minimum error rate are then

$$\lambda = \mathcal{O}\left(\sqrt{\frac{G^2}{B^2\, N}}\right) \quad \Rightarrow \quad F(\widehat{x}) - F(x^\star) \leq \mathcal{O}\left(\sqrt{\frac{G^2 B^2}{N}}\right). \tag{4.9}$$

An estimate of β can be obtained exploring the statistical similarity of the local empirical losses f_{i} in (4.8). Under the additional assumption that $\nabla^2 f(\bullet; z)$ is M-Lipchitz on $\mathcal{H}$, for all $z \in \mathcal{Z}$, a minor modification of [185, Lemma 6] applied to (4.7)-(4.8), yields: with high probability,

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

with

$$\beta = \begin{cases} \widetilde{\mathcal{O}}\left(\sqrt{\frac{L^2}{n}}\right), & \text{if } M = 0; \\ \\ \widetilde{\mathcal{O}}\left(\sqrt{\frac{L^2\, d}{n}}\right), & \text{otherwise,} \end{cases} \tag{4.10}$$

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

Based on (4.9)-(4.10), an estimate of β/μ and κ_g for (4.8) reads:

$$1 + \frac{\beta}{\mu} = 1 + \widetilde{\mathcal{O}}\left(L\sqrt{d\, m}\right) \quad \text{and} \quad \kappa_g = 1 + \widetilde{\mathcal{O}}\left(L\sqrt{d\, m\, n}\right). \tag{4.11}$$

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

4.1.2 Network setting

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

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

Assumption 4.1.2 (On the network)**.** *The graph $\mathcal{G}$ is connected.*

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

Assumption 4.1.3 (On the network)**.** *The graph sequence $\{\mathcal{G}^\nu\}$, $\nu = 0, 1, \dots$, is B-strongly connected, i.e., there exists a finite integer $B > 0$ such that the graph with edge set $\cup_{t=\nu B}^{(\nu+1)B-1} \mathcal{E}^t$ is strongly connected, for all $\nu = 0, 1, \dots$.*

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

4.2 The SONATA algorithm over undirected graphs

We recall here the SONATA/NEXT algorithm [144], [170], customized to undirected, static, graphs. Each agent i maintains and updates iteratively a local copy $x_\mathrm{i} \in \mathbb{R}^d$ of the global variable x, along with the auxiliary variable $y_\mathrm{i} \in \mathbb{R}^d$, which estimates the gradient of F. Denoting by x_i^ν (resp. y_i^ν) the values of x_i (resp. y_i) at iteration $\nu = 0, 1, \dots$, the SONATA algorithms is described in Algorithm 2. In words, each agent i, given the current iterates x_i^ν

Algorithm 2: SONATA over undirected graphs

Data: $x_{\mathrm{i}}^0 \in \mathcal{K}$ and $y_{\mathrm{i}}^0 = \nabla f_{\mathrm{i}}(x_{\mathrm{i}}^0)$, $\mathrm{i} \in [m]$.
Iterate: $\nu = 1, 2, \ldots$
`[S.1] [Distributed Local Optimization]` Each agent i solves

$$\widehat{x}_{\mathrm{i}}^{\nu} \triangleq \operatorname*{argmin}_{x_{\mathrm{i}} \in \mathcal{K}} \underbrace{\widetilde{f}_{\mathrm{i}}(x_{\mathrm{i}}; x_{\mathrm{i}}^{\nu}) + \Big(y_{\mathrm{i}}^{\nu} - \nabla f_{\mathrm{i}}(x_{\mathrm{i}}^{\nu})\Big)^{\top}(x_{\mathrm{i}} - x_{\mathrm{i}}^{\nu})}_{\widetilde{F}_{\mathrm{i}}(x_{\mathrm{i}}; x_{\mathrm{i}}^{\nu})} + G(x_{\mathrm{i}}), \tag{4.12a}$$

and updates

$$x_{\mathrm{i}}^{\nu+\frac{1}{2}} = x_{\mathrm{i}}^{\nu} + \alpha \cdot d_{\mathrm{i}}^{\nu}, \quad \text{with} \quad d_{\mathrm{i}}^{\nu} \triangleq \widehat{x}_{\mathrm{i}}^{\nu} - x_{\mathrm{i}}^{\nu}; \tag{4.12b}$$

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

(a) `Consensus`

$$x_{\mathrm{i}}^{\nu+1} = \sum_{\mathrm{j}=1}^{m} w_{\mathrm{ij}} x_{\mathrm{j}}^{\nu+\frac{1}{2}}, \tag{4.12c}$$

(b) `Gradient tracking`

$$y_{\mathrm{i}}^{\nu+1} = \sum_{\mathrm{j}=1}^{m} w_{\mathrm{ij}} \Big(y_{\mathrm{j}}^{\nu} + \nabla f_{\mathrm{j}}(x_{\mathrm{j}}^{\nu+1}) - \nabla f_{\mathrm{j}}(x_{\mathrm{j}}^{\nu})\Big). \tag{4.12d}$$

end

and y_{i}^{ν}, first solves a strongly convex optimization problem wherein $\widetilde{F}_{\mathrm{i}}$ is an approximation of the sum-cost F at x_{i}^{ν}; $\widetilde{f}_{\mathrm{i}}$ in (4.12a) is a strongly convex function, which plays the role of a surrogate of f_{i} (cf. Assumption 4.2.1 below) while y_{i}^{ν} acts as approximation of the gradient of F at x_{i}^{ν}, that is, $\nabla F(x_{\mathrm{i}}^{\nu}) \approx y_{\mathrm{i}}^{\nu}$ (see discussion below). Then, agent i updates x_{i}^{ν} along the local direction d_{i}^{ν} [cf. (4.12b)], using the step-size $\alpha \in (0, 1]$; the resulting point $x_{\mathrm{i}}^{\nu+1/2}$ is broadcast to its neighbors. The update $x_{\mathrm{i}}^{\nu+1/2} \to x_{\mathrm{i}}^{\nu+1}$ is obtained via the consensus step (4.12c) while the y-variables are updated via the perturbed consensus (4.12d), aiming at tracking $\nabla F(x_{\mathrm{i}}^{\nu})$.

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

• *On the subproblem (4.12a) and surrogate functions* $\widetilde{f}_{\mathrm{i}}$ The surrogate functions satisfy the following conditions.

Assumption 4.2.1. *Each* $\widetilde{f}_{\mathrm{i}} : \mathcal{O} \times \mathcal{O} \to \mathbb{R}$ *is* C^2 *and satisfies*

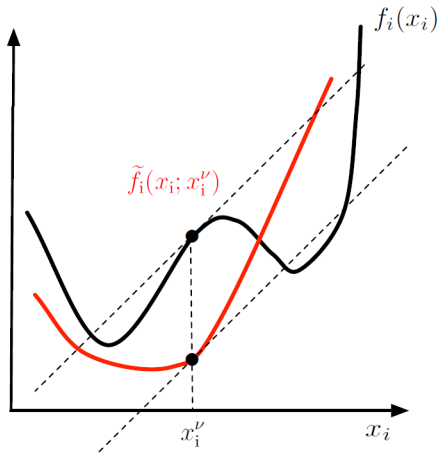


Figure 4.1. Illustration of surrogate function $\widetilde{f}_i$

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

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

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

where $\nabla \widetilde{f}_i(x;z)$ *is the partial gradient of* $\widetilde{f}_i$ *at* (x,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 x_i^ν the first order properties of f_i (see Fig. 4.1). Conditions (i)-(iii) are certainly satisfied if one uses the classical linearization of f_i, that is,

$$\widetilde{f}_i(x_i; x_i^\nu) = \nabla f_i(x_i^\nu)^\top (x_i - x_i^\nu) + \frac{\tau_i}{2}\|x_i - x_i^\nu\|^2, \tag{4.13}$$

with $\tau_i > 0$, which leads to the standard proximal-gradient update for $\widehat{x}_i$. Note that if, in addition, $G = 0$ and $\mathcal{K} = \mathbb{R}^d$, (4.12a)–(4.12c) reduces to the standard (ATC) consensus/gradient-tracking step (setting $\alpha = 1$ and absorbing $1/\tau_i$ into the common stepsize γ): $x_i^{\nu+1} = \sum_j w_{ij}(x_i^\nu - \gamma\, y_i^\nu)$ [73], [126], [145]. However, Assumption 4.2.1 allows us to cover a much wider array of approximations that better suit the geometry of the problem at hand, en-

hancing convergence speed. For instance, on the opposite side of (4.13), we have a surrogate retaining all the structure of f_{i}, such as

$$\widetilde{f}_{\mathrm{i}}(x_{\mathrm{i}}; x_{\mathrm{i}}^{\nu}) = f_{\mathrm{i}}(x_{\mathrm{i}}) + \frac{\tau_{\mathrm{i}}}{2}\|x_{\mathrm{i}} - x_{\mathrm{i}}^{\nu}\|^2, \tag{4.14}$$

with $\tau_{\mathrm{i}} > 0$. Using (4.14), one can rewrite (4.12a) as:

$$\widehat{x}_{\mathrm{i}}^{\nu} = \operatorname*{argmin}_{x_{\mathrm{i}} \in \mathcal{K}} \Big(1 \cdot \underbrace{(\nabla f_{\mathrm{i}}(x_{\mathrm{i}}^{\nu}) + y_{\mathrm{i}}^{\nu})}_{\nabla g(x_{\mathrm{i}}^{\nu})} - \underbrace{(\nabla f_{\mathrm{i}}(x_{\mathrm{i}}^{\nu}) + \tau_{\mathrm{i}} x_{\mathrm{i}}^{\nu})}_{\nabla \omega(x_{\mathrm{i}}^{\nu})}\Big)^{\top} x_{\mathrm{i}} + \underbrace{\Big(f_{\mathrm{i}}(x_{\mathrm{i}}) + \frac{\tau_{\mathrm{i}}}{2}\|x_{\mathrm{i}}\|^2\Big)}_{\omega(x_{\mathrm{i}})} + G(x_{\mathrm{i}}), \tag{4.15}$$

which can be interpreted as a mirror-descent update (with step-size one) for the composite minimization of $g(x_{\mathrm{i}}) \triangleq f_{\mathrm{i}}(x_{\mathrm{i}}) + (y_{\mathrm{i}}^{\nu})^{\top}(x_{\mathrm{i}} - x_{\mathrm{i}}^{\nu})$, based on the Bregman distance associated with the reference function $\omega(x_{\mathrm{i}}) \triangleq f_{\mathrm{i}}(x_{\mathrm{i}}) + \tau_{\mathrm{i}}/2\|x_{\mathrm{i}}\|^2$.

We refer the reader to [6], [186], [187] as good sources of examples of nonlinear surrogates satisfying Assumption 4.2.1; here we only anticipate that, when the f_{i}'s are sufficiently similar, higher order models such as (4.14) yield indeed faster rates of SONATA than those achievable using linear surrogates (4.13). Further intuition is provided next.

Under Assumption 4.2.1, it is not difficult to check that, for every $\mathrm{i} \in [m]$, there exist constants D_{i}^{ℓ} and D_{i}^{u}, $D_{\mathrm{i}}^{\ell} \leq D_{\mathrm{i}}^{u}$, such that

$$D_{\mathrm{i}}^{\ell}\, I \preceq \nabla^2 \widetilde{f}_{\mathrm{i}}(x, y) - \nabla^2 F(x) \preceq D_{\mathrm{i}}^{u}\, I, \quad \forall x, y \in \mathcal{K}; \qquad \text{let} \;\; D_{\mathrm{i}} \triangleq \max\{|D_{\mathrm{i}}^{\ell}|, |D_{\mathrm{i}}^{u}|\}. \tag{4.16}$$

For instance, (4.16) holds with $D_{\mathrm{i}} = \max\{|\widetilde{\mu}_{\mathrm{i}} - L|, |\widetilde{L}_{\mathrm{i}} - \mu|\}$. Roughly speaking, the smaller D_{i} the better $\widetilde{F}_{\mathrm{i}}$ in (4.12a) approximates F. To see this, compare F and $\widetilde{F}_{\mathrm{i}}$ up to the second order: there exist $\theta_1, \theta_2 \in (0, 1)$ such that

$$\begin{aligned}
\widetilde{F}_{\mathrm{i}}(x_{\mathrm{i}}; x_{\mathrm{i}}^{\nu}) &= \widetilde{f}_{\mathrm{i}}(x_{\mathrm{i}}^{\nu}; x_{\mathrm{i}}^{\nu}) + \Big(y_{\mathrm{i}}^{\nu} - \nabla f_{\mathrm{i}}(x_{\mathrm{i}}^{\nu}) + \nabla \widetilde{f}_{\mathrm{i}}(x_{\mathrm{i}}^{\nu}; x_{\mathrm{i}}^{\nu})\Big)^{\top}(x_{\mathrm{i}} - x_{\mathrm{i}}^{\nu}) \\
&\quad + \frac{1}{2}(x_{\mathrm{i}} - x_{\mathrm{i}}^{\nu})^{\top} \nabla^2 \widetilde{f}_{\mathrm{i}}\Big(x_{\mathrm{i}}^{\nu} + \theta_1(x_{\mathrm{i}} - x_{\mathrm{i}}^{\nu}); x_{\mathrm{i}}^{\nu}\Big)(x_{\mathrm{i}} - x_{\mathrm{i}}^{\nu}) \\
F(x_{\mathrm{i}}) &= F(x_{\mathrm{i}}^{\nu}) + \nabla F(x_{\mathrm{i}}^{\nu})^{\top}(x_{\mathrm{i}} - x_{\mathrm{i}}^{\nu}) \\
&\quad + \frac{1}{2}(x_{\mathrm{i}} - x_{\mathrm{i}}^{\nu})^{\top} \nabla^2 F\Big(x_{\mathrm{i}}^{\nu} + \theta_2(x_{\mathrm{i}} - x_{\mathrm{i}}^{\nu}); x_{\mathrm{i}}^{\nu}\Big)(x_{\mathrm{i}} - x_{\mathrm{i}}^{\nu}).
\end{aligned} \tag{4.17}$$

Noting that $\nabla \widetilde{f}_{\mathrm{i}}(x_{\mathrm{i}}^{\nu}; x_{\mathrm{i}}^{\nu}) = \nabla f_{\mathrm{i}}(x_{\mathrm{i}}^{\nu})$ [Assumption 4.2.1(i)] and $\nabla \widetilde{F}_{\mathrm{i}}(x_{\mathrm{i}}^{\nu}; x_{\mathrm{i}}^{\nu}) = y_{\mathrm{i}}^{\nu}$, and anticipating $\|\nabla F(x_{\mathrm{i}}^{\nu}) - y_{\mathrm{i}}^{\nu}\| \to 0$ as $\nu \to \infty$ (see discussion below), it follows that $\widetilde{F}_{\mathrm{i}}$ approximates F asymptotically, up to the first order. A better match, is achieved when D_{i} is sufficiently small. One can then expect that, if the local functions are sufficiently similar (β is small), surrogates $\widetilde{f}_{\mathrm{i}}$ exploiting higher order information of f_{i}, such as (4.14), may be more effective than mere linearization. Our theoretical findings confirm the above intuition–see Sec. 4.2.4.

- *Consensus and gradient tracking steps (4.12c)-(4.12d)* In the consensus and tracking steps, the weights w_{ij}'s satisfy the following standard assumption.

Assumption 4.2.2. *The weight matrix $W \triangleq (w_{\mathrm{ij}})_{\mathrm{i,j}=1}^{m}$ has a sparsity pattern compliant with $\mathcal{G}$, that is*

4.2.2.1 $w_{\mathrm{ii}} > 0$, *for all* $\mathrm{i} = 1, \ldots, m$;

4.2.2.2 $w_{\mathrm{ij}} > 0$, *if* $(\mathrm{i}, \mathrm{j}) \in \mathcal{E}$; *and* $w_{\mathrm{ij}} = 0$ *otherwise;*

Furthermore, W is doubly stochastic, that is, $1^{\top} W = 1^{\top}$ *and* $W1 = 1$.

Several rules have been proposed in the literature compliant with Assumption 4.2.2, such as the Laplacian, the Metropolis-Hasting, and the maximum-degree weights rules [188].

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

$$\bar{y}^{\nu} \triangleq \frac{1}{m} \sum_{\mathrm{i}=1}^{m} y_{\mathrm{i}}^{\nu} \quad \text{and} \quad \overline{\nabla F_c}^{\nu} \triangleq \frac{1}{m} \sum_{\mathrm{i}=1}^{m} \nabla f_{\mathrm{i}}(x_{\mathrm{i}}^{\nu}). \tag{4.18}$$

Summing (4.12d) over $\mathrm{i} \in [m]$ and invoking the doubly stochasticity of W; we have

$$\bar{y}^{\nu+1} = \bar{y}^{\nu} + \overline{\nabla F_c}^{\nu+1} - \overline{\nabla F_c}^{\nu}. \tag{4.19}$$

Applying (4.19) inductively and using the initial condition $y_{\mathrm{i}}^{0} = \nabla f_{\mathrm{i}}(x_{\mathrm{i}}^{0})$, $\mathrm{i} \in [m]$, yield

$$\bar{y}^{\nu} = \overline{\nabla F_c}^{\nu}, \quad \forall \nu = 0, 1, \ldots. \tag{4.20}$$

That is, the average of all the y_i^ν's in the network is equal to that of the $\nabla f_i(x_i^\nu)$'s, at every iteration ν. Assuming that consensus on x_i^ν's and y_i^ν's is asymptotically achieved, that is, $\|x_i^\nu - x_j^\nu\| \underset{\nu\to\infty}{\longrightarrow} 0$ and $\|y_i^\nu - y_j^\nu\| \underset{\nu\to\infty}{\longrightarrow} 0$, $i \neq j$, (4.20) would imply the desired gradient tracking property $\|\nabla F(x_i^\nu) - y_i^\nu\| \to 0$ as $\nu \to \infty$, for all $i \in [m]$.

4.2.1 A special instance: SONATA on star-networks

Although the main focus of this thesis is the study of SONATA over meshed-networks, it is worth discussing here its special instance over star networks. Specifically, consider a star (unidirected) graph with m nodes, where one of them (the master node) connects with all the others (workers). The workers still own only one function f_i of the sum-cost F. Two common approaches developed in the literature to solve (4.1) in this setting are: (i) based upon receiving the gradients ∇f_i from the workers, the master solves (4.1) and broadcasts the updated vector variables to the workers; (ii) based upon receiving the full gradient ∇F and the current iterate from the master, all the workers solve locally an instance of (4.1) and send their outcomes to the master that averages them out, producing then the new iterate. Here we follow the latter approach; the algorithm is described in Algorithm 3, which corresponds to SONATA (up to a proper initialization), with weight matrix $W = [1,\, 0_{m,m-1}]\,[1/m,\, 0_{m,m-1}]^\top$.

Connection with existing schemes SONATA-star, employing linear surrogates [cf. (4.13)] and $\alpha = 1$, reduces to the proximal gradient algorithm. When the surrogates (4.14) are used (and still $\alpha = 1$), SONATA-star coincides with the DANE algorithm [140] if $G = 0$ and to the CEASE (with averaging) algorithm [142] if $G \neq 0$. Nevertheless, our convergence rates improve on those of DANE and CEASE–see Sec. 4.2.4.1.

4.2.2 Intermediate definitions

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

$$p^\nu \triangleq \sum_{i=1}^m \Big(U(x_i^\nu) - U(x^\star)\Big), \tag{4.21}$$

Algorithm 3: SONATA on Star-Networks (SONATA-Star)

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

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

`[S.1]` Each worker i evaluates $\nabla f_{\mathrm{i}}(x^\nu)$ and sends it to the master node;

`[S.2]` The master broadcasts $\nabla F(x^\nu) = 1/m \sum_{\mathrm{i}=1}^m \nabla f_{\mathrm{i}}(x^\nu)$ to the workers;

`[S.3]` Each worker i computes

$$\widehat{x}_{\mathrm{i}}^\nu \triangleq \underset{x_{\mathrm{i}} \in \mathcal{K}}{\operatorname{argmin}} \, \widetilde{f}_{\mathrm{i}}(x_{\mathrm{i}}; x^\nu) + \left(\nabla F(x^\nu) - \nabla f_{\mathrm{i}}(x_{\mathrm{i}}^\nu)\right)^\top (x_{\mathrm{i}} - x^\nu) + G(x_{\mathrm{i}}),$$

and sends $\widehat{x}_{\mathrm{i}}^\nu$ to the master;

`[S.4]` The master computes

$$x^{\nu+1} = x^\nu + \alpha \left(\frac{1}{m} \sum_{\mathrm{i}=1}^m \widehat{x}_{\mathrm{i}}^\nu - x^\nu \right),$$

and sends it back to the workers.

end

where $x^\star$ is the unique solution of Problem (4.1).

We stack the local variables and gradients in the column vectors

$$x^\nu \triangleq [x_1^{\nu\top}, \ldots, x_m^{\nu\top}]^\top, \; y^\nu \triangleq [y_1^{\nu\top}, \ldots, y_m^{\nu\top}]^\top, \; \nabla F_c^\nu \triangleq [\nabla f_1(x_1^\nu)^\top, \ldots, \nabla f_m(x_m^\nu)^\top]^\top. \tag{4.22}$$

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

$$x_\perp^\nu \triangleq x^\nu - 1_m \otimes \bar{x}^\nu \quad \text{and} \quad y_\perp^\nu \triangleq y^\nu - 1_m \otimes \bar{y}^\nu, \tag{4.23}$$

respectively, while the gradient tracking error is defined as

$$\delta^\nu \triangleq [\delta_1^{\nu\top}, \ldots, \delta_m^{\nu\top}]^\top, \quad \text{with} \quad \delta_{\mathrm{i}}^\nu \triangleq \nabla F(x_{\mathrm{i}}^\nu) - y_{\mathrm{i}}^\nu, \quad \mathrm{i} = 1, \ldots, m. \tag{4.24}$$

Recalling L_{i}, $\widetilde{L}_{\mathrm{i}}$, $\widetilde{\mu}_{\mathrm{i}}$, D_{i}^{ℓ} and D_{i} as given in Assumptions 4.1.1 and 4.2.1 and (4.16), we introduce the following algorithm-dependent parameters

$$\begin{aligned} \widetilde{\mu}_{\mathrm{mn}} &\triangleq \min_{\mathrm{i}\in[m]} \widetilde{\mu}_{\mathrm{i}}, \quad & \widetilde{L}_{\mathrm{mx}} &\triangleq \max_{\mathrm{i}\in[m]}, \widetilde{L}_{\mathrm{i}}, \\ D_{\mathrm{mn}}^{\ell} &\triangleq \min_{\mathrm{i}\in[m]} D_{\mathrm{i}}^{\ell}, \quad & D_{\mathrm{mx}} &\triangleq \max_{\mathrm{i}\in[m]} D_{\mathrm{i}}. \end{aligned} \tag{4.25}$$

Finally, given the weight matrix W, we define

$$\widehat{W} \triangleq W \otimes I_d, \quad \text{and} \quad J \triangleq \frac{1}{m} 1_m 1_m^{\top} \otimes I_d. \tag{4.26}$$

Under Assumptions 4.1.2 and 4.2.2, it is well known that (see, e.g., [189])

$$\rho \triangleq \sigma(\widehat{W} - J) < 1, \tag{4.27}$$

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

4.2.3 Linear convergence rate

Our proof of linear rate of SONATA passes through the following steps. **Step 1:** We begin showing that the optimality gap p^{ν} converges linearly up to an error of the order of $\mathcal{O}(\|x_{\perp}^{\nu}\|^2 + \|y_{\perp}^{\nu}\|^2)$, see Proposition 4.2.1. **Step 2** proves that $\|x_{\perp}^{\nu}\|$ and $\|y_{\perp}^{\nu}\|$ are also linearly convergent up to an error $\mathcal{O}(\|d^{\nu}\|)$, see Proposition 4.2.2. In **Step 3** we close the loop establishing $\|d^{\nu}\| = \mathcal{O}(\sqrt{p^{\nu}} + \|y_{\perp}^{\nu}\|)$, see Proposition 4.2.3. Finally, in **Step 4**, we properly chain together the above inequalities (cf. Proposition 4.2.4), so that linear rate is proved for the sequences $\{p^{\nu}\}$, $\{\|x_{\perp}^{\nu}\|^2\}$, $\{\|y_{\perp}^{\nu}\|^2\}$, and $\{\|d^{\nu}\|^2\}$–see Theorems 4.2.1 and 4.2.2. We will tacitly assume that Assumptions 4.1.1, 4.1.2, 4.2.1, and 4.2.2 are satisfied.

Step 1: p^{ν} converges linearly up to $\mathcal{O}(\|x_{\perp}^{\nu}\|^2+\|y_{\perp}^{\nu}\|^2)$

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

$$p^{\nu+1} \leq \sum_{\mathrm{i}=1}^{m}\sum_{\mathrm{j}=1}^{m} w_{\mathrm{ij}}\Big(U\Big(x_{\mathrm{j}}^{\nu+\frac{1}{2}}\Big) - U(x^{\star})\Big) = \sum_{\mathrm{i}=1}^{m}\Big(U(x_{\mathrm{i}}^{\nu+\frac{1}{2}}) - U(x^{\star})\Big). \tag{4.28}$$

We can now bound $U(x_{\mathrm{i}}^{\nu+\frac{1}{2}})$, regarding the local optimization (4.12a)-(4.12b) as a perturbed descent on the objective, whose perturbation is due to the tracking error δ^ν. In fact, Lemma 4.2.1 below shows that, for sufficiently small α, the local update (4.12b) will decrease the objective value U up to some error, related to δ_{i}^ν.

Lemma 4.2.1. *Let $\{x_{\mathrm{i}}^\nu\}$ be the sequence generated by SONATA; there holds:*

$$U(x_{\mathrm{i}}^{\nu+\frac{1}{2}}) \leq U(x_{\mathrm{i}}^\nu) - \alpha\left(\left(1-\frac{\alpha}{2}\right)\widetilde{\mu}_{\mathrm{i}} + \frac{\alpha}{2}\cdot D_{\mathrm{i}}^\ell\right)\|d_{\mathrm{i}}^\nu\|^2 + \alpha\|d_{\mathrm{i}}^\nu\|\|\delta_{\mathrm{i}}^\nu\|, \tag{4.29}$$

with D_{i}^ℓ and δ_{i}^ν are defined in (4.16) and (4.24), respectively.

Proof. Consider the Taylor expansion of F:

$$\begin{aligned} F(x_{\mathrm{i}}^{\nu+\frac{1}{2}}) &= F(x_{\mathrm{i}}^\nu) + \nabla F(x_{\mathrm{i}}^\nu)^\top(\alpha d_{\mathrm{i}}^\nu) + (\alpha d_{\mathrm{i}}^\nu)^\top H(\alpha d_{\mathrm{i}}^\nu), \\ &\overset{(4.24)}{=} F(x_{\mathrm{i}}^\nu) + \left(\delta_{\mathrm{i}}^\nu\right)^\top(\alpha d_{\mathrm{i}}^\nu) + \left(y_{\mathrm{i}}^\nu\right)^\top(\alpha d_{\mathrm{i}}^\nu) + (\alpha d_{\mathrm{i}}^\nu)^\top H(\alpha d_{\mathrm{i}}^\nu), \end{aligned} \tag{4.30}$$

where $H \triangleq \int_0^1 (1-\theta)\nabla^2 F(\theta x_{\mathrm{i}}^{\nu+\frac{1}{2}} + (1-\theta)x_{\mathrm{i}}^\nu)d\theta$.

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

$$G(x_{\mathrm{i}}^\nu) - G(\widehat{x}_{\mathrm{i}}^\nu) \geq (d_{\mathrm{i}}^\nu)^\top\left(\nabla\widetilde{f}_{\mathrm{i}}(\widehat{x}_{\mathrm{i}}^\nu; x_{\mathrm{i}}^\nu) + y_{\mathrm{i}}^\nu - \nabla f_{\mathrm{i}}(x_{\mathrm{i}}^\nu)\right) = (d_{\mathrm{i}}^\nu)^\top\left(y_{\mathrm{i}}^\nu + \widetilde{H}_{\mathrm{i}} d_{\mathrm{i}}^\nu\right), \tag{4.31}$$

where the equality follows from $\nabla\widetilde{f}_{\mathrm{i}}(x_{\mathrm{i}}^\nu; x_{\mathrm{i}}^\nu) = \nabla f_{\mathrm{i}}(x_{\mathrm{i}}^\nu)$ and the integral form of the mean value theorem. Substituting (4.31) in (4.30) and using the convexity of G yield

$$\begin{aligned} &F(x_{\mathrm{i}}^{\nu+\frac{1}{2}}) \\ &\leq F(x_{\mathrm{i}}^\nu) + (\delta_{\mathrm{i}}^\nu)^\top(\alpha d_{\mathrm{i}}^\nu) + (\alpha d_{\mathrm{i}}^\nu)^\top H(\alpha d_{\mathrm{i}}^\nu) + \alpha\left(G(x_{\mathrm{i}}^\nu) - G(\widehat{x}_{\mathrm{i}}^\nu) - (d_{\mathrm{i}}^\nu)^\top\widetilde{H}_{\mathrm{i}} d_{\mathrm{i}}^\nu\right) \\ &\leq F(x_{\mathrm{i}}^\nu) + (\delta_{\mathrm{i}}^\nu)^\top(\alpha d_{\mathrm{i}}^\nu) + \alpha\left(-(d_{\mathrm{i}}^\nu)^\top\widetilde{H}_{\mathrm{i}} d_{\mathrm{i}}^\nu + (\alpha d_{\mathrm{i}}^\nu)^\top H(d_{\mathrm{i}}^\nu)\right) + G(x_{\mathrm{i}}^\nu) - G(x_{\mathrm{i}}^{\nu+\frac{1}{2}}). \end{aligned} \tag{4.32}$$

It remains to bound $\alpha H - \widetilde{H}_{\mathrm{i}}$. We proceed as follows:

$$\begin{aligned}
&\alpha H - \widetilde{H}_{\mathrm{i}} \\
&= \alpha \int_0^1 (1-\theta)\nabla^2 F(\theta x_{\mathrm{i}}^{\nu+\frac{1}{2}} + (1-\theta)x_{\mathrm{i}}^{\nu})d\theta - \int_0^1 \nabla^2 \widetilde{f}_{\mathrm{i}}(\theta\widehat{x}_{\mathrm{i}}^{\nu} + (1-\theta)x_{\mathrm{i}}^{\nu}; x_{\mathrm{i}}^{\nu})d\theta \\
&\overset{(4.12b)}{=} \int_0^{\alpha} (1-\theta/\alpha)\nabla^2 F(\theta\widehat{x}_{\mathrm{i}}^{\nu} + (1-\theta)x_{\mathrm{i}}^{\nu})d\theta - \int_0^1 \nabla^2 \widetilde{f}_{\mathrm{i}}(\theta\widehat{x}_{\mathrm{i}}^{\nu} + (1-\theta)x_{\mathrm{i}}^{\nu}; x_{\mathrm{i}}^{\nu})d\theta \\
&\overset{(a)}{\preceq} -\int_0^{\alpha} (1-\theta/\alpha)\cdot(D_{\mathrm{i}}^{\ell})\, I\, d\theta - \int_0^{\alpha} (\theta/\alpha)\nabla^2 \widetilde{f}_{\mathrm{i}}(\theta\widehat{x}_{\mathrm{i}} + (1-\theta)x_{\mathrm{i}}^{\nu}; x_{\mathrm{i}}^{\nu})d\theta \\
&\quad - \int_{\alpha}^1 \nabla^2 \widetilde{f}_{\mathrm{i}}(\theta\,\widehat{x}_{\mathrm{i}}^{\nu} + (1-\theta)\,x_{\mathrm{i}}^{\nu}; x_{\mathrm{i}}^{\nu})d\theta \\
&\overset{(b)}{\preceq} -\frac{1}{2}\alpha\,(D_{\mathrm{i}}^{\ell})\, I - \left(1-\frac{\alpha}{2}\right)\tilde{\mu}_{\mathrm{i}}\, I,
\end{aligned} \tag{4.33}$$

where in (a) we used $\nabla^2 F(\theta\widehat{x}_{\mathrm{i}}^{\nu} + (1-\theta)x_{\mathrm{i}}^{\nu}) \preceq -(D_{\mathrm{i}}^{\ell})I + \nabla^2 \widetilde{f}_{\mathrm{i}}(\theta\widehat{x}_{\mathrm{i}}^{\nu} + (1-\theta)x_{\mathrm{i}}^{\nu}; x_{\mathrm{i}}^{\nu})$ [cf. (4.16)] while (b) follows from Assumption 4.2.1(iii). Substituting (4.33) into (4.32) completes the proof □

We can now substitute (4.29) into (4.28) and get

$$p^{\nu+1} \leq p^{\nu} + \sum_{\mathrm{i}=1}^{m} \left\{ \alpha \|d_{\mathrm{i}}^{\nu}\| \|\delta_{\mathrm{i}}^{\nu}\| - \alpha\left(1-\frac{\alpha}{2}\right)\tilde{\mu}_{\mathrm{i}}\|d_{\mathrm{i}}^{\nu}\|^2 - \frac{D_{\mathrm{i}}^{\ell}}{2}\alpha^2\|d_{\mathrm{i}}^{\nu}\|^2 \right\} \tag{4.34a}$$

$$\overset{(a)}{\leq} p^{\nu} - \left(\left(1-\frac{\alpha}{2}\right)\tilde{\mu}_{\mathrm{mn}} + \frac{\alpha D_{\mathrm{mn}}^{\ell}}{2} - \frac{1}{2}\epsilon_{opt}\right)\alpha\|d^{\nu}\|^2 + \frac{1}{2}\epsilon_{opt}^{-1}\,\alpha\cdot\|\delta^{\nu}\|^2, \tag{4.34b}$$

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

$$\left(1-\frac{\alpha}{2}\right)\tilde{\mu}_{\mathrm{mn}} + \frac{\alpha D_{\mathrm{mn}}^{\ell}}{2} - \frac{1}{2}\epsilon_{opt} > 0; \tag{4.35}$$

and D_{mn}^{ℓ} is defined in (4.25).

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

Lemma 4.2.2. *The following lower bound holds for $\|d^{\nu}\|^2$:*

$$\alpha\,\|d^{\nu}\|^2 \geq \frac{\mu}{D_{\mathrm{mx}}^2}\left(p^{\nu+1} - (1-\alpha)p^{\nu} - \frac{\alpha}{\mu}\|\delta^{\nu}\|^2\right), \tag{4.36}$$

where D_{mx} is defined in (4.25).

Proof. Invoking the optimality condition of $\widehat{x}_{\mathrm{i}}^{\nu}$, yields

$$G(x^{\star}) - G(\widehat{x}_{\mathrm{i}}^{\nu}) \geq -(x^{\star} - \widehat{x}_{\mathrm{i}}^{\nu})^{\top}\Big(\nabla \widetilde{f}_{\mathrm{i}}(\widehat{x}_{\mathrm{i}}^{\nu}; x_{\mathrm{i}}^{\nu}) + y_{\mathrm{i}}^{\nu} - \nabla f_{\mathrm{i}}(x_{\mathrm{i}}^{\nu})\Big). \tag{4.37}$$

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

$$\begin{aligned}
U(x^{\star}) &\geq U(\widehat{x}_{\mathrm{i}}^{\nu}) + G(x^{\star}) - G(\widehat{x}_{\mathrm{i}}^{\nu}) + \nabla F(\widehat{x}_{\mathrm{i}}^{\nu})^{\top}(x^{\star} - \widehat{x}_{\mathrm{i}}^{\nu}) + \frac{\mu}{2}\|x^{\star} - \widehat{x}_{\mathrm{i}}^{\nu}\|^2 \\
&\overset{(4.37)}{\geq} U(\widehat{x}_{\mathrm{i}}^{\nu}) + \Big(\nabla F(\widehat{x}_{\mathrm{i}}^{\nu}) - \nabla \widetilde{f}_{\mathrm{i}}(\widehat{x}_{\mathrm{i}}^{\nu}; x_{\mathrm{i}}^{\nu}) - \big(y_{\mathrm{i}}^{\nu} - \nabla f_{\mathrm{i}}(x_{\mathrm{i}}^{\nu})\big)\Big)^{\top}(x^{\star} - \widehat{x}_{\mathrm{i}}^{\nu}) + \frac{\mu}{2}\|x^{\star} - \widehat{x}_{\mathrm{i}}^{\nu}\|^2 \\
&= U(\widehat{x}_{\mathrm{i}}^{\nu}) + \frac{\mu}{2}\Big\|x^{\star} - \widehat{x}_{\mathrm{i}}^{\nu} + \frac{1}{\mu}\Big(\nabla F(\widehat{x}_{\mathrm{i}}^{\nu}) - \nabla \widetilde{f}_{\mathrm{i}}(\widehat{x}_{\mathrm{i}}^{\nu}; x_{\mathrm{i}}^{\nu}) - \big(y_{\mathrm{i}}^{\nu} - \nabla f_{\mathrm{i}}(x_{\mathrm{i}}^{\nu})\big)\Big)\Big\|^2 \\
&\quad - \frac{1}{2\mu}\Big\|\nabla F(\widehat{x}_{\mathrm{i}}^{\nu}) - \nabla \widetilde{f}_{\mathrm{i}}(\widehat{x}_{\mathrm{i}}^{\nu}; x_{\mathrm{i}}^{\nu}) - \big(y_{\mathrm{i}}^{\nu} - \nabla f_{\mathrm{i}}(x_{\mathrm{i}}^{\nu})\big)\Big\|^2 \\
&\geq U(\widehat{x}_{\mathrm{i}}^{\nu}) - \frac{1}{2\mu}\Big\|\nabla F(\widehat{x}_{\mathrm{i}}^{\nu}) \pm \nabla F(x_{\mathrm{i}}^{\nu}) - \nabla \widetilde{f}_{\mathrm{i}}(\widehat{x}_{\mathrm{i}}^{\nu}; x_{\mathrm{i}}^{\nu}) - \big(y_{\mathrm{i}}^{\nu} - \nabla f_{\mathrm{i}}(x_{\mathrm{i}}^{\nu})\big)\Big\|^2 \\
&\geq U(\widehat{x}_{\mathrm{i}}^{\nu}) - \frac{1}{\mu}\Big\|\nabla F(\widehat{x}_{\mathrm{i}}^{\nu}) - \nabla F(x_{\mathrm{i}}^{\nu}) + \nabla f_{\mathrm{i}}(x_{\mathrm{i}}^{\nu}) - \nabla \widetilde{f}_{\mathrm{i}}(\widehat{x}_{\mathrm{i}}^{\nu}; x_{\mathrm{i}}^{\nu})\Big\|^2 - \frac{1}{\mu}\|\delta_{\mathrm{i}}^{\nu}\|^2 \\
&= U(\widehat{x}_{\mathrm{i}}^{\nu}) - \frac{1}{\mu}\Big\|\int_0^1 \Big(\nabla^2 F(\theta\widehat{x}_{\mathrm{i}}^{\nu} + (1-\theta)x_{\mathrm{i}}^{\nu}) - \nabla^2 \widetilde{f}_{\mathrm{i}}(\theta\widehat{x}_{\mathrm{i}}^{\nu} + (1-\theta)x_{\mathrm{i}}^{\nu}; x_{\mathrm{i}}^{\nu})\Big)\,(d_{\mathrm{i}}^{\nu})\,\mathrm{d}\theta\Big\|^2 - \frac{1}{\mu}\|\delta_{\mathrm{i}}^{\nu}\|^2 \\
&\geq U(\widehat{x}_{\mathrm{i}}^{\nu}) - \frac{D_{\mathrm{i}}^2}{\mu}\big\|d_{\mathrm{i}}^{\nu}\big\|^2 - \frac{1}{\mu}\|\delta_{\mathrm{i}}^{\nu}\|^2.
\end{aligned}$$

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

$$\|d^{\nu}\|^2 \geq \frac{\mu}{D_{\mathrm{mx}}^2}\left(\sum_{\mathrm{i}=1}^{m}\Big(U(\widehat{x}_{\mathrm{i}}^{\nu}) - U(x^{\star})\Big) - \frac{1}{\mu}\|\delta^{\nu}\|^2\right). \tag{4.38}$$

Using (4.28) in conjunction with $U(x_{\mathrm{i}}^{\nu+\frac{1}{2}}) \leq \alpha U(\widehat{x}_{\mathrm{i}}^{\nu}) + (1-\alpha)U(x_{\mathrm{i}}^{\nu})$ leads to

$$\alpha \sum_{\mathrm{i}=1}^{m}\left(U(\widehat{x}_{\mathrm{i}}^{\nu}) - U(x^{\star})\right) \geq p^{\nu+1} - (1-\alpha)p^{\nu}. \tag{4.39}$$

Combining (4.38) with (4.39) provides the desired result (4.36). □

As last step, we upper bound $\|\delta^{\nu}\|^2$ in (4.34) in terms of the consensus errors $\|x_{\perp}^{\nu}\|^2$ and $\|y_{\perp}^{\nu}\|^2$.

Lemma 4.2.3. *The following upper bound holds for the tracking error* $\|\delta^\nu\|^2$*:*

$$\|\delta^\nu\|^2 \leq 4L_{\mathrm{mx}}^2\|x_\perp^\nu\|^2 + 2\|y_\perp^\nu\|^2, \tag{4.40}$$

where L_{mx} *is defined in* (4.5).

Proof.

$$\begin{aligned}
\|\delta^\nu\|^2 &\overset{(4.24)}{=} \sum_{\mathrm{i}=1}^m \|\nabla F(x_\mathrm{i}^\nu) \pm \bar{y}^\nu - y_\mathrm{i}^\nu\|^2 \\
&\overset{(4.18)}{=} \frac{1}{m^2}\sum_{\mathrm{i}=1}^m \Big\| \sum_{\mathrm{j}=1}^m \nabla f_\mathrm{j}(x_\mathrm{i}^\nu) - \sum_{\mathrm{j}=1}^m \nabla f_\mathrm{j}(x_\mathrm{j}^\nu) + m\cdot\bar{y}^\nu - m\cdot y_\mathrm{i}^\nu\Big\|^2 \\
&\overset{(4.2),(4.5)}{\leq} \frac{1}{m^2}\sum_{\mathrm{i}=1}^m \left(2m\sum_{\mathrm{j}=1}^m L_{\mathrm{mx}}^2\|x_\mathrm{i}^\nu - x_\mathrm{j}^\nu\|^2 + 2m^2\|\bar{y}^\nu - y_\mathrm{i}^\nu\|^2\right) \\
&= 4L_{\mathrm{mx}}^2\|x_\perp^\nu\|^2 + 2\|y_\perp^\nu\|^2.
\end{aligned}$$

□

We are ready to prove the linear convergence of the optimality gap up to consensus errors. The result is summarized in Proposition 4.2.1 below. The proof follows readily multiplying (4.34) and (4.36) 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 $\|d^\nu\|$, and using (4.40) to bound $\|\delta^\nu\|^2$.

Proposition 4.2.1. *The optimality gap* p^ν *[cf. (4.21)] satisfies*

$$p^{\nu+1} \leq \sigma(\alpha)\cdot p^\nu + \eta(\alpha)\cdot\left(4L_{\mathrm{mx}}^2\|x_\perp^\nu\|^2 + 2\|y_\perp^\nu\|^2\right) \tag{4.41}$$

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

$$\sigma(\alpha) \triangleq 1 - \alpha\,\frac{\left(1-\frac{\alpha}{2}\right)\widetilde{\mu}_{\mathrm{mn}} + \frac{D_{\mathrm{mn}}^\ell}{2}\alpha - \frac{1}{2}\epsilon_{opt}}{\frac{D_{\mathrm{mx}}^2}{\mu} + \left(1-\frac{\alpha}{2}\right)\widetilde{\mu}_{\mathrm{mn}} + \frac{D_{\mathrm{mn}}^\ell}{2}\alpha - \frac{1}{2}\epsilon_{opt}}, \tag{4.42}$$

$$\eta(\alpha) \triangleq \frac{\frac{1}{2}\epsilon_{opt}^{-1}\alpha\cdot\frac{D_{\mathrm{mx}}^2}{\mu} + \frac{\alpha}{\mu}\cdot\left(\left(1-\frac{\alpha}{2}\right)\widetilde{\mu}_{\mathrm{mn}} + \frac{D_{\mathrm{mn}}^\ell}{2}\alpha - \frac{1}{2}\epsilon_{opt}\right)}{\frac{D_{\mathrm{mx}}^2}{\mu} + \left(1-\frac{\alpha}{2}\right)\widetilde{\mu}_{\mathrm{mn}} + \frac{D_{\mathrm{mn}}^\ell}{2}\alpha - \frac{1}{2}\epsilon_{opt}}; \tag{4.43}$$

ϵ_{opt} *satisfies* (4.35)*; and* L_{mx} *and* $\widetilde{\mu}_{\mathrm{mn}}$, D_{mn}^ℓ, D_{mx} *are defined in* (4.5) *and* (4.25)*, respectively.*

Step 2: $\|x_\perp^\nu\|$ and $\|y_\perp^\nu\|$ linearly converge up to $\mathcal{O}(\|d^\nu\|)$ We upper bound $\|x_\perp^\nu\|$ and $\|y_\perp^\nu\|$ in terms of $\|d^\nu\|$. We begin rewriting the SONATA algorithm (4.12a)-(4.12d) in vector-matrix form; using (4.22) and (4.26), we have

$$x^{\nu+1} = \widehat{W}(x^\nu + \alpha d^\nu) \tag{4.44a}$$

$$y^{\nu+1} = \widehat{W}(y^\nu + \nabla F_c^{\nu+1} - \nabla F_c^\nu). \tag{4.44b}$$

Noting that $x_\perp^\nu = (I - J)x^\nu$ [similarly, $y_\perp^\nu = (I - J)y^\nu$] and $(I - J)\widehat{W} = \widehat{W} - J$ (due to the doubly stochasticity of W), it follows from (4.44) that

$$x_\perp^{\nu+1} = (\widehat{W} - J)(x_\perp^\nu + \alpha d^\nu) \tag{4.45}$$

$$y_\perp^{\nu+1} = (\widehat{W} - J)(y_\perp^\nu + \nabla F_c^{\nu+1} - \nabla F_c^\nu). \tag{4.46}$$

Using (4.45)-(4.46), Proposition 4.2.2 below establishes linear convergence of the consensus errors $x_\perp^\nu$ and $y_\perp^\nu$, up to a perturbation.

Proposition 4.2.2. *There holds:*

$$\|x_\perp^{\nu+1}\| \le \rho\|x_\perp^\nu\| + \alpha\rho\|d^\nu\|, \tag{4.47a}$$

$$\|y_\perp^{\nu+1}\| \le \rho\|y_\perp^\nu\| + 2L_{\rm mx}\rho\|x_\perp^\nu\| + \alpha L_{\rm mx}\rho\|d^\nu\|, \tag{4.47b}$$

with ρ and $L_{\rm mx}$ defined in (4.27) and (4.5), respectively.

Proof. We prove next (4.47b); (4.47a) follows readily from (4.45). Using (4.44a), (4.46), and the Lipschitz continuity of $\nabla f_{\rm i}$ [cf. (4.2)], we can bound $\|y_\perp^{\nu+1}\|$ as

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

where in the last inequality we used $\|W\| \le 1$. □

Step 3: $\|d^\nu\| = \mathcal{O}(\sqrt{p^\nu} + \|y_\perp^\nu\|)$ **(closing the loop)** Given the inequalities in Propositions 4.2.1 and 4.2.2, to close the loop, one needs to link $\|d^\nu\|$ to the quantities in the aforementioned inequalities, which is done next.

Proposition 4.2.3. *The following upper bound holds for* $\|d^\nu\|$*:*

$$\|d^\nu\|^2 \leq \frac{6}{\mu}\left(\left(\frac{D_{\mathrm{mx}}}{\widetilde{\mu}_{\mathrm{mn}}}+1\right)^2 + \frac{4L_{\mathrm{mx}}^2}{\widetilde{\mu}_{\mathrm{mn}}^2}\right)p^\nu + \frac{3}{\widetilde{\mu}_{\mathrm{mn}}^2}\|y_\perp^\nu\|^2. \tag{4.48}$$

where L_{mx} *and* $\widetilde{L}_{\mathrm{mx}}$, $\widetilde{\mu}_{\mathrm{mn}}$, D_{mx} *are defined in* (4.5) *and* (4.25)*, respectively.*

Proof. By optimality of $\widehat{x}_{\mathrm{i}}^\nu$ and $x^\star$ we have

$$\begin{aligned}\left(\nabla\widetilde{f}_{\mathrm{i}}(\widehat{x}_{\mathrm{i}}^\nu; x_{\mathrm{i}}^\nu) + y_{\mathrm{i}}^\nu - \nabla f_{\mathrm{i}}(x_{\mathrm{i}}^\nu)\right)^\top(x^\star - \widehat{x}_{\mathrm{i}}^\nu) + G(x^\star) - G(\widehat{x}_{\mathrm{i}}^\nu) &\geq 0,\\ \nabla F(x^\star)^\top(\widehat{x}_{\mathrm{i}}^\nu - x^\star) + G(\widehat{x}_{\mathrm{i}}^\nu) - G(x^\star) &\geq 0.\end{aligned}$$

Summing the two inequalities above yields

$$\begin{aligned}0 &\leq \left(\nabla F(x^\star) - y_{\mathrm{i}}^\nu + \nabla f_{\mathrm{i}}(x_{\mathrm{i}}^\nu) - \nabla\widetilde{f}_{\mathrm{i}}(\widehat{x}_{\mathrm{i}}^\nu; x_{\mathrm{i}}^\nu) \pm \bar{y}^\nu\right)^\top(\widehat{x}_{\mathrm{i}}^\nu - x^\star)\\
&\leq \left(\nabla F(x^\star) - \frac{1}{m}\sum_{\mathrm{j}=1}^m \nabla f_{\mathrm{j}}(x_{\mathrm{j}}^\nu) + \nabla f_{\mathrm{i}}(x_{\mathrm{i}}^\nu) - \nabla\widetilde{f}_{\mathrm{i}}(\widehat{x}_{\mathrm{i}}^\nu; x_{\mathrm{i}}^\nu)\right)^\top(\widehat{x}_{\mathrm{i}}^\nu - x^\star)\\
&\quad + \|\bar{y}^\nu - y_{\mathrm{i}}^\nu\|\|\widehat{x}_{\mathrm{i}}^\nu - x^\star\|\\
&\leq \left(\nabla F(x^\star) - \nabla F(x_{\mathrm{i}}^\nu) + \nabla f_{\mathrm{i}}(x_{\mathrm{i}}^\nu) - \nabla\widetilde{f}_{\mathrm{i}}(\widehat{x}_{\mathrm{i}}^\nu; x_{\mathrm{i}}^\nu)\right)^\top(\widehat{x}_{\mathrm{i}}^\nu - x^\star)\\
&\quad + \|\bar{y}^\nu - y_{\mathrm{i}}^\nu\|\|\widehat{x}_{\mathrm{i}}^\nu - x^\star\| + \left\|\nabla F(x_{\mathrm{i}}^\nu) - \frac{1}{m}\sum_{\mathrm{j}=1}^m \nabla f_{\mathrm{j}}(x_{\mathrm{j}}^\nu)\right\|\left\|\widehat{x}_{\mathrm{i}}^\nu - x^\star\right\|\end{aligned}$$

$$\begin{aligned}&\leq \left(\nabla F(x^\star) - \nabla F(x_{\mathrm{i}}^\nu) + \nabla f_{\mathrm{i}}(x_{\mathrm{i}}^\nu) \pm \nabla\widetilde{f}_{\mathrm{i}}(x^\star; x_{\mathrm{i}}^\nu) - \nabla\widetilde{f}_{\mathrm{i}}(\widehat{x}_{\mathrm{i}}^\nu; x_{\mathrm{i}}^\nu)\right)^\top(\widehat{x}_{\mathrm{i}}^\nu - x^\star)\\
&\quad + \|\bar{y}^\nu - y_{\mathrm{i}}^\nu\|\|\widehat{x}_{\mathrm{i}}^\nu - x^\star\| + \left(\frac{1}{m}\sum_{\mathrm{j}=1}^m L_{\mathrm{j}}\left\|x_{\mathrm{i}}^\nu - x_{\mathrm{j}}^\nu\right\|\right)\left\|\widehat{x}_{\mathrm{i}}^\nu - x^\star\right\|\\
&\leq \left(\int_0^1 \left(\nabla^2 F(\theta x^\star + (1-\theta)x_{\mathrm{i}}^\nu) - \nabla^2\widetilde{f}_{\mathrm{i}}\left(\theta x^\star + (1-\theta)x_{\mathrm{i}}^\nu; x_{\mathrm{i}}^\nu\right)\right)(x^\star - x_{\mathrm{i}}^\nu)\,\mathrm{d}\theta\right)^\top(\widehat{x}_{\mathrm{i}}^\nu - x^\star)\\
&\quad - \widetilde{\mu}_{\mathrm{i}}\left\|\widehat{x}_{\mathrm{i}}^\nu - x^\star\right\|^2 + \|\bar{y}^\nu - y_{\mathrm{i}}^\nu\|\|\widehat{x}_{\mathrm{i}}^\nu - x^\star\| + \left(\frac{1}{m}\sum_{\mathrm{j}=1}^m L_{\mathrm{j}}\left\|x_{\mathrm{i}}^\nu - x_{\mathrm{j}}^\nu\right\|\right)\left\|\widehat{x}_{\mathrm{i}}^\nu - x^\star\right\|\end{aligned}$$

$$\leq D_{\mathrm{i}}\left\|x^\star - x_{\mathrm{i}}^\nu\right\|\left\|\widehat{x}_{\mathrm{i}}^\nu - x^\star\right\| - \widetilde{\mu}_{\mathrm{i}}\left\|\widehat{x}_{\mathrm{i}}^\nu - x^\star\right\|^2 + \|\bar{y}^\nu - y_{\mathrm{i}}^\nu\|\|\widehat{x}_{\mathrm{i}}^\nu - x^\star\| + \left(\frac{1}{m}\sum_{\mathrm{j}=1}^m L_{\mathrm{j}}\left\|x_{\mathrm{i}}^\nu - x_{\mathrm{j}}^\nu\right\|\right)\left\|\widehat{x}_{\mathrm{i}}^\nu - x^\star\right\|.$$

Rearranging terms and using the reverse triangle inequality we obtain the following bound for $\|d_{\mathrm{i}}^\nu\|$:

$$D_{\mathrm{i}}\left\|x^\star - x_{\mathrm{i}}^\nu\right\| + \|\bar{y}^\nu - y_{\mathrm{i}}^\nu\| + \left(\frac{1}{m}\sum_{\mathrm{j}=1}^m L_{\mathrm{j}}\left\|x_{\mathrm{i}}^\nu - x_{\mathrm{j}}^\nu\right\|\right) \geq \widetilde{\mu}_{\mathrm{i}}\left\|\widehat{x}_{\mathrm{i}}^\nu - x^\star\right\| \geq \widetilde{\mu}_{\mathrm{i}}\left(\|d_{\mathrm{i}}^\nu\| - \|x^\star - x_{\mathrm{i}}^\nu\|\right). \tag{4.49}$$

Therefore,

$$\begin{aligned}\|d_{\mathrm{i}}^\nu\|^2 \leq\ & 3\left(\frac{D_{\mathrm{i}}}{\widetilde{\mu}_{\mathrm{i}}} + 1\right)^2\left\|x^\star - x_{\mathrm{i}}^\nu\right\|^2 + \frac{3}{\widetilde{\mu}_{\mathrm{i}}^2}\|\bar{y}^\nu - y_{\mathrm{i}}^\nu\|^2 + \frac{3}{\widetilde{\mu}_{\mathrm{i}}^2}\left(\frac{1}{m}\sum_{\mathrm{j}=1}^m L_{\mathrm{j}}\left\|x_{\mathrm{i}}^\nu - x_{\mathrm{j}}^\nu\right\|\right)^2 \\ \leq\ & 3\left(\frac{D_{\mathrm{i}}}{\widetilde{\mu}_{\mathrm{i}}} + 1\right)^2\left\|x^\star - x_{\mathrm{i}}^\nu\right\|^2 + \frac{3}{\widetilde{\mu}_{\mathrm{i}}^2}\|\bar{y}^\nu - y_{\mathrm{i}}^\nu\|^2 + \frac{6L_{\mathrm{mx}}^2}{\widetilde{\mu}_{\mathrm{i}}^2 m}\left(\sum_{\mathrm{j}=1}^m \|x_{\mathrm{j}}^\nu - x^\star\|^2 + m\|x_{\mathrm{i}}^\nu - x^\star\|^2\right).\end{aligned}$$

Summing over $\mathrm{i} = 1, \ldots, m$, yields

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

□

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

We are now ready to prove linear rate of the SONATA algorithm. We build on the following intermediate result, introduced in [126].

Lemma 4.2.4. *Given the sequence $\{s^\nu\}$, define the transformations*

$$S^K(z) \triangleq \max_{\nu=0,\ldots,K}|s^\nu|z^{-\nu} \quad \textit{and} \quad S(z) \triangleq \sup_{\nu\in\mathbb{N}}|s^\nu|z^{-\nu}, \tag{4.50}$$

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 (4.41), (4.47) and (4.48) so that Lemma 4.2.4 can be applied to the sequences $\{p^\nu\}$, $\{\|x_\perp^\nu\|^2\}$, $\{\|y_\perp^\nu\|^2\}$ and $\{\|d^\nu\|^2\}$, establishing thus their linear convergence.

Proposition 4.2.4. *Let $P^K(z)$, $X_\perp^K(z)$, $Y_\perp^K(z)$ and $D^K(z)$ denote the transformation (4.50) applied to the sequences $\{p^\nu\}$, $\{\|x_\perp^\nu\|^2\}$, $\{\|y_\perp^\nu\|^2\}$ and $\{\|d^\nu\|^2\}$, respectively. Given the constants $\sigma(\alpha)$ and $\eta(\alpha)$ (defined in Proposition 4.2.1) and the free parameters $\epsilon_x, \epsilon_y > 0$ (to be determined), the following hold*

$$P^K(z) \leq G_P(\alpha, z) \cdot \Big(4L_{\text{mx}}^2 X_\perp^K(z) + 2Y_\perp^K(z)\Big) + \omega_p, \tag{4.51a}$$
$$X_\perp^K(z) \leq G_X(z) \cdot \rho^2\alpha^2 D^K(z) + \omega_x, \tag{4.51b}$$
$$Y_\perp^K(z) \leq G_Y(z) \cdot 8L_{\text{mx}}^2\rho^2 X_\perp^K(z) + G_Y(z) \cdot 2L_{\text{mx}}^2\rho^2\alpha^2 D^K(z) + \omega_y, \tag{4.51c}$$
$$D^K(z) \leq C_1 \cdot P^K(z) + C_2 \cdot Y_\perp^K(z), \tag{4.51d}$$

for all

$$z \in \Big(\max\{\sigma(\alpha), \rho^2(1+\epsilon_x), \rho^2(1+\epsilon_y)\}, 1\Big), \tag{4.52}$$

where

$$G_P(\alpha, z) \triangleq \frac{\eta(\alpha)}{z - \sigma(\alpha)}, \qquad \omega_p \triangleq \frac{z}{z - \sigma(\alpha)} \cdot p^0 \tag{4.53a}$$
$$G_X(z) \triangleq \frac{(1+\epsilon_x^{-1})}{z - \rho^2(1+\epsilon_x)}, \qquad \omega_x \triangleq \frac{z}{z - \rho^2(1+\epsilon_x)} \cdot \|x_\perp^0\|^2, \tag{4.53b}$$
$$G_Y(z) \triangleq \frac{(1+\epsilon_y^{-1})}{z - \rho^2(1+\epsilon_y)}, \qquad \omega_y \triangleq \frac{z}{z - \rho^2(1+\epsilon_y)} \cdot \|y_\perp^0\|^2, \tag{4.53c}$$
$$C_1 \triangleq \frac{6}{\mu}\left(\left(\frac{D_{\text{mx}}}{\widetilde{\mu}_{\text{mn}}} + 1\right)^2 + \frac{4L_{\text{mx}}^2}{\widetilde{\mu}_{\text{mn}}^2}\right), \qquad C_2 \triangleq \frac{4}{\widetilde{\mu}_{\text{mn}}^2}. \tag{4.53d}$$

Proof. Squaring (4.47) and using Young's inequality yield

$$\begin{aligned}
\|x_\perp^{\nu+1}\|^2 &\leq \rho^2(1+\epsilon_x)\|x_\perp^\nu\|^2 + \rho^2(1+\epsilon_x^{-1})\alpha^2\|d^\nu\|^2 \\
\|y_\perp^{\nu+1}\|^2 &\leq \rho^2(1+\epsilon_y)\|y_\perp^\nu\|^2 + \rho^2(1+\epsilon_y^{-1})\Big(8L_{\text{mx}}^2\|x_\perp^\nu\|^2 + 2\alpha^2 L_{\text{mx}}^2\|d^\nu\|^2\Big),
\end{aligned} \tag{4.54}$$

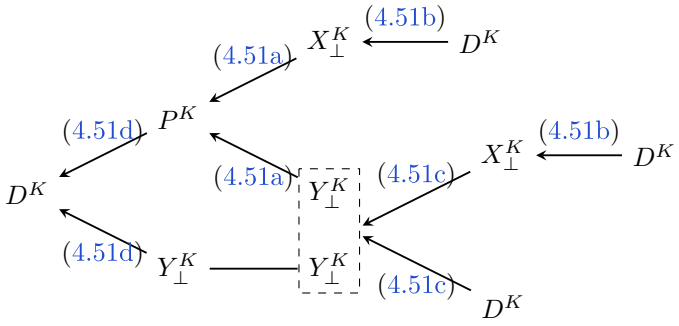


Figure 4.2. Chain of the inequalities in Proposition 4.2.4 leading to (4.55).

for arbitrary $\epsilon_x, \epsilon_y > 0$. The proof is completed by taking the maximum of both sides of (4.41), (4.48), and (4.54) 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 $z \in (0,1)$. □

Chaining the inequalities in Proposition 4.2.4 in the way shown in Fig. 4.2, we can bound $D^K(z)$ as (see Appendix 4.6.1 for the proof)

$$D^K(z) \leq \mathcal{P}(\alpha, z) \cdot D^K(z) + \mathcal{R}(\alpha, z), \tag{4.55}$$

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

$$\begin{aligned} \mathcal{P}(\alpha, z) \triangleq\ & G_P(\alpha, z) \cdot G_X(z) \cdot C_1 \cdot 4L_{\text{mx}}^2 \cdot \rho^2 \cdot \alpha^2 \\ & + (G_P(\alpha, z) \cdot 2C_1 + C_2) \cdot G_Y(z) \cdot 2L_{\text{mx}}^2 \rho^2 \cdot \alpha^2 \\ & + (G_P(\alpha, z) \cdot 2C_1 + C_2) \cdot G_Y(z) \cdot 8L_{\text{mx}}^2 \rho^2 \cdot G_X(z) \cdot \rho^2 \cdot \alpha^2, \end{aligned} \tag{4.56}$$

and $\mathcal{R}(\alpha, z)$ is a remainder, which is bounded under (4.52).

Therefore, as long as $\mathcal{P}(\alpha, z) < 1$, (4.55) implies

$$D^K(z) \leq \frac{\mathcal{R}(\alpha, z)}{1 - \mathcal{P}(\alpha, z)} \leq B < +\infty \tag{4.57}$$

where B is a constant independent of K. Therefore, $D(z) \leq B$ and thus $\{\|d^\nu\|^2\}$ converges R-linearly to zero at rate at least z (cf. Lemma 4.2.4). Applying the same argument to

the other inequalities in Proposition 4.2.4, one can conclude that also the sequences $\{p^\nu\}$, $\{\|x_\perp^\nu\|^2\}$ and $\{\|y_\perp^\nu\|\}$ converge R-linearly to zero.

The last step consists to showing that there exist a sufficiently small step-size $\alpha \in (0,1]$ and $z \in (0,1)$ satisfying (4.52), such that $\mathcal{P}(\alpha, z) < 1$. This is proved in the Theorem 4.2.1 below.

Theorem 4.2.1. *Consider Problem (4.1) under Assumptions 4.1.1 and 4.1.2; and the SONATA algorithm (4.12a)-(4.12d), under Assumptions 4.2.1 and 4.2.2, with $\widetilde{\mu}_{\rm mn} \geq D^\ell_{\rm mn}$. Then, there exists a sufficiently small step-size $\bar{\alpha} \in (0,1]$ [see the proof for its expression] such that for all $\alpha < \bar{\alpha}$, $\{U(x_{\rm i}^\nu)\}$ converges to $U^\star$ at an R-linear rate, ${\rm i} \in [m]$.*

Proof. The proof is organized in following two steps: **Step 1)** We first consider the "marginal" stable case by letting $z = 1$, and show that there exists $\bar{\alpha} > 0$ so that $\mathcal{P}(\alpha, 1) < 1$, for all $\alpha \in (0, \bar{\alpha})$; **Step 2)** Then, invoking the continuity of $\mathcal{P}(\alpha, z)$, we argue that, for any $\alpha \in (0, \bar{\alpha})$, one can find $\bar{z}(\alpha) < 1$ such that $\mathcal{P}\Big(\alpha, \bar{z}(\alpha)\Big) < 1$. This implies the boundedness of $D^K\Big(\bar{z}(\alpha)\Big)$, and thus $\|d^\nu\|^2 = \mathcal{O}\Big(\bar{z}(\alpha)^\nu\Big)$ (cf. Lemma 4.2.4).

• **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,

$$\epsilon^\star = \operatorname*{argmin}_{\epsilon > 0} \frac{1 + \epsilon^{-1}}{1 - \rho^2(1+\epsilon)} = \frac{1-\rho}{\rho}. \tag{4.58}$$

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 4.2.1) and the constraint (4.35), the problem boils down to minimize

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

subject to $\epsilon_{opt} \in (0, 2\widetilde{\mu}_{\rm mn} - \alpha(\widetilde{\mu}_{\rm mn} - D^\ell_{\rm mn}))$. To have a nonempty feasible set, we require $\alpha < 2\widetilde{\mu}_{\rm mn}/(\widetilde{\mu}_{\rm mn} - D^\ell_{\rm mn})$ (recall that it is assumed $\widetilde{\mu}_{\rm mn} \geq D^\ell_{\rm mn}$). Setting the derivative of

$G_P(\alpha,1)$ with respect to ϵ_{opt} to zero, yields $\epsilon^\star_{opt} = \left(1-\frac{\alpha}{2}\right)\widetilde{\mu}_{\rm mn} + \alpha D^\ell_{\rm mn}/2$, 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

$$\begin{aligned}\mathcal{P}^\star(\alpha,1) \triangleq\ & G^\star_P(\alpha)\cdot C_1\cdot 4L^2_{\rm mx}\cdot\frac{\rho^2}{(1-\rho)^2}\cdot\alpha^2 \\ &+ \left(G^\star_P(\alpha)\cdot 2C_1 + C_2\right)\cdot 2L^2_{\rm mx}\cdot\frac{\rho^2}{(1-\rho)^2}\cdot\alpha^2 \\ &+ \left(G^\star_P(\alpha)\cdot 2C_1 + C_2\right)\cdot 8L^2_{\rm mx}\cdot\frac{\rho^4}{(1-\rho)^4}\cdot\alpha^2,\end{aligned} \tag{4.59}$$

where

$$G^\star_P(\alpha) \triangleq \frac{\frac{D^2_{\rm mx}}{\mu} + \frac{1}{\mu}\cdot\left(\left(1-\frac{\alpha}{2}\right)\widetilde{\mu}_{\rm mn} + \frac{D^\ell_{\rm mn}}{2}\alpha\right)^2}{\left(\left(1-\frac{\alpha}{2}\right)\widetilde{\mu}_{\rm mn} + \frac{D^\ell_{\rm mn}}{2}\alpha\right)^2}. \tag{4.60}$$

• **Step 2:** Since $\mathcal{P}^\star(\bullet,1)$ is continuous and monotonically increasing on $(0, 2\widetilde{\mu}_{\rm mn}/\ (\widetilde{\mu}_{\rm mn} - D^\ell_{\rm mn})$, with $\mathcal{P}^\star(0,1)=0$, there exists some $\bar{\alpha} < 2\widetilde{\mu}_{\rm mn}/(\widetilde{\mu}_{\rm mn} - D^\ell_{\rm mn})$ such that $\mathcal{P}^\star(\alpha,1)<1$, for all $\alpha\in(0,\bar{\alpha})$. One can verify that, for any $\alpha\in(0, 2\widetilde{\mu}_{\rm mn}/(\widetilde{\mu}_{\rm mn} - D^\ell_{\rm mn}))$, $\mathcal{P}^\star(\alpha,z)$ 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 $\alpha \le \widetilde{\mu}_{\rm mn}/(\widetilde{\mu}_{\rm mn} - D^\ell_{\rm mn})$, we upper bound $G^\star_P(\alpha)$ by $G^\star_P(\widetilde{\mu}_{\rm mn}/(\widetilde{\mu}_{\rm mn} - D^\ell_{\rm mn}))$. Using for $G^\star_P(\alpha)$ this upper bound in (4.59) and solving the resulting $\mathcal{P}^\star(\alpha,1)<1$ for α, yield

$$\begin{aligned}\alpha < \alpha_1 \triangleq \Bigg(& G^\star_P\left(\frac{\widetilde{\mu}_{\rm mn}}{\widetilde{\mu}_{\rm mn} - D^\ell_{\rm mn}}\right)\cdot C_1\cdot 4L^2_{\rm mx}\cdot\frac{\rho^2}{(1-\rho)^2} \\ &+ \left(G^\star_P\left(\frac{\widetilde{\mu}_{\rm mn}}{\widetilde{\mu}_{\rm mn} - D^\ell_{\rm mn}}\right)\cdot 2C_1 + C_2\right)\cdot 2L^2_{\rm mx}\cdot\frac{\rho^2}{(1-\rho)^2} \\ &+ \left(G^\star_P\left(\frac{\widetilde{\mu}_{\rm mn}}{\widetilde{\mu}_{\rm mn} - D^\ell_{\rm mn}}\right)\cdot 2C_1 + C_2\right)\cdot 8L^2_{\rm mx}\cdot\frac{\rho^4}{(1-\rho)^4}\Bigg)^{-\frac{1}{2}}.\end{aligned} \tag{4.61}$$

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

The next theorem provides an explicit expression of the convergence rate in Theorem 4.2.1 in terms of the step-size α; the constants J, $A_{\frac{1}{2}}$, and α^* therein are defined in (4.104), (4.102) with $\theta = 1/2$, and (4.106), respectively.

Theorem 4.2.2. *In the setting of Theorem 4.2.1, 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, $U(x_{\rm i}^\nu) - U^\star = \mathcal{O}(z^\nu)$, for all ${\rm i} \in [m]$, where*

$$z = \begin{cases} 1 - J \cdot \alpha & \text{for } \alpha \in (0, \min\{\alpha^*, \alpha_{\rm mx}\}), \\ \left(\rho + \sqrt{\alpha A_{\frac{1}{2}}}\right)^2 & \text{for } \alpha \in [\min\{\alpha^*, \alpha_{\rm mx}\}, \alpha_{\rm mx}). \end{cases} \tag{4.62}$$

Proof. See Appendix 4.6.2. □

4.2.4 Discussion

Theorem 4.2.2 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 4.2.2 to specific network topologies and surrogate functions. We begin considering star-networks (cf. Sec. 4.2.4.1) and then move to general graph topologies with no master node (cf. Sec. 4.2.4.2). We will customize the rate achieved by SONATA employing the following two surrogate functions $\widetilde{f}_{\rm i}$, representing the two extreme choices in the spectrum of admissible surrogates:

- **Linearization:**

$$\widetilde{f}_{\rm i}(x_{\rm i}; x_{\rm i}^\nu) \triangleq \nabla f_{\rm i}(x_{\rm i}^\nu)^\top (x_{\rm i} - x_{\rm i}^\nu) + \frac{L}{2}\|x_{\rm i} - x_{\rm i}^\nu\|^2; \tag{4.63}$$

- **Local $f_{\rm i}$:**

$$\widetilde{f}_{\rm i}(x_{\rm i}; x_{\rm i}^\nu) \triangleq f_{\rm i}(x_{\rm i}) + \frac{\beta}{2}\|x_{\rm i} - x_{\rm i}^\nu\|^2. \tag{4.64}$$

4.2.4.1 Star-networks: SONATA-Star

Convergence of SONATA-Star (Algorithm 3) is established in Corollary 4.2.1 below.

Corollary 4.2.1. *Consider Problem (4.1) under Assumption 4.1.1 over a star-network; let $\{x^\nu\}$ be the sequence generated by SONATA-Star (Algorithm 3), based on the surrogate functions satisfying Assumption 4.2.1 and step-size $\alpha \in (0, \min(2\widetilde{\mu}_{\rm mn}/(\widetilde{\mu}_{\rm mn} - D^\ell_{\rm mn}), 1)]$. Then, for all $\mathrm{i} = 1, \ldots, m$,*

$$U(x^\nu) - U^\star = \mathcal{O}(z^\nu), \quad \text{with} \quad z = 1 - \alpha \cdot \frac{\left(1 - \frac{\alpha}{2}\right)\widetilde{\mu}_{\rm mn} + \frac{\alpha D^\ell_{\rm mn}}{2}}{\frac{D^2_{\rm mx}}{2\mu} + \left(1 - \frac{\alpha}{2}\right)\widetilde{\mu}_{\rm mn} + \frac{\alpha D^\ell_{\rm mn}}{2}}. \tag{4.65}$$

In particular, when the surrogates (4.63) and (4.64) are employed along with $\alpha = 1$, the rate above reduces to the following expressions:

- ***Linearization*** *(4.63)**:** $z \leq 1 - \kappa_g^{-1}$. Therefore, $U(x^\nu) - U^\star \leq \epsilon$ in at most $\mathcal{O}\Big(\kappa_g \log(1/\epsilon)\Big)$ iterations (communications);*

- ***Local*** *$f_{\rm i}$ (4.64)**:***

$$z \leq 1 - \frac{1}{1 + 4 \cdot \frac{\beta}{\mu} \cdot \min\{1, \frac{\beta}{\mu}\}}. \tag{4.66}$$

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

$$\begin{cases} \mathcal{O}\left(1 \cdot \log\left(1/\epsilon\right)\right), & \text{if } \beta \leq \mu, \\ \mathcal{O}\left(\frac{\beta}{\mu} \cdot \log\left(1/\epsilon\right)\right), & \text{if } \beta > \mu, \end{cases} \tag{4.67}$$

  iterations (communications).

Proof. See Appendix 4.6.3. □

The following comments are in order. When linearization is employed, SONATA-Star matches the iteration complexity of the centralized proximal-gradient algorithm. When the $f_{\rm i}$'s are sufficiently similar, (4.66)-(4.67) proves that faster rates can be achieved if surrogates

(4.64) are chosen over first-order approximations: when $\beta \ll L$, (4.67) is significantly faster than $\mathcal{O}\big(\kappa_g \log(1/\epsilon)\big)$. As case study, consider Example 2 (cf. Sec. 4.1.1.2): plugging (4.11) into Corollary 4.2.1 shows that using the surrogates (4.64) yields $\widetilde{\mathcal{O}}\big(L\sqrt{d\,m}\cdot\log(1/\epsilon)\big)$ iterations (communications); this contrasts with $\widetilde{\mathcal{O}}\big(L\sqrt{d\,m\,n}\cdot\log(1/\epsilon)\big)$, achieved by first-order methods (and SONATA-Star using linearization), which instead increases with the sample size n.

Comparison with DANE & CEASE Since SONATA-Star contains as special cases the DANE [140] and CEASE [142] algorithms, we contrast here Corollary 4.2.1 with their convergence rates. We recall that DANE is applicable to (4.1) when $G = 0$: For quadratic losses, it achieves an ϵ-optimal objective value in $\mathcal{O}\big((\beta/\mu)^2\cdot\log(1/\epsilon)\big)$ iterations/communications (here $\beta/\mu \geq 1$). This rate is worse than (4.67). For nonquadratic losses, [140] did not show any rate improvement of DANE over plain gradient algorithms, i.e., $\mathcal{O}\big(\kappa_g\cdot\log(1/\epsilon)\big)$ while SONATA-star still retains $\mathcal{O}\big(\beta/\mu\cdot\log(1/\epsilon)\big)$. The CEASE algorithm is proved to achieve an ϵ-solution on the iterates in $\mathcal{O}\big((\beta/\mu)^2\cdot\log(1/\epsilon)\big)$ iterations/communications (with $\beta/\mu \geq 1$); SONATA reaches the same error on the iterates in $\mathcal{O}\big(\beta/\mu\cdot\log(\kappa_g/\epsilon)\big)$ iterations/communications, which matches the order of the mirror-decent algorithm.

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

4.2.4.2 The general case

The convergence rate of SONATA over general graphs is summarized in Corollary 4.2.2 for the linearization surrogates (4.63) while Corollaries 4.2.3 and 4.2.4 consider the surrogates (4.64) based on local f_i, with Corollary 4.2.3 addressing the case $\beta \leq \mu$ and Corollary 4.2.4 the case $\beta > \mu$. The step-size α is tuned to obtain favorable rate expressions.

Corollary 4.2.2 (Linearization surrogates)**.** *In the setting of Theorem 4.2.2, let $\{x^\nu\}$ be the sequence generated by SONATA, using the surrogates (4.63) 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* $U(x_{\rm i}^\nu) - U^\star \leq \epsilon$, $\mathrm{i} \in [m]$, *is*

$$\textbf{Case I:} \qquad \mathcal{O}\left(\kappa_g \log(1/\epsilon)\right), \qquad \text{if} \quad \frac{\rho}{(1-\rho)^2} \leq \frac{1}{110\,\kappa_g \left(1+\frac{\beta}{L}\right)^2}, \tag{4.68}$$

$$\textbf{Case II:} \qquad \mathcal{O}\left(\frac{\left(\kappa_g + \beta/\mu\right)^2 \rho}{(1-\rho)^2} \log(1/\epsilon)\right), \qquad \text{otherwise.} \tag{4.69}$$

Proof. See Appendix 4.6.4. □

Corollary 4.2.3 (local $f_{\rm i}$, $\beta \leq \mu$)**.** *Instate assumptions of Theorem 4.2.2 and suppose* $\beta \leq \mu$. *Consider SONATA using the surrogates (4.64) and step-size* $\alpha = c \cdot \alpha_{\rm mx}$, $c \in (0,1)$, *with* $\alpha_{\rm mx} = \min\{1, (1-\rho)^2/(M\rho)\}$ *and* $M = 193\left(1+\frac{\beta}{\mu}\right)^2\left(\kappa_g + \frac{\beta}{\mu}\right)^2$. *The number of iterations (communications) needed for* $U(x_{\rm i}^\nu) - U^\star \leq \epsilon$, $\mathrm{i} \in [m]$, *is*

$$\textbf{Case I:} \qquad \mathcal{O}\left(1 \cdot \log(1/\epsilon)\right), \qquad \text{if} \quad \frac{\rho}{(1-\rho)^2} \leq \frac{1}{193\left(1+\frac{\beta}{\mu}\right)^2\left(\kappa_g + \frac{\beta}{\mu}\right)^2}, \tag{4.70}$$

$$\textbf{Case II:} \qquad \mathcal{O}\left(\frac{\kappa_g^2\,\rho}{(1-\rho)^2} \log(1/\epsilon)\right), \qquad \text{otherwise.} \tag{4.71}$$

Corollary 4.2.4 (local $f_{\rm i}$, $\beta > \mu$)**.** *Instate assumptions of Theorem 4.2.2 and suppose* $\beta > \mu$. *Consider SONATA using the surrogates (4.64) and step-size* $\alpha = c \cdot \alpha_{\rm mx}$, $c \in (0,1)$, *with* $\alpha_{\rm mx} = \min\{1, (1-\rho)^2/(M\rho)\}$ *and* $M = 253\left(1+\frac{L}{\beta}\right)\left(\kappa_g + \frac{\beta}{\mu}\right)$. *The number of iterations (communications) needed for* $U(x_{\rm i}^\nu) - U^\star \leq \epsilon$, $\mathrm{i} \in [m]$, *is*

$$\textbf{Case I:} \quad \mathcal{O}\left(\frac{\beta}{\mu} \cdot \log(1/\epsilon)\right) \qquad \text{if} \quad \frac{\rho}{(1-\rho)^2} \leq \frac{1}{253\left(1+\frac{L}{\beta}\right)\left(\kappa_g + \frac{\beta}{\mu}\right)}, \tag{4.72}$$

$$\textbf{Case II:} \quad \mathcal{O}\left(\frac{\left(\kappa_g + (\beta/\mu)\right)^2 \rho}{(1-\rho)^2} \log(1/\epsilon)\right), \qquad \text{otherwise}\,. \tag{4.73}$$

The proof of Corollaries 4.2.3 and 4.2.4 can be found in Appendix 4.6.5.

Several comments are in order.

• **Order of the rate of centralized (nonaccelerated) methods (Case I):** For a fixed optimization problem, if the network is sufficiently connected (ρ "small"), its impact on

the rate becomes negligible (the bottleneck is the optimization), and SONATA matches the *network-independent* rate order achieved on star-topologies (cf. Corollary 4.2.1) by the proximal gradient algorithm when linearization is employed [cf. (4.68)] and by the mirror-descent scheme when the local f_i's are used in the surrogates [cf. (4.70) and (4.72)].

• **Network-dependent rates (Case II):** As expected, the convergence rate deteriorates as ρ increases, i.e., the network connectivity gets worse. This translates in a less favorable dependence of the complexity on κ_g and β/μ (by a square factor) and network scalability of the order of $\rho/(1-\rho)^2$. When $\beta\sqrt{\rho} = \mathcal{O}(L)$ (e.g., the network is decently connected or $\beta = \mathcal{O}(L)$), the complexity becomes $\mathcal{O}\left(\kappa_g^2(1-\rho)^{-2}\log(1/\epsilon)\right)$, which compares favorably with that of existing distributed schemes, determined instead by the more pessimistic local quantities (4.4). The scalability of the rate with the network connectivity, $(1-\rho)^{-2}$, can be improved leveraging multiple rounds of communications or accelerated consensus protocols, as discussed below.

• **Linearization (4.63) vs. local f_i (4.64) surrogates:** As already observed in the setting of star-networks, the use of the local losses as surrogates employs a form of preconditioning in the local agents subproblems. When the f_i's are sufficiently similar to each other, so that $1 + \beta/\mu < \kappa_g$, exploiting local Hessian information via (4.64) provably reduces the iteration/communication complexity over linear models (4.63)–contrast (4.68) with (4.70) and (4.72). 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 (4.64) are no longer informative of the average-loss F and using linearization might yield better rates. Although these design recommendations are based on sufficient conditions, numerical results seem to confirm the above conclusions–see Sec. 4.4.

• **Multiple communications rounds and acceleration:** The discussion above shows that rates of the order of those of centralized methods can be achieved if the network is sufficiently connected (Case I). When this is not the case, one can still achieve the same iteration complexity at the cost of multiple, finite, rounds of communications per iteration. Specifically, let ρ_0 be the connectivity of the given network and suppose we run K steps of communications per iteration (computation) in (4.44a)-(4.44b); this yields an effective

network with improved connectivity $\rho = \rho_0^K$. One can then choose K so that the ratio $\rho_0^K/(1-\rho_0^K)^2$ satisfies the condition triggering Case I in the Corollaries 4.2.2–4.2.4, as briefly summarized next.

1) Linearization: Invoking Corollary 4.2.2, 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 (4.63) reaches an ϵ-solution in $\mathcal{O}\left(\kappa_g \log(1/\epsilon)\right)$ iterations and $\mathcal{O}\left(\kappa_g \cdot (1-\rho_0)^{-1} \log(\kappa_g(1+\beta/L)^2) \log(1/\epsilon)\right)$ communications. The dependence on the network connectivity ρ_0 can be further improved leveraging Chebyshev polynomials (see, e.g., [163], [164]): the final communication complexity of SONATA reads

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

2) Local f_i surrogates: Considering the case $\beta \geq \mu$ (Corollary 4.2.4), we can show that SONATA using the surrogates (4.64) 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

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

communications. If Chebyshev polynomials are used to accelerate the communications, the communication complexity further improves to

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

4.3 The SONATA algorithm over directed time-varying graphs

In this section we extend SONATA and its convergence analysis to solve Problem (4.1) over *directed, time-varying graphs* (Assumption 4.1.3). Note that (4.12a)-(4.12d) is not readily applicable to this setting, as constructing a doubly stochastic weight matrix compliant with a directed graph is generally infeasible or computationally costly–see e.g. [190]. Conditions on the weight matrices can be relaxed if the consensus/tracking schemes (4.12c)-(4.12d) are properly changed to deal with the lack of doubly stochasticity.

Here, we consider the perturbed push-sum protocols as proposed in [144] (but in the Adapt-Then-Combine (ATC) form). The resulting distributed algorithm, still termed SONATA, is formally described in Algorithm 4.

Algorithm 4: SONATA over time-varying directed graphs

Data: $x_{\mathrm{i}}^0 \in \mathcal{K}$, $y_{\mathrm{i}}^0 = \nabla f_{\mathrm{i}}(x_{\mathrm{i}}^0)$, and $\phi_{\mathrm{i}}^0 = 1$, $\mathrm{i} \in [m]$.

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

`[S.1] [Distributed Local Optimization]` Each agent i solves

$$\widehat{x}_{\mathrm{i}}^{\nu} \triangleq \operatorname*{argmin}_{x_{\mathrm{i}} \in \mathcal{K}} \ \widetilde{f}_{\mathrm{i}}(x_{\mathrm{i}}; x_{\mathrm{i}}^{\nu}) + \left(y_{\mathrm{i}}^{\nu} - \nabla f_{\mathrm{i}}(x_{\mathrm{i}}^{\nu})\right)^{\top}(x_{\mathrm{i}} - x_{\mathrm{i}}^{\nu}) + G(x_{\mathrm{i}}), \tag{4.74a}$$

and updates

$$x_{\mathrm{i}}^{\nu+\frac{1}{2}} = x_{\mathrm{i}}^{\nu} + \alpha \cdot d_{\mathrm{i}}^{\nu}, \quad \text{with} \quad d_{\mathrm{i}}^{\nu} \triangleq \widehat{x}_{\mathrm{i}}^{\nu} - x_{\mathrm{i}}^{\nu}; \tag{4.74b}$$

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

(a) `Consensus`

$$\phi_{\mathrm{i}}^{\nu+1} = \sum_{\mathrm{j}=1}^{m} c_{\mathrm{ij}}^{\nu} \phi_{\mathrm{j}}^{\nu}, \quad x_{\mathrm{i}}^{\nu+1} = \frac{1}{\phi_{\mathrm{i}}^{\nu+1}} \sum_{\mathrm{j}=1}^{m} c_{\mathrm{ij}}^{\nu} \phi_{\mathrm{j}}^{\nu} x_{\mathrm{j}}^{\nu+\frac{1}{2}}, \tag{4.74c}$$

(b) `Gradient tracking`

$$y_{\mathrm{i}}^{\nu+1} = \frac{1}{\phi_{\mathrm{i}}^{\nu+1}} \sum_{\mathrm{j}=1}^{m} c_{\mathrm{ij}}^{\nu} \left(\phi_{\mathrm{j}}^{\nu}\, y_{\mathrm{j}}^{\nu} + \nabla f_{\mathrm{j}}(x_{\mathrm{j}}^{\nu+1}) - \nabla f_{\mathrm{j}}(x_{\mathrm{j}}^{\nu})\right), \tag{4.74d}$$

end

In the perturbed push-sum protocols (4.74c)-(4.74d), $C^{\nu} \triangleq (c_{\mathrm{ij}}^{\nu})_{\mathrm{i,j}=1}^{m}$ satisfies the assumption below.

Assumption 4.3.1. *For each $\nu \geq 0$, the weight matrix $C^{\nu} \triangleq (c_{\mathrm{ij}}^{\nu})_{\mathrm{i,j}=1}^{m}$ has a sparsity pattern compliant with $\mathcal{G}^{\nu}$, i.e., there exists a constant c_{ℓ} such that, for all $\nu = 0, 1, \ldots,$*

4.3.1.1 $c_{\mathrm{ii}}^{\nu} \geq c_{\ell} > 0$, for all $\mathrm{i} \in [m]$;

4.3.1.2 $c_{\mathrm{ij}}^{\nu} \geq c_{\ell} > 0$, if $(\mathrm{j}, \mathrm{i}) \in \mathcal{E}^{\nu}$; and $c_{\mathrm{ij}}^{\nu} = 0$ otherwise.

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

We conclude this section stating the counterparts of the definitions introduced in Sec. 4.1, adjusted here to the case of directed time-varying graphs. Using the column stochasticity of C^ν and (4.74d), one can see that opposed to (4.19), the average gradient is now preserved on the weighted average of the y_i's:

$$\frac{1}{m}\sum_{\mathrm{i}=1}^{m}\phi_\mathrm{i}^{\nu+1}y_\mathrm{i}^{\nu+1}=\frac{1}{m}\sum_{\mathrm{i}=1}^{m}\phi_\mathrm{i}^{\nu}y_\mathrm{i}^{\nu}+\overline{\nabla F_c}^{\nu+1}-\overline{\nabla F_c}^{\nu}, \tag{4.75}$$

where $\overline{\nabla F_c}^{\nu}$ is defined in (4.18). This suggests to decompose y^ν into its weighted average and the consensus error, defined respectively as

$$\bar{y}_\phi^\nu \triangleq \frac{1}{m}\sum_{\mathrm{i}=1}^{m}\phi_\mathrm{i}^{\nu}y_\mathrm{i}^{\nu} \quad \text{and} \quad y_{\phi,\perp}^{\nu} \triangleq y^\nu - 1_m \otimes \bar{y}_\phi^\nu. \tag{4.76}$$

Accordingly, we define the weighted average of x^ν and the consensus error as

$$\bar{x}_\phi^\nu \triangleq \frac{1}{m}\sum_{\mathrm{i}=1}^{m}\phi_\mathrm{i}^{\nu}x_\mathrm{i}^{\nu} \quad \text{and} \quad x_{\phi,\perp}^{\nu} \triangleq x^\nu - 1_m \otimes \bar{x}_\phi^\nu. \tag{4.77}$$

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

$$p_\phi^\nu \triangleq \sum_{\mathrm{i}=1}^{m}\phi_\mathrm{i}^{\nu}p_\mathrm{i}^{\nu}, \quad \text{with} \quad p_\mathrm{i}^{\nu} \triangleq \Big(U(x_\mathrm{i}^{\nu}) - U^\star\Big). \tag{4.78}$$

Finally, apart from the problem parameters L_i, L_mx, L, μ [cf. (4.5)] and algorithm parameters $\widetilde{\mu}_\mathrm{mn}$, $\widetilde{L}_\mathrm{mx}$, D_mn^ℓ, D_mx [cf. (4.25)], we introduce the following network parameters, borrowed from [144, Prop. 1]:

$$\phi_{lb} \triangleq c_\ell^{2(m-1)B}, \quad \phi_{ub} \triangleq m - c_\ell^{2(m-1)B}, \tag{4.79}$$

with c_ℓ and B given in Assumptions 4.3.1 and 4.1.3, respectively; and

$$c_0 \triangleq 2m\cdot\frac{1+\tilde{c}_\ell^{-(m-1)B}}{1-\tilde{c}_\ell^{-(m-1)B}}, \quad \rho_B \triangleq (1-\tilde{c}_\ell^{(m-1)B})^{\frac{1}{(m-1)B}}, \quad \tilde{c}_\ell \triangleq c_\ell^{2(m-1)B+1}/m. \tag{4.80}$$

Furthermore, we will use the following lower and upper bounds of ϕ_{i}^{ν} [144, Prop. 1]

$$\phi_{lb} \leq \phi_{\mathrm{i}}^{\nu} \leq \phi_{ub}, \quad \text{for all } \mathrm{i} \in [m], \quad \nu = 0, 1, \dots.$$

4.3.1 Linear convergence rate

The proof of linear convergence of SONATA (Algorithm 4) follows the same path of the one developed in Sec. 4.2.3 for the case of undirected graphs. Hence, we omit similar derivations and highlight only the key differences. We will tacitly assume that Assumptions 4.1.1, 4.1.3, 4.2.1, and 4.3.1 are satisfied.

Step 1: p_{ϕ}^{ν} converges linearly up $\mathcal{O}(\|x_{\phi,\perp}^{\nu}\|^2 + \|y_{\phi,\perp}^{\nu}\|^2)$ This is counterpart of Proposition 4.2.1 (cf. Sec. 4.2.3), and stated as follows.

Proposition 4.3.1. *The optimality gap sequence $\{p_{\phi}^{\nu}\}$ satisfies:*

$$p_{\phi}^{\nu+1} \leq \sigma(\alpha) \cdot p_{\phi}^{\nu} + \eta(\alpha) \cdot \phi_{ub} \cdot \left(8L_{\mathrm{mx}}^2 \|x_{\phi,\perp}^{\nu}\|^2 + 2\|y_{\phi,\perp}^{\nu}\|^2\right), \tag{4.81}$$

where the constants L_{mx} and $\widetilde{\mu}_{\mathrm{mn}}$ are defined in (4.5) and (4.25), respectively; and $\sigma(\alpha) \in (0,1)$ and $\eta(\alpha) > 0$ are defined in (4.42).

Proof. The proof follows closely that of Proposition 4.2.1 and thus is omitted. For completeness, we report it in the supporting materials. Here, we only notice that, instead of (4.28), we built on: $\sum_{\mathrm{i}=1}^{m} \phi_{\mathrm{i}}^{\nu+1} U(x_{\mathrm{i}}^{\nu+1}) \leq \sum_{\mathrm{i}=1}^{m} \phi_{\mathrm{i}}^{\nu} U\left(x_{\mathrm{i}}^{\nu+\frac{1}{2}}\right)$, where we used $\sum_{\mathrm{j}=1}^{m} c_{\mathrm{ij}}^{\nu} \phi_{\mathrm{j}}^{\nu} / \phi_{\mathrm{i}}^{\nu+1} = 1$, for all $\mathrm{i} \in [m]$. □

Step 2: Decay of $\|x_{\phi,\perp}^{\nu}\|$ and $\|y_{\phi,\perp}^{\nu}\|$

Lemma 4.3.1. *The following bounds hold for $\|x_{\phi,\perp}^{\nu}\|$ and $\|y_{\phi,\perp}^{\nu}\|$:*

$$\left\|x_{\phi,\perp}^{\nu}\right\|^2 \leq 2c_0^2 \rho_B^{2\nu} \left\|x_{\phi,\perp}^{0}\right\|^2 + \frac{2c_0^2 \rho_B^2}{1-\rho_B} \sum_{t=0}^{\nu-1} \rho_B^{\nu-1-t} \alpha^2 \left\|d^t\right\|^2 \tag{4.82a}$$

$$\left\|y_{\phi,\perp}^{\nu}\right\|^2 \leq 2c_0^2 \rho_B^{2\nu} \left\|y_{\phi,\perp}^{0}\right\|^2 + \frac{2c_0^2 \rho_B^2 m L_{\mathrm{mx}}^2 \phi_{lb}^{-2}}{1-\rho_B} \sum_{t=0}^{\nu-1} \rho_B^{\nu-1-t} \left(8\left\|x_{\phi,\perp}^{t}\right\|^2 + 2\alpha^2 \left\|d^t\right\|^2\right). \tag{4.82b}$$

where B and ρ_B are defined in (4.79), and ϵ_x and ϵ_y are arbitrary positive constants (to be determined).

Proof. Using the result in [191, Lemma 5] and [144, Lemma 3, 11], we obtain

$$\left\|x^{\nu}_{\phi,\perp}\right\| \leq c_0 \left(\rho_B^{\nu}\left\|x^{0}_{\phi,\perp}\right\| + \sum_{t=0}^{\nu-1} \rho_B^{(\nu-1)-t}(\rho_B \alpha \left\|d^t\right\|)\right) \tag{4.83}$$

$$\left\|y^{\nu}_{\phi,\perp}\right\| \leq c_0 \left(\rho_B^{\nu}\left\|y^{0}_{\phi,\perp}\right\| + \sqrt{m} L_{\mathrm{mx}} \phi_{lb}^{-1} \sum_{t=0}^{\nu-1} \rho_B^{(\nu-1)-t} \cdot \rho_B \left(2\left\|x^{t}_{\phi,\perp}\right\| + \alpha\left\|d^t\right\|\right)\right). \tag{4.84}$$

The rest of the proof follows similar steps as [73, Lemma 2], hence it is omitted. □

Step 3: $\|d^{\nu}\| = \mathcal{O}(\sqrt{p^{\nu}_{\phi}} + \|y^{\nu}_{\phi,\perp}\|)$

Proposition 4.3.2. *The following upper bound holds for $\|d^{\nu}\|$:*

$$\|d^{\nu}\|^2 \leq \frac{6}{\mu\phi_{lb}} \left(\left(\frac{D_{\mathrm{mx}}}{\widetilde{\mu}_{\mathrm{mn}}} + 1\right)^2 + \frac{4L^2_{\mathrm{mx}}}{\widetilde{\mu}^2_{\mathrm{mn}}}\right) p^{\nu}_{\phi} + \frac{3}{\widetilde{\mu}^2_{\mathrm{mn}}} \|y^{\nu}_{\phi,\perp}\|^2, \tag{4.85}$$

where L_{mx}, $\widetilde{L}_{\mathrm{mx}}$, $\widetilde{\mu}_{\mathrm{mn}}$, and D_{mx} are defined in (4.5) and (4.25), respectively.

Proof. The proof follows similar path of that of Proposition 4.2.3 and thus is omitted. □

4.3.2 Establishing linear rate

We can now prove linear rate following the path introduced in Sec. 4.2.3; for sake of simplicity, we will use the same notation as in Sec. 4.2.3. We begin applying the transformation (4.50) to the sequences $\{p^{\nu}_{\phi}\}_{\nu\in\mathbb{N}_+}$, $\{\|x^{\nu}_{\phi,\perp}\|^2\}$, $\{\|y^{\nu}_{\phi,\perp}\|^2\}$, and $\{\|d^{\nu}\|^2\}$, satisfying the inequalities (4.81), (4.82a), (4.82b), and (4.85), respectively.

Proposition 4.3.3. *Let $P^K_{\phi}(z)$, $D^K(z)$, $X^K_{\phi,\perp}(z)$, and $Y^K_{\phi,\perp}(z)$ denote the transformation (4.50) of the sequences $\{p^{\nu}_{\phi}\}$, $\{\|d^{\nu}\|^2\}$, $\{\|x^{\nu}_{\phi,\perp}\|^2\}$ and $\{\|y^{\nu}_{\phi,\perp}\|^2\}$. Given the constants $\sigma(\alpha)$ and $\eta(\alpha)$, defined in Proposition 4.3.1, and the free parameters $\epsilon_x, \epsilon_y > 0$, the following holds:*

$$P^K_{\phi}(z) \leq G_P(\alpha, z) \cdot \left(8\phi_{ub} L^2_{\mathrm{mx}} X^K_{\phi,\perp}(z) + 2\phi_{ub} Y^K_{\phi,\perp}(z)\right) + \omega_p \tag{4.86}$$

$$X^K_{\phi,\perp}(z) \leq G_X(z) \cdot \rho_B^2 \alpha^2 D^K(z) + \omega_x \tag{4.87}$$

$$Y^K_{\phi,\perp}(z) \leq G_Y(z) \cdot 2m\phi_{lb}^{-2} L^2_{\text{mx}} \rho_B^2 \left(4X^K_{\phi,\perp}(z) + \alpha^2 D^K(z)\right) + \omega_y \tag{4.88}$$

$$D^K(z) \leq C_1 \cdot P^K_\phi(z) + C_2 \cdot Y^K_{\phi,\perp}(z), \tag{4.89}$$

for all

$$z \in (\max\{\sigma(\alpha), \rho_B\}, 1), \tag{4.90}$$

where

$$G_P(\alpha, z) \triangleq \frac{\eta(\alpha)}{z - \sigma(\alpha)}, \qquad \omega_p \triangleq \frac{z}{z - \sigma(\alpha)} \cdot p^0_\phi \tag{4.91}$$

$$G_X(z) \triangleq \frac{2c_0^2}{(1-\rho_B)(z-\rho_B)}, \qquad \omega_x \triangleq 2c_0^2 \left\| x^0_{\phi,\perp} \right\|^2 \tag{4.92}$$

$$G_Y(z) \triangleq \frac{2c_0^2}{(1-\rho_B)(z-\rho_B)}, \qquad \omega_y \triangleq 2c_0^2 \left\| y^0_{\phi,\perp} \right\|^2 \tag{4.93}$$

$$C_1 \triangleq \frac{6}{\mu\phi_{lb}} \left(\left(\frac{D_{\text{mx}}}{\widetilde{\mu}_{\text{mn}}} + 1 \right)^2 + \frac{4L^2_{\text{mx}}}{\widetilde{\mu}^2_{\text{mn}}} \right), \qquad C_2 \triangleq \frac{4}{\widetilde{\mu}^2_{\text{mn}}}. \tag{4.94}$$

Proof. The proof of the first two inequalities (4.86) and (4.89) follows the same steps of those used to prove Proposition 4.2.4. Applying [192, Lemma 21] to (4.82a) and (4.82b) respectively gives (4.87) and (4.88).

□

Chaining the inequalities in Proposition 4.3.3 as done in for (4.51) (cf. Fig. 4.2), we can bound $D^K(z)$ as

$$D^K(z) \leq \mathcal{P}(\alpha, z) \cdot D^K(z) + \mathcal{R}(\alpha, z), \tag{4.95}$$

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

$$\begin{aligned} \mathcal{P}(\alpha, z) \triangleq\ & G_P(\alpha, z) \cdot G_X(z) \cdot C_1 \cdot 8\phi_{ub} L^2_{\text{mx}} \cdot \rho_B^2 \cdot \alpha^2 \\ & + (G_P(\alpha, z) \cdot 2\phi_{ub} \cdot C_1 + C_2) \cdot G_Y(z) \cdot 2m\phi_{lb}^{-2} L^2_{\text{mx}} \cdot \rho_B^2 \cdot \alpha^2 \\ & + (G_P(\alpha, z) \cdot 2\phi_{ub} \cdot C_1 + C_2) \cdot G_Y(z) \cdot 8m\phi_{lb}^{-2} L^2_{\text{mx}} \cdot G_X(z) \cdot \rho_B^4 \cdot \alpha^2 \end{aligned} \tag{4.96}$$

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

Comparing (4.96) to (4.56) we can see that they share the same form and only differ in coefficients. Therefore, with the same argument as in the proof of Theorem 4.2.1 we can easily arrive at the following conclusion.

Theorem 4.3.1. *Consider Problem* (4.1) *under Assumptions 4.1.1, and 4.1.3; and SONATA (Algorithm 4) under Assumptions 4.2.1 and 4.3.1, with* $\widetilde{\mu}_{\mathrm{mn}} \geq D^{\ell}_{\mathrm{mn}}$. *Then, there exists a sufficiently small step-size* $\bar{\alpha} \in (0, 1]$ *such that, for all* $\alpha < \bar{\alpha}$, $\{U(x_{\mathrm{i}}^{\nu})\}$ *converges to* $U^{\star}$ *at an R-linear rate,* $\mathrm{i} \in [m]$.

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

For sake of completeness, we provide an explicit expression of the linear rates in terms of the step-size α in the supporting material–see Theorem 4.6.1. Table 4.4 summarizes the expression of the rates achieved by SONATA using the surrogate functions (4.63) and (4.64)–a formal statement of these results along with the proofs can be found in the supporting material-see Corollaries 4.6.1, 4.6.2 and 4.6.3.

The rate estimates in Table 4.4 are almost identical to those obtained in Sec. 4.2.4.2, with the difference that the network dependence now is expressed throughout ρ_B rather than ρ. Therefore, similar comments–as those stated in Sec. 4.2.4.2–apply to the rates in Table 4.4. For example, if the network is sufficiently connected (ρ_B "small"), its impact on the rate becomes negligible and SONATA matches the *network-independent* rate achieved on star-topology (cf. Corollary 4.2.1) or centralized settings. Specifically, when linearization surrogate (4.63) is used, this rate coincides with the rates of centralized proximal gradient algorithm.

4.4 Numerical Results

In this section, we corroborate numerically the complexity results proved in Corollaries 4.2.2–4.2.4. As a test problem, we consider the distributed ridge regression:

$$\min_{x \in \mathbb{R}^d} \frac{1}{m} \left\{ \frac{1}{2n} \|A_{\mathrm{i}} x - b_{\mathrm{i}}\|^2 + \lambda \|x\|^2 \right\}, \tag{4.97}$$

Table 4.4. Summary of convergence rates of SONATA over time-varying directed graphs: number of communication rounds to reach ϵ-accuracy.

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

where the loss function of agent i is $f_{\mathrm{i}}(x) = \frac{1}{2n}\|A_{\mathrm{i}}x - b_{\mathrm{i}}\|^2 + \lambda\|x\|^2$ [agent i owns data $(A_{\mathrm{i}}, b_{\mathrm{i}})$]. Problem parameters are generated as follows. Each row of the measurement matrix A_{i} is independently and identically drawn from distribution $\mathcal{N}(0,\Sigma)$; and b_{i} is generated according to the linear model $b_{\mathrm{i}} = A_{\mathrm{i}}x^* + n_{\mathrm{i}}$, where x^* is the ground truth, generated according to $\mathcal{N}(5\cdot 1, I)$, and $n_{\mathrm{i}} \sim \mathcal{N}(0, 0.1\cdot I)$ is the measurement noise. The covariance matrix Σ is constructed according to the eigenvalue decomposition $\Sigma = \sum_{\mathrm{j}=1}^{d}\lambda_{\mathrm{j}}u_{\mathrm{j}}u_{\mathrm{j}}^\top$, where the eigenvalues $\{\lambda_{\mathrm{j}}\}_{\mathrm{j}=1}^{d}$ are uniformly distributed in $[\mu_0, L_0]$. The eigenvectors, forming $U = [u_1,\ldots,u_d]$, are obtained via the QR decomposition of a random $d\times d$ matrix with standard Gaussian i.i.d. elements. The network is generated using an Erdős-Rényi model $G(m,p)$,

with $m = 30$ nodes and each edge independently included in the graph with probability $p = 0.5$.

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

(S.I) **Changing κ_g with fixed β**: We generate a sequence of instances of (4.97) with fixed β and increasing κ_g. To do so, we use the same data set $\{A_{\mathrm{i}}, b_{\mathrm{i}}\}$ across the different instances and change the regularization parameter λ, so that the condition number κ_g ranges in $[K_\ell, K_u]$.

(S.II) **Changing β with (almost) fixed κ_g**: We generate instances of (4.97) with decreasing β and (almost) fixed κ_g. To do so, we set $\lambda = 0$ and increased the local sample size n from N_ℓ to N_u; we set N_ℓ sufficiently large so that the empirical condition number κ_g is close to L_0/μ_0 for all instances.

We run SONATA using surrogates (4.63) (linearization) and (4.64) (local f_{i})–we term it as SONATA-L and SONATA-F, respectively. The simulations parameters of the different experiments are summarized in Table 4.5; and the algorithmic parameters are set according to Corollaries 4.2.2–4.2.4.[1] We measure the algorithm's complexity using $T_\epsilon = \inf\left\{\nu \geq 0 \,|\, \frac{1}{m}\sum_{\mathrm{i}=1}^{m}(F(x_{\mathrm{i}}^{\nu}) - F^\star) \leq 10^{-7}\right\}$.

In Table 4.6, we report the corresponding iteration complexity of SONATA for each simulation setup (s.1)-(s.6) in Table 4.5. Each figure is generated under one particular realization of the problem setting. Further, in order to compare the complexity of SONATA across different settings, all the simulations share the same network parameters, as well as the same data set whenever the problem parameters are the same. The results of our experiments are reported in Table 4.6; the curve are generated using only one random realization for visualization clarity. However, the behavior of the curves (e.g., scalability with respect to the parameters) is representative and consistent across all the random experiments we conducted.

The following comments are in order.

[1]↑ The expressions are not tight in terms of the absolute constants. To show convergence rate in both Cases I and II in Corollary 4.2.2-4.2.4, we enlarged the second term in the expression of α_{mx} by a constant factor.

Table 4.5. Simulation setup and parameter setting.

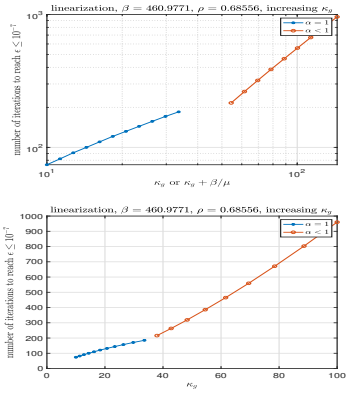
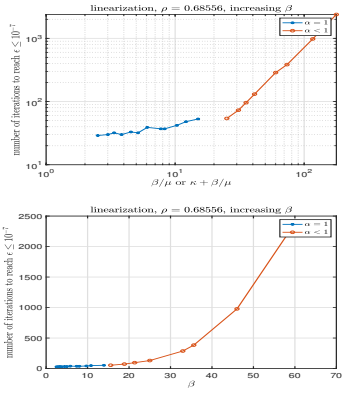
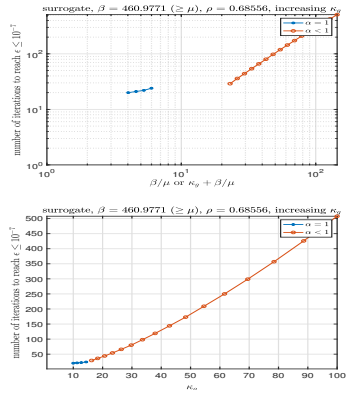
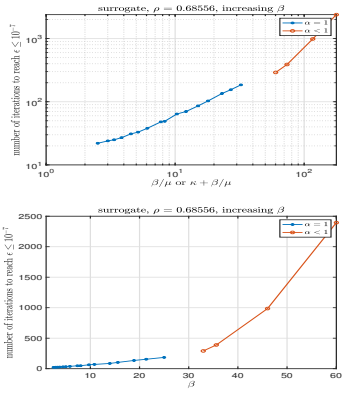
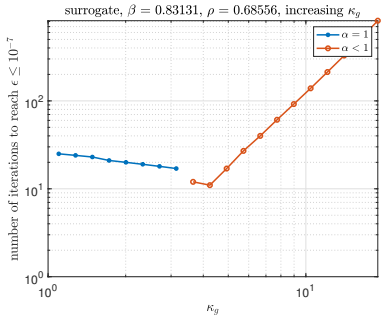
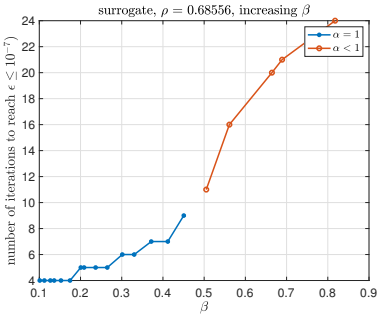
	Setting (S.I)	Setting (S.II)
Linearization	(s.1) $n = 10^3$ $\mu_0 = 1,\ L_0 = 10^3$ $K_\ell = 10,\ K_u = 100$	(s.4) $\lambda = 0$ $\mu_0 = 1,\ L_0 = 5,\ \kappa_g \approx 5$ $N_\ell = 10,\ N_u = 10^3$
Local f_{i} ($\beta \geq \mu$)	(s.2) same as above	(s.5) same as above
Local f_{i} ($\beta < \mu$)	(s.3)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

• **Scalability with respect to κ_g.** Consider setting (S.I) wherein β is fixed and λ is changing. Figures for (s.1)-(s.3) show that when $\alpha = 1$ (blue curve), the iteration complexity of SONATA-L scales linearly with respect to κ_g [as predicted by Corollary 4.2.2], while that of SONATA-F is invariant whenever $\beta < \mu$ [as stated in Corollary 4.2.3]. When $\beta \geq \mu$, the iteration complexity of SONATA-F grows as λ increases since β/μ decreases [cf. Corollary 4.2.4]. However, the increasing rate is much slower than SONATA-L, due to the fact that $(\beta/\mu)/\kappa_g = \beta/L \ll 1$ for large λ. When $\alpha < 1$, the iteration complexity scales quadratically with respect to κ_g, in all settings, as predicted by our theory.

• **Scalability with respect to β.** Consider now setting (S.II), where we decrease the local sample size n to increase β. In contrast to setting (S.I), Figures for (s.4) and (s.5) show that, with $\alpha = 1$, the iteration complexity of SONATA-F scales linearly with β/μ when $\beta > \mu$, while that of SONATA-L is invariant–this is consistent with Corollaries 4.2.2 and 4.2.4. When $\alpha < 1$, the iteration complexity scales quadratically with respect to β/μ. Finally, the plot associated with (s.6) simply reveals that when $\beta < \mu$, iteration complexity of SONATA-F remains bounded, as stated in Corollary 4.2.3.

• **Linearization versus Local f_{i}.** We compare the performance of SONATA-L and SONATA-F in the setting (S.II), with parameters $\lambda = 0$, $\mu_0 = 1$, $L_0 = 100$, $N_\ell = 10$, $N_u = 10^5$. We consider a relatively connected network with edge activation probability

Table 4.6. Iteration complexity of SONATA under the simulation settings in Table 4.5. Left (S.I): scalability of iteration complexity with respect to the condition number κ_g; Right (S.II): scalability of the iteration complexity with respect to the similarity parameter β.

	Setting (S.I)	Setting (S.II)
Linearization	(s.1) 	(s.4) 
Local f_{i} ($\beta \geq \mu$)	(s.2) 	(s.5) 
Local f_{i} ($\beta < \mu$)	(s.3) 	(s.6) 

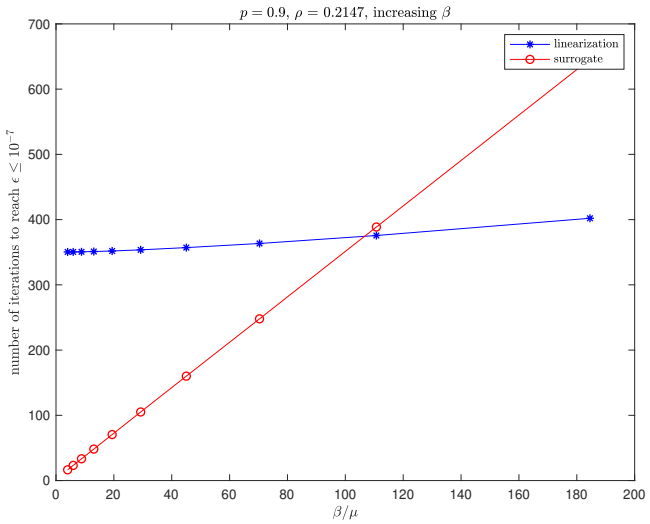


Figure 4.3. Complexity of SONATA-L versus SONATA-F.

$p = 0.9$ so that the step-size can be set to $\alpha = 1$, for all experiments. Note that such connectivity can also be achieved with a less connected network by running multiple but fixed rounds of consensus steps. Fig. 4.3 compares the iteration complexity as β increases, averaged over 100 Monte-Carlo realizations. We can see that for small β SONATA-F converges faster than SONATA-L; while for large β SONATA-L is faster. This can be explained using our results in Corollaries 4.2.2 and 4.2.4. As the complexity of SONATA-F and SONATA-L scales proportionally to β/μ and κ_g, respectively, when β/μ is comparatively smaller than κ_g, SONATA-F enjoys a better rate. But as β/μ increases, the rate deteriorates and eventually gets worse than that of SONATA-L.

4.5 Conclusions

In this chapter, we studied distributed multiagent optimization over (directed, time-varying) graphs. We considered the minimization of $F + G$ subject to convex constraints, where F is the smooth strongly convex sum of the agent's losses and G is a nonsmooth convex function. We build on the SONATA algorithm: the algorithm employs the use of surrogate objective functions in the agents' subproblems (going thus beyond linearization, such as proximal-gradient) coupled with a perturbed (push-sum) consensus mechanism that aims to

track locally the gradient of F. SONATA achieves precision $\epsilon > 0$ on the objective value in $\mathcal{O}(\kappa_g \log(1/\epsilon))$ gradient computations at each node and $\tilde{\mathcal{O}}\big(\kappa_g(1-\rho)^{-1/2}\log(1/\epsilon)\big)$ communication steps, where κ_g is the condition number of F and ρ characterizes the connectivity of the network. This is the first linear rate result for distributed composite optimization; it also improves on existing (non-accelerated) schemes just minimizing F, whose rate depends on much larger quantities than κ_g (e.g., the worst-case condition number among the agents). When considering in particular empirical risk minimization problems with statistically similar data across the agents, SONATA employing high-order surrogates achieves precision $\epsilon > 0$ in $\mathcal{O}\big((\beta/\mu)\log(1/\epsilon)\big)$ iterations and $\tilde{\mathcal{O}}\big((\beta/\mu)(1-\rho)^{-1/2}\log(1/\epsilon)\big)$ communication steps, where β measures the degree of similarity of the agents' losses and μ is the strong convexity constant of F. Therefore, when $\beta/\mu < \kappa_g$, the use of high-order surrogates yields provably faster rates than what achievable by first-order models; this is without exchanging any Hessian matrix over the network.

4.6 Proof of technical results

4.6.1 Proof of (4.55)

Chaining the inequalities in (4.51) as shown in Fig. 4.2, we have

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

$$+ C_1 \cdot \omega_p + (C_1 \cdot G_P(\alpha, z) \cdot 2 + C_2) \cdot \omega_y + C_1 \cdot G_P(\alpha, z) \cdot 4L_{\text{mx}}^2 \cdot \omega_x$$
$$+ (C_1 \cdot G_P(\alpha, z) \cdot 2 + C_2) \cdot G_Y(z) \cdot 8L_{\text{mx}}^2 \rho^2 \cdot \omega_x.$$

Notice that, under (4.52), $G_P(\alpha, z)$, $G_X(z)$, $G_Y(z)$, and ω_p, ω_x, ω_y are all bounded, which implies that the reminder $\mathcal{R}(\alpha, z)$ in (4.51) is bounded as well. □

4.6.2 Proof of Theorem 4.2.2

We find the smallest z satisfying (4.52) such that $\mathcal{P}(\alpha, z) < 1$, for $\alpha \in (0, \alpha_{\text{mx}})$, with $\alpha_{\text{mx}} \in (0, 1)$ to be determined.

Let us begin considering the condition $z > \sigma(\alpha)$ in (4.52). To simplify the analysis, we impose instead the following stronger version

$$z \geq \sigma(\alpha) + \frac{(\theta \cdot \alpha) \cdot \left(\left(1 - \frac{\alpha}{2}\right) \widetilde{\mu}_{\text{mn}} + \frac{D_{\text{mn}}^{\ell}}{2} \alpha - \frac{1}{2} \epsilon_{opt}\right)}{\frac{D_{\text{mx}}^2}{\mu} + \left(1 - \frac{\alpha}{2}\right) \widetilde{\mu}_{\text{mn}} + \frac{D_{\text{mn}}^{\ell}}{2} \alpha - \frac{1}{2} \epsilon_{opt}} \tag{4.98}$$

for some $\theta \in (0, 1)$, which will be chosen to tighten the bound. Notice that the RHS of (4.98) is strictly larger than $\sigma(\alpha)$ but still strictly less than one, for any $\alpha \in (0, (2\widetilde{\mu}_{\text{mn}} - \epsilon_{opt})/(\widetilde{\mu}_{\text{mn}} - D_{\text{mn}}^{\ell}))$, with given $\epsilon_{opt} \in (0, 2\widetilde{\mu}_{\text{mn}})$.

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 (4.98), $G_P(\alpha, z)$ can be upper bounded as

$$\begin{aligned} G_P(\alpha, z) &\leq \inf_{\epsilon_{opt} \in (0, 2\widetilde{\mu}_{\text{mn}} - \alpha(\widetilde{\mu}_{\text{mn}} - D_{\text{mn}}^{\ell}))} \frac{\frac{1}{2}\epsilon_{opt}^{-1} \cdot \frac{D_{\text{mx}}^2}{\mu} + \frac{1}{\mu} \cdot \left(\left(1 - \frac{\alpha}{2}\right) \widetilde{\mu}_{\text{mn}} + \frac{D_{\text{mn}}^{\ell}}{2} \alpha - \frac{1}{2} \epsilon_{opt}\right)}{\left(1 - \frac{\alpha}{2}\right) \widetilde{\mu}_{\text{mn}} + \frac{D_{\text{mn}}^{\ell}}{2} \alpha - \frac{1}{2} \epsilon_{opt}} \cdot \theta^{-1} \\ &= G_P^{\star}(\alpha) \cdot \theta^{-1}, \end{aligned} \tag{4.99}$$

where the minimum is attained at $\epsilon_{opt}^{\star} \triangleq \widetilde{\mu}_{\text{mn}} - \frac{\alpha}{2}(\widetilde{\mu}_{\text{mn}} - D_{\text{mn}}^{\ell})$; and $G_P^{\star}(\alpha)$ is defined in (4.60). Substituting the upper bound (4.99) in $\mathcal{P}(\alpha, z)$ and setting therein $\epsilon_{opt} = \epsilon_{opt}^{\star}$, we get the following sufficient condition for $\mathcal{P}(\alpha, z) < 1$:

$$G_P^{\star}(\alpha) \cdot \theta^{-1} \cdot C_1 \cdot 4L_{\text{mx}}^2 \cdot G_X(z) \cdot \rho^2 \cdot \alpha^2$$

$$+\left(G_P^\star(\alpha)\cdot\theta^{-1}\cdot 2C_1+C_2\right)\cdot G_Y(z)\cdot 2L_{\rm mx}^2\rho^2\cdot\alpha^2$$
$$+\left(G_P^\star(\alpha)\cdot\theta^{-1}\cdot 2C_1+C_2\right)\cdot G_Y(z)\cdot 8L_{\rm mx}^2\rho^2\cdot G_X(z)\cdot\rho^2\cdot\alpha^2<1. \quad (4.100)$$

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 (4.100) is

$$\alpha\le\alpha(z)\triangleq\left(A_{1,\theta}\frac{1}{(\sqrt{z}-\rho)^2}+A_{2,\theta}\frac{1}{(\sqrt{z}-\rho)^2}+A_{3,\theta}\frac{1}{(\sqrt{z}-\rho)^4}\right)^{-1/2}, \quad (4.101)$$

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

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

Condition (4.101) shows the rate z must satisfy

$$z\ge\left(\rho+\sqrt{A_\theta\alpha}\right)^2, \quad \text{with} \quad A_\theta\triangleq\sqrt{A_{1,\theta}+A_{2,\theta}+A_{3,\theta}}. \quad (4.102)$$

Notice that, under $\epsilon_x=\epsilon_y=(\sqrt{z}-\rho)/\rho$, (4.102) implies $z>\rho^2(1+\epsilon_x)=\rho^2(1+\epsilon_y)=\rho\sqrt{z}$, which are the other two conditions on z in (4.52). Therefore, overall, z must satisfy (4.98) and (4.102). Letting $\epsilon_{opt}=\epsilon_{opt}^\star$ in (4.98), the condition simplifies to

$$z\ge 1-\frac{\widetilde{\mu}_{\rm mn}-\frac{\alpha}{2}(\widetilde{\mu}_{\rm mn}-D_{\rm mn}^\ell)}{\frac{2D_{\rm mx}^2}{\mu}+\widetilde{\mu}_{\rm mn}-\frac{\alpha}{2}(\widetilde{\mu}_{\rm mn}-D_{\rm mn}^\ell)}\cdot(1-\theta)\alpha.$$

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

$$\bar{z}=\inf_{\theta\in(0,1)}\max\left\{\left(\rho+\sqrt{A_\theta\alpha}\right)^2,1-\frac{\widetilde{\mu}_{\rm mn}-\frac{\alpha}{2}(\widetilde{\mu}_{\rm mn}-D_{\rm mn}^\ell)}{\frac{2D_{\rm mx}^2}{\mu}+\widetilde{\mu}_{\rm mn}-\frac{\alpha}{2}(\widetilde{\mu}_{\rm mn}-D_{\rm mn}^\ell)}\cdot(1-\theta)\alpha\right\}. \quad (4.103)$$

Finally, we further simplify (4.103). Letting $\theta = 1/2$ and using $\alpha \in (0, \widetilde{\mu}_{\mathrm{mn}}/(\widetilde{\mu}_{\mathrm{mn}} - D^{\ell}_{\mathrm{mn}})]$, the second term in (4.103) can be upper bounded by

$$1 - \underbrace{\frac{\widetilde{\mu}_{\mathrm{mn}}\mu}{4D^2_{\mathrm{mx}} + \widetilde{\mu}_{\mathrm{mn}}\mu} \cdot \frac{1}{2}}_{\triangleq J} \alpha. \tag{4.104}$$

The condition $\bar{z} < 1$ imposes the following upper bound on α: $\alpha < \alpha_{\mathrm{mx}} = \min\{(1-\rho)^2/A_{\frac{1}{2}}, \widetilde{\mu}_{\mathrm{mn}}/(\widetilde{\mu}_{\mathrm{mn}} - D^{\ell}_{\mathrm{mn}}), 1\}$. Eq. (4.103) then simplifies to

$$\bar{z} = \max\left\{\left(\rho + \sqrt{\alpha A_{\frac{1}{2}}}\right)^2, 1 - J\alpha\right\}. \tag{4.105}$$

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

$$\alpha^* = \left(\frac{-\rho\sqrt{A_{\frac{1}{2}}} + \sqrt{A_{\frac{1}{2}} + J(1-\rho^2)}}{A_{\frac{1}{2}} + J}\right)^2. \tag{4.106}$$

To conclude, given the step-size satisfying $\alpha \in (0, \alpha_{\mathrm{mx}})$, the sequence $\{\|d^\nu\|^2\}$ converges at rate $\mathcal{O}(z^\nu)$, with z given in (4.62). □

4.6.3 Proof of Corollary 4.2.1

Since $W = J$, we have $\delta^\nu = 0$; then (4.34a) and (4.36) reduce to

$$p^{\nu+1} \leq p^\nu - \left(\left(1 - \frac{\alpha}{2}\right)\widetilde{\mu}_{\mathrm{mn}} + \frac{\alpha D^{\ell}_{\mathrm{mn}}}{2}\right)\alpha\|d^\nu\|^2 \tag{4.107}$$

and

$$\alpha\,\|d^\nu\|^2 \geq \frac{2\mu}{D^2_{\mathrm{mx}}}\left(p^{\nu+1} - (1-\alpha)p^\nu\right), \tag{4.108}$$

respectively. Combining (4.107) and (4.108) and using $\alpha < 2\widetilde{\mu}_{\rm mn}/(\widetilde{\mu}_{\rm mn} - D_{\rm mn})$, yield

$$p^{\nu+1} \leq \left(1 - \alpha \cdot \frac{\left(1-\frac{\alpha}{2}\right)\widetilde{\mu}_{\rm mn} + \frac{\alpha D^{\ell}_{\rm mn}}{2}}{\frac{D^2_{\rm mx}}{2\mu} + \left(1-\frac{\alpha}{2}\right)\widetilde{\mu}_{\rm mn} + \frac{\alpha D^{\ell}_{\rm mn}}{2}}\right) p^{\nu}, \tag{4.109}$$

which proves (4.65).

We customize next (4.65) to the specific choices of the surrogate functions.

• **Linearization:** Consider the choice of $\widetilde{f}_{\rm i}$ as in (4.63). 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 in (4.65), we obtain $z \leq 1 - \kappa_g^{-1}$.

• **Local $f_{\rm i}$:** Consider now $\widetilde{f}_{\rm i}$ as in (4.64). By $\nabla^2 f_{\rm i}(x) \succeq 0$, for all $x \in \mathcal{K}$, and Definition 4.1.1, we have $0 \preceq \nabla^2 \widetilde{f}_{\rm i}(x,y) - \nabla^2 F(x) \preceq 2\beta I$, for all $x, y \in \mathcal{K}$. 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 (4.65), yields

$$z \begin{cases} = 1 - \alpha \cdot \frac{\beta\left(1-\frac{\alpha}{2}\right)}{\frac{2\beta^2}{\mu} + \beta\left(1-\frac{\alpha}{2}\right)}, & \text{if } \mu \leq \beta \\ \leq 1 - \alpha \cdot \frac{\mu\left(1-\frac{\alpha}{2}\right)}{\frac{2\beta^2}{\mu} + \mu\left(1-\frac{\alpha}{2}\right)}, & \text{if } \mu > \beta. \end{cases} \tag{4.110}$$

Finally, setting $\alpha = \min\{1, 2\widetilde{\mu}_{\rm mn}/((\mu-\beta)_+ + \beta)\} = 1$ in the expression above, yields (4.66). □

4.6.4 Proof of Corollary 4.2.2

According to Theorem 4.2.2, the rate z can be bounded as

$$z \leq \max\{z_1, z_2\}, \quad \text{with} \quad z_1 \triangleq 1 - \alpha \cdot J \ \text{ and } \ z_2 \triangleq \left(\rho + \sqrt{\alpha A_{\frac{1}{2}}}\right)^2, \tag{4.111}$$

where J and $A_{\frac{1}{2}}$ are defined in (4.104) and (4.102), respectively.

The proof consists in bounding properly z_1 and z_2 based upon the surrogate (4.63) postulated in the corollary. We begin particularizing the expressions of J and $A_{\frac{1}{2}}$. Since $\nabla^2 \widetilde{f}_{\rm i}(x_{\rm i}; x_{\rm i}^{\nu}) = L$, one can set $\widetilde{\mu}_{\rm mn} = L$, and (4.16) holds with $D^{\ell}_{\rm mn} = 0$ and $D_{\rm mx} = L - \mu$. Furthermore, by Assumption 4.1.1, it follows that $\beta \geq \lambda_{\max}(\nabla^2 f_{\rm i}(x)) - L$, for all $x \in \mathcal{K}$;

hence, one can set $L_{\mathrm{mx}} = L+\beta$. Next, we will substitute the above values into the expressions of J and $A_{\frac{1}{2}}$.

To do so, we need to particularize first the quantities $G_P^\star\left(\frac{\widetilde{\mu}_{\mathrm{mn}}}{\widetilde{\mu}_{\mathrm{mn}}-D_{\mathrm{mn}}^\ell}\right)$ [cf. (4.60)], C_1 and C_2 [cf. (4.53d)]:

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

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

$$
J = \frac{1}{2}\frac{\kappa_g}{4(\kappa_g-1)^2+\kappa_g} \in \left[\frac{1}{8\kappa_g}, \frac{1}{2}\right], \tag{4.112}
$$

and

$$
\begin{aligned}
&(A_{\frac{1}{2}})^2\\
=&\, G_P^\star(1)\cdot 2\cdot C_1\cdot 4L_{\mathrm{mx}}^2\cdot\rho^2 + (G_P^\star(1)\cdot 4\cdot C_1 + C_2)\cdot 2L_{\mathrm{mx}}^2\rho^2\\
&+ (G_P^\star(1)\cdot 4\cdot C_1 + C_2)\cdot 8L_{\mathrm{mx}}^2\rho^4\\
=&\, (24G_P^\star(1)\cdot C_1 + 5C_2)\cdot 2L_{\mathrm{mx}}^2\rho^2\\
=&\left(24\cdot\frac{4(L-\mu)^2+L^2}{\mu L^2}\cdot\frac{6}{\mu L^2}\left((2L-\mu)^2+4(L+\beta)^2\right)+20L^{-2}\right)\cdot 2(L+\beta)^2\rho^2\\
\le&\left(24\cdot\frac{5}{\mu}\cdot\frac{24}{\mu L^2}\left(L^2+(L+\beta)^2\right)+20L^{-2}\right)\cdot 2(L+\beta)^2\rho^2\\
=&\left(24\cdot 24\cdot 5\left(1+\left(1+\frac{\beta}{L}\right)^2\right)\left(1+\frac{\beta}{L}\right)^2\kappa_g^2+20\left(1+\frac{\beta}{L}\right)^2\right)\cdot 2\rho^2\\
\le&110^2\cdot\kappa_g^2\left(1+\frac{\beta}{L}\right)^4\rho^2,
\end{aligned}
\tag{4.113}
$$

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

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

By (4.113), we have

$$z_2 \leq \bar{z}_2 \triangleq \left(\rho + \sqrt{\alpha M \rho}\right)^2, \quad \text{with} \quad M \triangleq 110 \cdot \kappa_g (1+\beta/L)^2. \tag{4.114}$$

Since $\alpha \in (0,1]$ must be chosen so that $z \in (0,1]$, we impose $\max\{z_1, \bar{z}_2\} < 1$, implying $\alpha \leq \min\{J^{-1}, (1-\rho)^2/(M\rho), 1\}$. Since $J^{-1} > 1$ [cf. (4.112)], the condition on α reduces to $\alpha \leq \alpha_{\text{mx}} \triangleq \min\{(1-\rho)^2/(M\rho), 1\}$. Choose $\alpha = c \cdot \alpha_{\text{mx}}$, for some given $c \in (0,1)$. Depending on the value of ρ, either $\alpha_{\text{mx}} = 1$ or $\alpha_{\text{mx}} = (1-\rho)^2/(M\rho)$.

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

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

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

Therefore, z can be bounded as

$$\begin{aligned}
z \leq \max\{z_1, \bar{z}_2\} &\leq 1 - c \cdot \left(1 - \sqrt{c}\right)^2 \cdot (1-1/110)^2 \cdot J \\
&\leq 1 - c \cdot \left(1 - \sqrt{c}\right)^2 \cdot (1-1/110)^2 \cdot \frac{1}{8\kappa_g}.
\end{aligned} \tag{4.115}$$

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

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

We claim that $(J\,c)/(M\rho) < 1$. Suppose this is not the case, that is, $M\rho \leq Jc$. Since $Jc < 1/2$ [cf. (4.112)] and $M \geq 110\,\kappa$, $M\rho \leq Jc$ would imply $\rho < 1/(220\kappa_g)$. This however is in contradiction with the assumption $M\rho \geq (1-\rho)^2$, as it would lead to $1/2 > M\rho \geq (1-\rho)^2 > (1-1/(220\kappa_g))^2$.

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

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

4.6.5 Proof of Corollaries 4.2.3 and 4.2.4

We follow similar steps as in Appendix 4.6.4 but customized to the surrogate (4.64). We begin particularizing the expressions of J and $A_{\frac{1}{2}}$.

In the setting of the corollary, we have: $\nabla^2 \widetilde{f}_{\mathrm{i}}(x;y) = \nabla^2 f_{\mathrm{i}}(x) + \beta I$, for all $y \in \mathcal{K}$; $\nabla^2 f_{\mathrm{i}}(x) \succeq 0$, for all $x \in \mathcal{K}$; and, by Assumption 4.1.1, $0 \preceq \nabla^2 \widetilde{f}_{\mathrm{i}}(x,y) - \nabla^2 F(x) \preceq 2\beta I$, for all $x, y \in \mathcal{K}$. Therefore, we can set $D^{\ell}_{\mathrm{mn}} = 0$, $D_{\mathrm{mx}} = 2\beta$, $\widetilde{\mu}_{\mathrm{mn}} = \beta + (\mu-\beta)_+ = \max\{\beta, \mu\}$, and $L_{\mathrm{mx}} = L+\beta$. Using these values, $G^{\star}_P\left(\frac{\widetilde{\mu}_{\mathrm{mn}}}{\widetilde{\mu}_{\mathrm{mn}} - D^{\ell}_{\mathrm{mn}}}\right)$, C_1, and C_2 can be simplified as follows:

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

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

$$J = \frac{1}{2}\frac{1}{1 + 16\left(\frac{\beta}{\mu}\right) \cdot \min\left\{1, \frac{\beta}{\mu}\right\}}, \tag{4.116}$$

and

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

where

$$
M = \begin{cases}253\left(1+\frac{L}{\beta}\right)\left(\kappa_g+\frac{\beta}{\mu}\right), & \beta>\mu,\\ 193\left(1+\frac{\beta}{\mu}\right)^2\left(\kappa_g+\frac{\beta}{\mu}\right)^2, & \beta\leq\mu.\end{cases} \tag{4.117}
$$

Similarly to the proof of Corollary 4.2.2, we bound $z\leq\max\{z_1,z_2\}$ as

$$
z\leq\max\{z_1,\bar{z}_2\},\quad \text{with}\quad z_1\triangleq 1-\alpha\cdot J\ \text{ and }\ \bar{z}_2\triangleq\left(\rho+\sqrt{\alpha\,M\,\rho}\right)^2, \tag{4.118}
$$

where J and M are now given by (4.116) and (4.117), respectively. For $\max\{z_1,z_2\}<1$, we require $\alpha\leq\alpha_{\mathrm{mx}}\triangleq\min\{1,(1-\rho)^2/(M\rho)\}$, and choose $\alpha=c\cdot\alpha_{\mathrm{mx}}$, with arbitrary $c\in(0,1)$. We study separately the cases $\beta>\mu$ and $\beta\leq\mu$.

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

$$
M=253\left(1+\frac{L}{\beta}\right)\left(\kappa_g+\frac{\beta}{\mu}\right)\quad\text{and}\quad J=\frac{1}{2}\frac{1}{1+16\,(\beta/\mu)}\geq\frac{1}{34(\beta/\mu)}. \tag{4.119}
$$

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

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

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

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

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

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

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

Using the same argument as in the proof of Corollary 4.2.2–Case II, one can show that $(c\,J)/(M\rho) < 1$. Therefore,

$$\begin{aligned} z \leq \max\{z_1, \bar{z}_2\} &\leq 1 - \left(1 - \sqrt{c}\right)^2 \cdot c\,J \cdot \frac{(1-\rho)^2}{M\rho} \\ &\overset{(4.119)}{\leq} 1 - c \cdot \left(1 - \sqrt{c}\right)^2 \cdot \frac{1}{34} \cdot \frac{(1-\rho)^2}{253\left(\kappa_g + \frac{\beta}{\mu}\right)^2 \rho}. \end{aligned}$$

2) $\beta \leq \mu$**.** In this case we have

$$M = 193\left(1 + \frac{\beta}{\mu}\right)^2 \left(\kappa_g + \frac{\beta}{\mu}\right)^2 \quad \text{and} \quad J = \frac{1}{2}\frac{1}{1 + 16\,(\beta/\mu)^2}. \tag{4.120}$$

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

$$z \leq \max\{z_1, \bar{z}_2\} \leq 1 - c \cdot \left(1 - \sqrt{c}\right)^2 \cdot \left(1 - \frac{1}{193}\right)^2 \cdot \frac{1}{2 + 32\left(\frac{\beta}{\mu}\right)^2}. \tag{4.121}$$

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

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

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

4.6.6 Proof of Proposition 4.3.1

We begin introducing some intermediate results.

Lemma 4.6.1. *Consider Problem (4.1) under Assumption 4.1.1; and SONATA (Algorithm 4) under Assumptions 4.2.1 and 4.3.1. Then, there holds*

$$
U(x_{\mathrm{i}}^{\nu+\frac{1}{2}}) \leq U(x_{\mathrm{i}}^{\nu}) - \alpha\left(\left(1-\frac{\alpha}{2}\right)\tilde{\mu}_{\mathrm{i}} + \frac{\alpha}{2}\cdot D_{\mathrm{i}}^{\ell}\right)\|d_{\mathrm{i}}^{\nu}\|^2 + \alpha\|d_{\mathrm{i}}^{\nu}\|\|\delta_{\mathrm{i}}^{\nu}\|, \tag{4.122}
$$

with $\delta_{\mathrm{i}}^{\nu}$ defined in (4.24).

Proof. Consider the Taylor expansion of F:

$$
\begin{aligned}
F(x_{\mathrm{i}}^{\nu+\frac{1}{2}}) &= F(x_{\mathrm{i}}^{\nu}) + \nabla F(x_{\mathrm{i}}^{\nu})^{\top}(\alpha d_{\mathrm{i}}^{\nu}) + (\alpha d_{\mathrm{i}}^{\nu})^{\top} H(\alpha d_{\mathrm{i}}^{\nu}), \\
&\overset{(4.24)}{=} F(x_{\mathrm{i}}^{\nu}) + \left(\delta_{\mathrm{i}}^{\nu}\right)^{\top}(\alpha d_{\mathrm{i}}^{\nu}) + \left(y_{\mathrm{i}}^{\nu}\right)^{\top}(\alpha d_{\mathrm{i}}^{\nu}) + (\alpha d_{\mathrm{i}}^{\nu})^{\top} H(\alpha d_{\mathrm{i}}^{\nu}),
\end{aligned} \tag{4.123}
$$

where $H \triangleq \int_0^1 (1-\theta)\nabla^2 F(\theta x_{\mathrm{i}}^{\nu+\frac{1}{2}} + (1-\theta)x_{\mathrm{i}}^{\nu})d\theta$.

Invoking the optimality of $\widehat{x}_{\mathrm{i}}^{\nu}$, we have

$$
G(x_{\mathrm{i}}^{\nu}) - G(\widehat{x}_{\mathrm{i}}^{\nu}) \geq (d_{\mathrm{i}}^{\nu})^{\top}\left(\nabla \widetilde{f}_{\mathrm{i}}(\widehat{x}_{\mathrm{i}}^{\nu}; x_{\mathrm{i}}^{\nu}) + y_{\mathrm{i}}^{\nu} - \nabla f_{\mathrm{i}}(x_{\mathrm{i}}^{\nu})\right) = (d_{\mathrm{i}}^{\nu})^{\top}\left(y_{\mathrm{i}}^{\nu} + \widetilde{H}_{\mathrm{i}} d_{\mathrm{i}}^{\nu}\right) \tag{4.124}
$$

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

Substituting (4.124) in (4.123) and using the convexity of G yield

$$\begin{aligned}
&F(x_{\mathrm{i}}^{\nu+\frac{1}{2}}) \\
&\leq F(x_{\mathrm{i}}^{\nu}) + (\delta_{\mathrm{i}}^{\nu})^{\top}(\alpha d_{\mathrm{i}}^{\nu}) + (\alpha d_{\mathrm{i}}^{\nu})^{\top} H(\alpha d_{\mathrm{i}}^{\nu}) + \alpha\left(G(x_{\mathrm{i}}^{\nu}) - G(\widehat{x}_{\mathrm{i}}^{\nu}) - (d_{\mathrm{i}}^{\nu})^{\top}\widetilde{H}_{\mathrm{i}} d_{\mathrm{i}}^{\nu}\right) \\
&\leq F(x_{\mathrm{i}}^{\nu}) + (\delta_{\mathrm{i}}^{\nu})^{\top}(\alpha d_{\mathrm{i}}^{\nu}) + \alpha\left(-(d_{\mathrm{i}}^{\nu})^{\top}\widetilde{H}_{\mathrm{i}} d_{\mathrm{i}}^{\nu} + (\alpha d_{\mathrm{i}}^{\nu})^{\top} H(d_{\mathrm{i}}^{\nu})\right) + G(x_{\mathrm{i}}^{\nu}) - G(x_{\mathrm{i}}^{\nu+\frac{1}{2}}).
\end{aligned} \tag{4.125}$$

It remains to bound $\alpha H - \widetilde{H}_{\mathrm{i}}$. We proceed as follows:

$$\begin{aligned}
&\alpha H - \widetilde{H}_{\mathrm{i}} \\
&= \alpha\int_0^1 (1-\theta)\nabla^2 F(\theta x_{\mathrm{i}}^{\nu+\frac{1}{2}} + (1-\theta)x_{\mathrm{i}}^{\nu})d\theta - \int_0^1 \nabla^2 \widetilde{f}_{\mathrm{i}}(\theta\widehat{x}_{\mathrm{i}}^{\nu} + (1-\theta)x_{\mathrm{i}}^{\nu}; x_{\mathrm{i}}^{\nu})d\theta \\
&\overset{(4.12b)}{=} \int_0^{\alpha} (1-\theta/\alpha)\nabla^2 F(\theta\widehat{x}_{\mathrm{i}}^{\nu} + (1-\theta)x_{\mathrm{i}}^{\nu})d\theta - \int_0^1 \nabla^2 \widetilde{f}_{\mathrm{i}}(\theta\widehat{x}_{\mathrm{i}}^{\nu} + (1-\theta)x_{\mathrm{i}}^{\nu}; x_{\mathrm{i}}^{\nu})d\theta \\
&\overset{(a)}{\preceq} -\int_0^{\alpha} (1-\theta/\alpha)\cdot(D_{\mathrm{i}}^{\ell})\, I\, d\theta - \int_0^{\alpha} (\theta/\alpha)\nabla^2 \widetilde{f}_{\mathrm{i}}(\theta\widehat{x}_{\mathrm{i}} + (1-\theta)x_{\mathrm{i}}^{\nu}; x_{\mathrm{i}}^{\nu})d\theta \\
&\quad - \int_{\alpha}^1 \nabla^2 \widetilde{f}_{\mathrm{i}}(\theta\,\widehat{x}_{\mathrm{i}}^{\nu} + (1-\theta)\,x_{\mathrm{i}}^{\nu}; x_{\mathrm{i}}^{\nu})d\theta \\
&\overset{(b)}{\preceq} -\frac{1}{2}\alpha\,(D_{\mathrm{i}}^{\ell})\, I - \left(1-\frac{\alpha}{2}\right)\widetilde{\mu}_{\mathrm{i}}\, I,
\end{aligned} \tag{4.126}$$

where in (a) we used $\nabla^2 F(\theta\widehat{x}_{\mathrm{i}}^{\nu} + (1-\theta)x_{\mathrm{i}}^{\nu}) \preceq -(D_{\mathrm{i}}^{\ell})I + \nabla^2 \widetilde{f}_{\mathrm{i}}(\theta\widehat{x}_{\mathrm{i}}^{\nu} + (1-\theta)x_{\mathrm{i}}^{\nu}; x_{\mathrm{i}}^{\nu})$ [cf. (4.16)] while (b) follows from the fact that $\widetilde{f}_{\mathrm{i}}$ is $\widetilde{\mu}_{\mathrm{i}}$-strongly convex (cf. Assumption 4.2.1). Substituting (4.126) into (4.125) completes the proof □

We connect now the individual decreases in (4.122) with that of the optimality gap p_{ϕ}^{ν}, defined in (4.78). Notice that

$$\sum_{\mathrm{i}=1}^{m} \phi_{\mathrm{i}}^{\nu+1} U(x_{\mathrm{i}}^{\nu+1}) \leq \sum_{\mathrm{i}=1}^{m}\sum_{\mathrm{j}=1}^{m} c_{\mathrm{ij}}\phi_{\mathrm{j}}^{\nu} U\left(x_{\mathrm{j}}^{\nu+\frac{1}{2}}\right) = \sum_{\mathrm{i}=1}^{m} \phi_{\mathrm{i}}^{\nu} U(x_{\mathrm{i}}^{\nu+\frac{1}{2}}), \tag{4.127}$$

due to the convexity of U, column-stochasticity of $\{c_{\mathrm{ij}}^{\nu}\}_{\mathrm{i,j}}$ and $\sum_{\mathrm{j}=1}^{m} c_{\mathrm{ij}}^{\nu}\phi_{\mathrm{j}}^{\nu}/\phi_{\mathrm{i}}^{\nu+1} = 1$, for all $\mathrm{i} = 1, \ldots, m$. Summing (4.122) over $\mathrm{i} = 1, \ldots m$, and using (4.127), we obtain

$$\begin{aligned}
p_{\phi}^{\nu+1} &\leq p_{\phi}^{\nu} + \sum_{\mathrm{i}=1}^{m} \phi_{\mathrm{i}}^{\nu}\left\{\alpha\|d_{\mathrm{i}}^{\nu}\|\|\delta_{\mathrm{i}}^{\nu}\| - \alpha\left(1-\frac{\alpha}{2}\right)\widetilde{\mu}_{\mathrm{i}}\|d_{\mathrm{i}}^{\nu}\|^2 - \frac{D_{\mathrm{i}}^{\ell}}{2}\alpha^2\|d_{\mathrm{i}}^{\nu}\|^2\right\} \\
&\overset{(a)}{\leq} p_{\phi}^{\nu} - \left(\left(1-\frac{\alpha}{2}\right)\widetilde{\mu}_{\mathrm{mn}} + \frac{\alpha D_{\mathrm{mn}}}{2} - \frac{1}{2}\epsilon_{opt}\right)\alpha\sum_{\mathrm{i}=1}^{m}\phi_{\mathrm{i}}^{\nu}\|d_{\mathrm{i}}^{\nu}\|^2 + \frac{1}{2}\epsilon_{opt}^{-1}\,\alpha\cdot\phi_{ub}\cdot\|\delta^{\nu}\|^2,
\end{aligned} \tag{4.128}$$

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

$$\left(1 - \frac{\alpha}{2}\right)\tilde{\mu}_{\mathrm{mn}} + \frac{\alpha D_{\mathrm{mn}}^{L}}{2} - \frac{1}{2}\epsilon_{opt} > 0. \tag{4.129}$$

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

Lemma 4.6.2. *In the setting of Lemma 4.2.1, there holds:*

$$\alpha \sum_{\mathrm{i}=1}^{m} \phi_{\mathrm{i}}^{\nu} \|d_{\mathrm{i}}^{\nu}\|^2 \geq \frac{\mu}{D_{\mathrm{mx}}^2} \left(p_{\phi}^{\nu+1} - (1-\alpha) p_{\phi}^{\nu} - \frac{\alpha}{\mu} \sum_{\mathrm{i}=1}^{m} \phi_{\mathrm{i}}^{\nu} \|\delta_{\mathrm{i}}^{\nu}\|^2 \right) \tag{4.130}$$

with D_{mx} defined in (4.25).

Proof. Invoking the optimality condition of $\widehat{x}_{\mathrm{i}}^{\nu}$, yields

$$G(x^{\star}) - G(\widehat{x}_{\mathrm{i}}^{\nu}) \geq -(x^{\star} - \widehat{x}_{\mathrm{i}}^{\nu})^{\top} \Big(\nabla \widetilde{f}_{\mathrm{i}}(\widehat{x}_{\mathrm{i}}^{\nu}; x_{\mathrm{i}}^{\nu}) + y_{\mathrm{i}}^{\nu} - \nabla f_{\mathrm{i}}(x_{\mathrm{i}}^{\nu}) \Big). \tag{4.131}$$

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

$$\begin{aligned}
U(x^{\star}) &\geq U(\widehat{x}_{\mathrm{i}}^{\nu}) + G(x^{\star}) - G(\widehat{x}_{\mathrm{i}}^{\nu}) + \nabla F(\widehat{x}_{\mathrm{i}}^{\nu})^{\top}(x^{\star} - \widehat{x}_{\mathrm{i}}^{\nu}) + \frac{\mu}{2}\|x^{\star} - \widehat{x}_{\mathrm{i}}^{\nu}\|^2 \\
&\overset{(4.37)}{\geq} U(\widehat{x}_{\mathrm{i}}^{\nu}) + \Big(\nabla F(\widehat{x}_{\mathrm{i}}^{\nu}) - \nabla \widetilde{f}_{\mathrm{i}}(\widehat{x}_{\mathrm{i}}^{\nu}; x_{\mathrm{i}}^{\nu}) - \Big(y_{\mathrm{i}}^{\nu} - \nabla f_{\mathrm{i}}(x_{\mathrm{i}}^{\nu}) \Big) \Big)^{\top} (x^{\star} - \widehat{x}_{\mathrm{i}}^{\nu}) + \frac{\mu}{2}\|x^{\star} - \widehat{x}_{\mathrm{i}}^{\nu}\|^2 \\
&= U(\widehat{x}_{\mathrm{i}}^{\nu}) + \frac{\mu}{2} \Big\| x^{\star} - \widehat{x}_{\mathrm{i}}^{\nu} + \frac{1}{\mu} \Big(\nabla F(\widehat{x}_{\mathrm{i}}^{\nu}) - \nabla \widetilde{f}_{\mathrm{i}}(\widehat{x}_{\mathrm{i}}^{\nu}; x_{\mathrm{i}}^{\nu}) - \Big(y_{\mathrm{i}}^{\nu} - \nabla f_{\mathrm{i}}(x_{\mathrm{i}}^{\nu}) \Big) \Big) \Big\|^2 \\
&\quad - \frac{1}{2\mu} \Big\| \nabla F(\widehat{x}_{\mathrm{i}}^{\nu}) - \nabla \widetilde{f}_{\mathrm{i}}(\widehat{x}_{\mathrm{i}}^{\nu}; x_{\mathrm{i}}^{\nu}) - \Big(y_{\mathrm{i}}^{\nu} - \nabla f_{\mathrm{i}}(x_{\mathrm{i}}^{\nu}) \Big) \Big\|^2
\end{aligned}$$

$$\begin{aligned}
&\geq U(\widehat{x}_{\mathrm{i}}^{\nu}) - \frac{1}{2\mu} \Big\| \nabla F(\widehat{x}_{\mathrm{i}}^{\nu}) \pm \nabla F(x_{\mathrm{i}}^{\nu}) - \nabla \widetilde{f}_{\mathrm{i}}(\widehat{x}_{\mathrm{i}}^{\nu}; x_{\mathrm{i}}^{\nu}) - \Big(y_{\mathrm{i}}^{\nu} - \nabla f_{\mathrm{i}}(x_{\mathrm{i}}^{\nu}) \Big) \Big\|^2 \\
&\geq U(\widehat{x}_{\mathrm{i}}^{\nu}) - \frac{1}{\mu} \Big\| \nabla F(\widehat{x}_{\mathrm{i}}^{\nu}) - \nabla F(x_{\mathrm{i}}^{\nu}) + \nabla f_{\mathrm{i}}(x_{\mathrm{i}}^{\nu}) - \nabla \widetilde{f}_{\mathrm{i}}(\widehat{x}_{\mathrm{i}}^{\nu}; x_{\mathrm{i}}^{\nu}) \Big\|^2 - \frac{1}{\mu}\|\delta_{\mathrm{i}}^{\nu}\|^2 \\
&= U(\widehat{x}_{\mathrm{i}}^{\nu}) - \frac{1}{\mu} \Big\| \int_0^1 \Big(\nabla^2 F(\theta \widehat{x}_{\mathrm{i}}^{\nu} + (1-\theta) x_{\mathrm{i}}^{\nu}) - \nabla^2 \widetilde{f}_{\mathrm{i}}(\theta \widehat{x}_{\mathrm{i}}^{\nu} + (1-\theta) x_{\mathrm{i}}^{\nu}; x_{\mathrm{i}}^{\nu}) \Big) (d_{\mathrm{i}}^{\nu}) \, \mathrm{d}\theta \Big\|^2 - \frac{1}{\mu}\|\delta_{\mathrm{i}}^{\nu}\|^2 \\
&\geq U(\widehat{x}_{\mathrm{i}}^{\nu}) - \frac{D_{\mathrm{i}}^2}{\mu} \Big\| d_{\mathrm{i}}^{\nu} \Big\|^2 - \frac{1}{\mu}\|\delta_{\mathrm{i}}^{\nu}\|^2,
\end{aligned}$$

where $D_{\mathrm{i}} = \max\{|D_{\mathrm{i}}^{\ell}|, |D_{\mathrm{i}}^{u}|\}$.

Rearranging the terms and summing over $\mathrm{i} = 1, \dots, m$, yields

$$\sum_{\mathrm{i}=1}^{m} \phi_{\mathrm{i}}^{\nu} \|d_{\mathrm{i}}^{\nu}\|^2 \geq \frac{\mu}{D_{\mathrm{mx}}^2} \left(\sum_{\mathrm{i}=1}^{m} \phi_{\mathrm{i}}^{\nu} \Big(U(\widehat{x}_{\mathrm{i}}^{\nu}) - U(x^{\star}) \Big) - \frac{1}{\mu} \sum_{\mathrm{i}=1}^{m} \phi_{\mathrm{i}}^{\nu} \|\delta_{\mathrm{i}}^{\nu}\|^2 \right). \tag{4.132}$$

Using (4.28) in conjunction with $U(x_{\mathrm{i}}^{\nu+\frac{1}{2}}) \leq \alpha U(\widehat{x}_{\mathrm{i}}^{\nu}) + (1-\alpha) U(x_{\mathrm{i}}^{\nu})$ leads to

$$\alpha \sum_{\mathrm{i}=1}^{m} \phi_{\mathrm{i}}^{\nu} \left(U(\widehat{x}_{\mathrm{i}}^{\nu}) - U(x^{\star}) \right) \geq p_{\phi}^{\nu+1} - (1-\alpha) p_{\phi}^{\nu}. \tag{4.133}$$

Combining (4.132) with (4.133) yields the desired result (4.130). □

As last step, we upper bound $\|\delta^{\nu}\|^2$ in (4.34) in terms of the consensus errors $\|x_{\perp}^{\nu}\|^2$ and $\|y_{\perp}^{\nu}\|^2$.

Lemma 4.6.3. *The tracking error $\|\delta^{\nu}\|^2$ can be bounded as*

$$\|\delta^{\nu}\|^2 \leq 8 L_{\mathrm{mx}}^2 \|x_{\phi,\perp}^{\nu}\|^2 + 2 \|y_{\phi,\perp}^{\nu}\|^2., \tag{4.134}$$

where L_{mx} is defined in (4.5).

Proof.

$$\begin{aligned}
\|\delta^{\nu}\|^2 &\overset{(4.24)}{=} \sum_{\mathrm{i}=1}^{m} \|\nabla F(x_{\mathrm{i}}^{\nu}) \pm \bar{y}_{\phi}^{\nu} - y_{\mathrm{i}}^{\nu}\|^2 \\
&\overset{(4.75)}{=} \frac{1}{m^2} \sum_{\mathrm{i}=1}^{m} \Big\| \sum_{\mathrm{j}=1}^{m} \nabla f_{\mathrm{j}}(x_{\mathrm{i}}^{\nu}) - \sum_{\mathrm{j}=1}^{m} \nabla f_{\mathrm{j}}(x_{\mathrm{j}}^{\nu}) + m \cdot \bar{y}_{\phi}^{\nu} - m \cdot y_{\mathrm{i}}^{\nu} \Big\|^2 \\
&\overset{A2,(4.5)}{\leq} \frac{1}{m^2} \sum_{\mathrm{i}=1}^{m} \left(2m \sum_{\mathrm{j}=1}^{m} L_{\mathrm{mx}}^2 \|x_{\mathrm{i}}^{\nu} - x_{\mathrm{j}}^{\nu}\|^2 + 2m^2 \|\bar{y}_{\phi}^{\nu} - y_{\mathrm{i}}^{\nu}\|^2 \right) \\
&\leq 8 L_{\mathrm{mx}}^2 \|x_{\phi,\perp}^{\nu}\|^2 + 2 \|y_{\phi,\perp}^{\nu}\|^2.
\end{aligned}$$

□

The linear convergence of the optimality gap up to consensus errors as stated in Proposition follows readily multiplying (4.130) by $\left(1 - \frac{\alpha}{2}\right) \widetilde{\mu}_{\mathrm{mn}} + \frac{\alpha D_{\mathrm{mn}}}{2} - \frac{1}{2}\epsilon_{opt}$ and adding with (4.128) to cancel out $\|d^{\nu}\|$, and using (4.40) to bound $\|\delta^{\nu}\|^2$.

4.6.7 Proof of Theorem 4.3.1

Following the same steps as in the proof of Theorem 4.2.1, we derive the optimal ϵ_{opt} appearing in $\eta(\alpha)$ and $\sigma(\alpha)$:

$$\epsilon^{\star}_{opt} = \left(1 - \frac{\alpha}{2}\right)\widetilde{\mu}_{\mathrm{mn}} + \alpha D^{\ell}_{\mathrm{mn}}/2, \tag{4.135}$$

where α must satisfy

$$\alpha < 2\widetilde{\mu}_{\mathrm{mn}}/(\widetilde{\mu}_{\mathrm{mn}} - D^{\ell}_{\mathrm{mn}}). \tag{4.136}$$

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

$$\begin{aligned}\mathcal{P}^{\star}(\alpha,1) \triangleq\ & G^{\star}_P(\alpha)\cdot C_1 \cdot 8\phi_{ub} L^2_{\mathrm{mx}} \cdot \frac{2c_0^2\rho_B^2}{(1-\rho_B)^2}\alpha^2 \\ &+ (G^{\star}_P(\alpha)\cdot 2\phi_{ub}\cdot C_1 + C_2)\cdot 2m\phi_{lb}^{-2} L^2_{\mathrm{mx}} \cdot \frac{2c_0^2\rho_B^2}{(1-\rho_{\bar{B}})^2}\alpha^2 \\ &+ (G^{\star}_P(\alpha)\cdot 2\phi_{ub}\cdot C_1 + C_2)\cdot 8m\phi_{lb}^{-2} L^2_{\mathrm{mx}} \cdot \frac{4c_0^4\rho_B^4}{(1-\rho_{\bar{B}})^4}\alpha^2,\end{aligned} \tag{4.137}$$

where

$$G^{\star}_P(\alpha) \triangleq \frac{\frac{D^2_{\mathrm{mx}}}{\mu} + \frac{1}{\mu}\cdot\left(\left(1-\frac{\alpha}{2}\right)\widetilde{\mu}_{\mathrm{mn}} + \frac{D^{\ell}_{\mathrm{mn}}}{2}\alpha\right)^2}{\left(\left(1-\frac{\alpha}{2}\right)\widetilde{\mu}_{\mathrm{mn}} + \frac{D^{\ell}_{\mathrm{mn}}}{2}\alpha\right)^2}. \tag{4.138}$$

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

$$\alpha < \alpha_2 \triangleq \left(G^{\star}_P\left(\frac{\widetilde{\mu}_{\mathrm{mn}}}{\widetilde{\mu}_{\mathrm{mn}} - D^{\ell}_{\mathrm{mn}}}\right)\cdot C_1 \cdot 8\phi_{ub} L^2_{\mathrm{mx}} \cdot \frac{2c_0^2\rho_B^2}{(1-\rho_B)^2}\alpha^2 \right. \tag{4.139}$$

$$+ \left(G^{\star}_P\left(\frac{\widetilde{\mu}_{\mathrm{mn}}}{\widetilde{\mu}_{\mathrm{mn}} - D^{\ell}_{\mathrm{mn}}}\right)\cdot 2\phi_{ub}\cdot C_1 + C_2\right)\cdot 2m\phi_{lb}^{-2} L^2_{\mathrm{mx}} \cdot \frac{2c_0^2\rho_B^2}{(1-\rho_{\bar{B}})^2}\alpha^2 \tag{4.140}$$

$$\left. + \left(G^{\star}_P\left(\frac{\widetilde{\mu}_{\mathrm{mn}}}{\widetilde{\mu}_{\mathrm{mn}} - D^{\ell}_{\mathrm{mn}}}\right)\cdot 2\phi_{ub}\cdot C_1 + C_2\right)\cdot 8m\phi_{lb}^{-2} L^2_{\mathrm{mx}} \cdot \frac{4c_0^4\rho_B^4}{(1-\rho_{\bar{B}})^4}\right)^{-1/2}. \tag{4.141}$$

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

4.6.8 Explicit expression of the linear rate in time-varying directed networks

The following theorem provides an explicit expression of the convergence rate in Theorem 4.3.1, in terms of the step-size α; the constants J and $A_{\frac{1}{2}}$ therein are defined in (4.149) and (4.146) with $\theta = 1/2$, respectively.

Theorem 4.6.1. *In the setting of Theorem 4.3.1, suppose that the step-size α satisfies $\alpha \in (0, \alpha_{\rm mx})$, with $\alpha_{\rm mx} \triangleq \min\{(1-\rho_B)/A_{\frac{1}{2}}, , \widetilde{\mu}_{\rm mn}/(\widetilde{\mu}_{\rm mn} - D^{\ell}_{\rm mn}), 1\}$. Then $\{U(x^{\nu}_{\rm i})\}$ converges to $U^{\star}$ at the R-linear rate $\mathcal{O}(z^{\nu})$, for all ${\rm i} = 1, \ldots, m$, where*

$$z = \begin{cases} 1 - J \cdot \alpha, & \text{if } \alpha \in (0, \min\{\alpha^*, \alpha_{\rm mx}\}), \\ \rho_B + A_{\frac{1}{2}}\alpha, & \text{if } \alpha \in [\min\{\alpha^*, \alpha_{\rm mx}\}, \alpha_{\rm mx}). \end{cases} \tag{4.142}$$

Proof. The proof follows similar steps as the proof of Theorem 4.2.2. For sake of simplicity, we used the same notation as therein. We find the smallest z satisfying (4.90) such that $\mathcal{P}(\alpha, z) < 1$, for $\alpha \in (0, \alpha_{\rm mx})$, and $\alpha_{\rm mx} \in (0, 1)$ to be determined[recall that $\mathcal{P}(\alpha, z)$ is defined in (4.96)].

Using exactly the same argument as Theorem 4.2.2 we have the following two conditions on z:

$$z \geq \sigma(\alpha) + \frac{(\theta \cdot \alpha) \cdot \left(\left(1 - \frac{\alpha}{2}\right)\widetilde{\mu}_{\rm mn} + \frac{D^{\ell}_{\rm mn}}{2}\alpha - \frac{1}{2}\epsilon_{opt}\right)}{\frac{D^2_{\rm mx}}{\mu} + \left(1 - \frac{\alpha}{2}\right)\widetilde{\mu}_{\rm mn} + \frac{D^{\ell}_{\rm mn}}{2}\alpha - \frac{1}{2}\epsilon_{opt}} \tag{4.143}$$

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

$$\begin{aligned} &G^{\star}_P(\alpha) \cdot \theta^{-1} \cdot G_X(z) \cdot C_1 \cdot 8\phi_{ub} L^2_{\rm mx} \cdot \rho^2_B \cdot \alpha^2 \\ &+ \left(G^{\star}_P(\alpha) \cdot \theta^{-1} \cdot 2\phi_{ub} \cdot C_1 + C_2\right) \cdot G_Y(z) \cdot 2m\phi^{-2}_{lb} L^2_{\rm mx} \cdot \rho^2_B \cdot \alpha^2 \\ &+ \left(G^{\star}_P(\alpha) \cdot \theta^{-1} \cdot 2\phi_{ub} \cdot C_1 + C_2\right) \cdot G_Y(z) \cdot 8m\phi^{-2}_{lb} L^2_{\rm mx} \cdot G_X(z) \cdot \rho^4_B \cdot \alpha^2 < 1. \end{aligned} \tag{4.144}$$

Using the fact that $G_P^\star(\alpha)$ is monotonically increasing on $\alpha \in (0, 2\widetilde{\mu}_{\mathrm{mn}}/(\widetilde{\mu}_{\mathrm{mn}} - D_{\mathrm{mn}}^\ell))$, and restricting $\alpha \in (0, \widetilde{\mu}_{\mathrm{mn}}/(\widetilde{\mu}_{\mathrm{mn}} - D_{\mathrm{mn}}^\ell)]$, a sufficient condition for (4.144) is

$$\alpha \leq \alpha(z) \triangleq \left(A_{1,\theta} \frac{1}{z - \rho_B} + A_{2,\theta} \frac{1}{z - \rho_B} + A_{3,\theta} \frac{1}{(z - \rho_B)^2} \right)^{-1/2}, \tag{4.145}$$

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

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

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

$$z \geq \rho_B + A_\theta \alpha, \quad \text{with} \quad A_\theta \triangleq \sqrt{A_{1,\theta} + A_{2,\theta} + A_{3,\theta}}. \tag{4.146}$$

Letting $\epsilon_{opt} = \epsilon_{opt}^\star$ in (4.143), the condition reduces to

$$z \geq 1 - \frac{\widetilde{\mu}_{\mathrm{mn}} - \frac{\alpha}{2}(\widetilde{\mu}_{\mathrm{mn}} - D_{\mathrm{mn}}^\ell)}{\frac{2D_{\mathrm{mx}}^2}{\mu} + \widetilde{\mu}_{\mathrm{mn}} - \frac{\alpha}{2}(\widetilde{\mu}_{\mathrm{mn}} - D_{\mathrm{mn}}^\ell)} \cdot (1 - \theta)\alpha. \tag{4.147}$$

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

$$\bar{z} = \inf_{\theta \in (0,1)} \max \left\{ \rho_B + A_\theta \alpha, 1 - \frac{\widetilde{\mu}_{\mathrm{mn}} - \frac{\alpha}{2}(\widetilde{\mu}_{\mathrm{mn}} - D_{\mathrm{mn}}^\ell)}{\frac{2D_{\mathrm{mx}}^2}{\mu} + \widetilde{\mu}_{\mathrm{mn}} - \frac{\alpha}{2}(\widetilde{\mu}_{\mathrm{mn}} - D_{\mathrm{mn}}^\ell)} \cdot (1 - \theta)\alpha \right\}, \tag{4.148}$$

with A_θ defined in (4.146).

Finally, we further simplify (4.148). Letting $\theta = 1/2$ and using $\alpha \in (0, \widetilde{\mu}_{\mathrm{mn}}/(\widetilde{\mu}_{\mathrm{mn}} - D_{\mathrm{mn}}^\ell)]$, the second term in the max of (4.148) can be upper bounded by

$$1 - \underbrace{\frac{\widetilde{\mu}_{\mathrm{mn}}\mu}{4D_{\mathrm{mx}}^2 + \widetilde{\mu}_{\mathrm{mn}}\mu} \cdot \frac{1}{2}}_{\triangleq J} \alpha. \tag{4.149}$$

The condition $\bar{z} < 1$ imposes the following upper bound on α: $\alpha < \alpha_{\rm mx} = \min\{(1-\rho_B)/A_{\frac{1}{2}}, \widetilde{\mu}_{\rm mn}/(\widetilde{\mu}_{\rm mn} - D^{\ell}_{\rm mn}), 1\}$. Eq. (4.148) then simplifies to (4.142) with $\alpha^* = (1-\rho_B)/(A_{\frac{1}{2}} + J)$ that equates $1 - J\alpha$ and $\rho_B + A_{\frac{1}{2}}\alpha$. □

4.6.9 Rate estimate using linearization surrogate (4.63) (time-varying directed network case)

Corollary 4.6.1 (Linearization surrogates)**.** *In the setting of Theorem 4.6.1, let $\{x^\nu\}$ be the sequence generated by SONATA (Algorithm 4), using the surrogates (4.63) and step-size $\alpha = c \cdot \alpha_{\rm mx}$, $c \in (0,1)$, where $\alpha_{\rm mx} = \min\{1, (1-\rho_B)^2/(C_M \cdot \kappa_g (1+\beta/L)^2)\}$ and C_M is a constant defined in (4.155). The number of iterations (communications) needed for $U(x_{\rm i}^\nu) - U^\star \leq \epsilon$, ${\rm i} \in [m]$, is*

$$\mathcal{O}\left(\kappa_g \log(1/\epsilon)\right), \qquad \text{if} \quad \frac{\rho_B}{(1-\rho_B)^2} \leq \frac{1}{C_M \cdot \kappa_g \left(1+\frac{\beta}{L}\right)^2}, \tag{4.150}$$

$$\mathcal{O}\left(\frac{\left(\kappa_g + \beta/\mu\right)^2 \rho_B}{(1-\rho_B)^2} \log(1/\epsilon)\right), \qquad \text{otherwise.} \tag{4.151}$$

Proof. According to Theorem 4.6.1, the rate z can be bounded as

$$z \leq \max\{z_1, z_2\}, \quad \text{with} \quad z_1 \triangleq 1 - \alpha \cdot J \ \text{ and } \ z_2 \triangleq \rho_B + A_{\frac{1}{2}}\alpha, \tag{4.152}$$

where J and $A_{\frac{1}{2}}$ are defined in (4.149) and (4.146), respectively.

The proof consists in bounding properly z_1 and z_2 based upon the surrogate (4.63) postulated in the corollary. We begin particularizing the expressions of J and $A_{\frac{1}{2}}$. Since $\nabla^2 \widetilde{f}_{\rm i}(x_{\rm i}; x_{\rm i}^\nu) = L$, one can set $\widetilde{\mu}_{\rm mn} = L$, and (4.16) holds with $D^{\ell}_{\rm mn} = 0$ and $D_{\rm mx} = L - \mu$. Furthermore, by Assumption 4.1.1, it follows that $\beta \geq \lambda_{\max}(\nabla^2 f_{\rm i}(x)) - L$, for all $x \in \mathcal{K}$; 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 particularize first the quantities $G_P^\star\left(\frac{\widetilde{\mu}_{\mathrm{mn}}}{\widetilde{\mu}_{\mathrm{mn}}-D_{\mathrm{mn}}^\ell}\right)$ [cf. (4.138)], C_1 and C_2 [cf. (4.94)]:

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

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

$$
J=\frac{1}{2}\frac{\kappa_g}{4(\kappa_g-1)^2+\kappa_g}\in\left[\frac{1}{8\kappa_g},\frac{1}{2}\right],
\tag{4.153}
$$

and

$$
\begin{aligned}
&(A_{\frac{1}{2}})^2\\
=&G_P^\star(1)\cdot 2\cdot C_1\cdot 8\phi_{ub}\cdot L_{\mathrm{mx}}^2\cdot\frac{2c_0^2\rho_B^2}{1-\rho_B}\\
&+\left(G_P^\star(1)\cdot 2\cdot C_1\cdot 2\phi_{ub}+C_2\right)\cdot 2m\phi_{lb}^{-2}L_{\mathrm{mx}}^2\cdot\frac{2c_0^2\rho_B^2}{1-\rho_B}\\
&+\left(G_P^\star(1)\cdot 2\cdot C_1\cdot 2\phi_{ub}+C_2\right)\cdot 8m\phi_{lb}^{-2}L_{\mathrm{mx}}^2\cdot\frac{4c_0^4\rho_B^4}{(1-\rho_B)^2}\\
\le&\left(G_P^\star(1)\cdot 2\cdot C_1\cdot 12\phi_{ub}+C_2\right)\cdot 8m\phi_{lb}^{-2}L_{\mathrm{mx}}^2\cdot\frac{4c_0^4\rho_B^2}{(1-\rho_B)^2}\\
\le&\left[\frac{4(L-\mu)^2+L^2}{\mu L^2}\cdot\frac{12}{\mu\phi_{lb}L^2}\left((2L-\mu)^2+4(L+\beta)^2\right)\cdot 12\phi_{ub}+\frac{4}{L^2}\right]\\
&\qquad\qquad\cdot 8m\phi_{lb}^{-2}\left(L+\beta\right)^2\cdot\frac{4c_0^4\rho_B^2}{(1-\rho_B)^2}\\
\le&C_M^2\cdot\kappa_g^2\left(1+\frac{\beta}{L}\right)^4\cdot\frac{\rho_B^2}{(1-\rho_B)^2},
\end{aligned}
\tag{4.154}
$$

where

$$
C_M\triangleq 608\cdot\phi_{lb}^{-1}\cdot c_0\sqrt{\frac{\phi_{ub}}{\phi_{lb}}\cdot m},
\tag{4.155}
$$

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

By (4.154), we have

$$z_2 \leq \bar{z}_2 \triangleq \rho_B + \alpha M \cdot \frac{\rho_B}{1-\rho_B}, \quad \text{with} \quad M \triangleq C_M \cdot \kappa_g (1+\beta/L)^2. \tag{4.156}$$

Since $\alpha \in (0,1]$ must be chosen so that $z \in (0,1]$, we impose $\max\{z_1, \bar{z}_2\} < 1$, implying $\alpha \leq \min\{J^{-1}, (1-\rho_B)^2/(M\rho_B), 1\}$. Since $J^{-1} > 1$ [cf. (4.153)], the condition on α reduces to $\alpha \leq \alpha_{\text{mx}} \triangleq \min\{1, (1-\rho_B)^2/(M\rho_B)\} < 1$. Choose $\alpha = c \cdot \alpha_{\text{mx}}$, for some given $c \in (0,1)$. Depending on the value of ρ_B, either $\alpha_{\text{mx}} = 1$ or $\alpha_{\text{mx}} = (1-\rho_B)^2/(M\rho_B)$.

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

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

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

Therefore, z can be bounded as

$$\begin{aligned} z \leq \max\{z_1, \bar{z}_2\} &\leq 1 - c \cdot (1-c)\left(1 - \frac{1}{C_M}\right) \cdot J \\ &\leq 1 - c \cdot (1-c)\left(1 - \frac{1}{C_M}\right) \cdot \frac{1}{8\kappa_g}. \end{aligned} \tag{4.157}$$

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

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

Now we can bound z. Since $Jc/(M\rho_B) < 1$ (by the same reasoning as in proof of Proposition 4.2.2),

$$\begin{aligned} z &\leq \max\{z_1, \bar{z}_2\} \leq 1 - \frac{Jc}{M\rho_B} \cdot (1-c)\,(1-\rho_B)^2 \\ &\overset{(4.153)}{\leq} 1 - \frac{c\,(1-c)}{8C_M} \cdot \frac{(1-\rho_B)^2}{\kappa_g^2(1+\beta/L)^2\rho_B} \\ &= 1 - \frac{c\,(1-c)}{8C_M} \cdot \frac{(1-\rho_B)^2}{(\kappa_g+\beta/\mu)^2\rho_B}. \end{aligned} \tag{4.158}$$

□

4.6.10 Rate estimate using local f_{i} (4.64) (time-varying directed network case)

Corollary 4.6.2 (Local f_{i}, $\beta \leq \mu$)**.** *Instate assumptions of Theorem 4.6.1 and suppose $\beta \leq \mu$. Consider SONATA (Algorithm 4) using the surrogates (4.64) and step-size $\alpha = c{\cdot}\alpha_{\mathrm{mx}}$, $c \in (0,1)$, with $\alpha_{\mathrm{mx}} = \min\{1, (1-\rho_B)^2/(\tilde{M}_2\rho_B)\}$ where $\tilde{M}_2 = 1087\tilde{C}_M\left(1+\frac{\beta}{\mu}\right)^2\left(\kappa_g+\frac{\beta}{\mu}\right)^2$ and the constant $\tilde{C}_M$ is defined in (4.165). The number of iterations (communications) needed for $U(x_{\mathrm{i}}^\nu) - U^\star \leq \epsilon$, $\mathrm{i} \in [m]$, is*

$$\mathcal{O}\left(1 \cdot \log(1/\epsilon)\right), \qquad \text{if} \quad \frac{\rho_B}{(1-\rho_B)^2} \leq \frac{1}{1087 \cdot \tilde{C}_M \cdot \left(1+\frac{\beta}{\mu}\right)^2\left(\kappa_g+\frac{\beta}{\mu}\right)^2}, \tag{4.159}$$

$$\mathcal{O}\left(\frac{\kappa_g^2\rho_B}{(1-\rho_B)^2}\log(1/\epsilon)\right), \qquad \text{otherwise.} \tag{4.160}$$

Corollary 4.6.3 (Local f_{i}, $\beta > \mu$)**.** *Instate assumptions of Theorem 4.6.1 and suppose $\beta > \mu$. Consider SONATA (Algorithm 4) using the surrogates (4.64) and step-size $\alpha = c{\cdot}\alpha_{\mathrm{mx}}$, $c \in (0,1)$, where $\alpha_{\mathrm{mx}} = \min\{1, (1-\rho_B)^2/(\tilde{M}_1\rho_B)\}$ with $\tilde{M}_1 = 1428\tilde{C}_M\left(1+\frac{L}{\beta}\right)\left(\kappa_g+\frac{\beta}{\mu}\right)$ and the constant $\tilde{C}_M$ is defined in (4.165). The number of iterations (communications) needed for $U(x_{\mathrm{i}}^\nu) - U^\star \leq \epsilon$, $\mathrm{i} \in [m]$, is*

$$\mathcal{O}\left(\frac{\beta}{\mu}\log(1/\epsilon)\right), \qquad \text{if} \quad \frac{\rho_B}{(1-\rho_B)^2} \leq \frac{1}{1428 \cdot \tilde{C}_M \cdot \left(1+\frac{L}{\beta}\right)\left(\kappa_g+\frac{\beta}{\mu}\right)}, \tag{4.161}$$

$$\mathcal{O}\left(\frac{\left(\kappa_g+(\beta/\mu)\right)^2\rho_B}{(1-\rho_B)^2}\log(1/\epsilon)\right), \qquad \text{otherwise.} \tag{4.162}$$

Proof. In the setting of the corollary, we have: $\nabla^2 \widetilde{f}_{\mathrm{i}}(x;y) = \nabla^2 f_{\mathrm{i}}(x) + \beta I$, for all $y \in \mathcal{K}$; $\nabla^2 f_{\mathrm{i}}(x) \succeq 0$, for all $x \in \mathcal{K}$; and, by Assumption 4.1.1, $0 \preceq \nabla^2 \widetilde{f}_{\mathrm{i}}(x,y) - \nabla^2 F(x) \preceq 2\beta I$, for all $x, y \in \mathcal{K}$. Therefore, we can set $D^{\ell}_{\mathrm{mn}} = 0$, $D_{\mathrm{mx}} = 2\beta$, $\widetilde{\mu}_{\mathrm{mn}} = \beta + (\mu - \beta)_+ = \max\{\beta, \mu\}$, and $L_{\mathrm{mx}} = L + \beta$.

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

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

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

$$
J = \frac{1}{2}\frac{1}{1 + 16\left(\frac{\beta}{\mu}\right)\cdot \min\left\{1, \frac{\beta}{\mu}\right\}}, \tag{4.163}
$$

and

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

where

$$\tilde{M} \triangleq \begin{cases} 1428 \cdot \tilde{C}_M \left(1 + \frac{L}{\beta}\right)\left(\kappa_g + \frac{\beta}{\mu}\right), & \beta > \mu, \\ 1087 \cdot \tilde{C}_M \left(1 + \frac{\beta}{\mu}\right)^2 \left(\kappa_g + \frac{\beta}{\mu}\right)^2, & \beta \leq \mu, \end{cases} \tag{4.164}$$

and

$$\tilde{C}_M \triangleq c_0^2 \phi_{lb}^{-1} \sqrt{\frac{\phi_{ub}}{\phi_{lb}} \cdot m}, \tag{4.165}$$

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

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

$$z \leq \max\{z_1, \bar{z}_2\}, \quad \text{with} \quad z_1 \triangleq 1 - \alpha \cdot J \ \text{ and } \ \bar{z}_2 \triangleq \rho_B + \alpha\, \tilde{M} \cdot \frac{\rho_B}{1 - \rho_B}, \tag{4.166}$$

where J and $\tilde{M}$ are now given by (4.163) and (4.164), respectively. For $\max\{z_1, z_2\} < 1$, we require $\alpha \leq \alpha_{\text{mx}} \triangleq \min\{1, (1-\rho_B)^2/(\tilde{M}\rho_B)\}$, and choose $\alpha = c \cdot \alpha_{\text{mx}}$, with arbitrary $c \in (0, 1)$.

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

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

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

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

when $\beta > \mu$, and

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

when $\beta \leq \mu$.

- **Case II:** $\alpha_{\text{mx}} = (1-\rho_B)^2/(\tilde{M}\rho_B)$. This corresponds to $\tilde{M}\rho_B > (1-\rho_B)^2$. We have $\alpha = c \cdot \alpha_{\text{mx}}$. Similarly to inequality (4.158) in proof of Corollary 4.6.1, we have

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

which yields

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

when $\beta > \mu$, and

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

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

□

5. DECENTRALIZED SECOND-ORDER ALGORITHMS FOR (STRONGLY) CONVEX OPTIMIZATION OVER NETWORKS

In this chapter, we study the class of Empirical Risk Minimization (ERM) problems over a network of m agents, modeled as undirected graph. Differently from master/slave systems, no centralized node is assumed in the network (which will be referred to as *meshed* network). Each agent i has access to n i.i.d. samples $z_{\mathrm{i}}^{(1)}, \ldots, z_{\mathrm{i}}^{(n)}$ drawn from an unknown, common distribution on $\mathcal{Z} \subseteq \mathbb{R}^p$; the associated empirical risk reads

$$f_{\mathrm{i}}(x) \triangleq \frac{1}{n} \sum_{\mathrm{j}=1}^{n} \ell\left(x; z_{\mathrm{i}}^{(\mathrm{j})}\right), \tag{5.1}$$

where $\ell : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}$ is the loss function, assumed to be (strongly) convex in x, for any given $z \in \mathcal{Z}$. Agents aim to minimize the total empirical risk over the $N = mn$ samples, resulting in the following ERM over networks:

$$\widehat{x} \in \operatorname*{argmin}_{x \in \mathcal{K}} F(x) \triangleq \frac{1}{m} \sum_{\mathrm{i}=1}^{m} f_{\mathrm{i}}(x), \tag{5.2}$$

where $\mathcal{K} \subseteq \mathbb{R}^d$ is convex and known to the agents.

Since the functions f_{i} can be accessed only locally and routing local data to other agents is infeasible or highly inefficient, solving (5.2) calls for the design of distributed algorithms that alternate between a local computation procedure at each agent's side, and a round of communication among neighboring nodes. The cost of communications is often considered the bottleneck for distributed computing, if compared with local (possibly parallel) computations (e.g., [193], [194]). Therefore, our goal is developing *communication-efficient* distributed algorithms that solve (5.2) within the *statistical* precision.

The provably faster convergence rates of second order methods over gradient-based algorithms make them potential candidates for communication saving (at the cost of more computations). Despite the success of Newton-like methods to solve ERM in a centralized setting (e.g., [195], [196]), including master/slave architectures [140], [141], [197]–[199], their distributed counterparts on meshed networks are not on par: convergence rates provably

faster than those of first order methods are achieved at high communication costs [200], [201], cf. Sec. 5.0.2.

We claim that stronger guarantees of second order methods over meshed networks can be certified if a *statistically-informed* design/analysis is put forth, in contrast with statistically agnostic approaches that look at (5.2) as deterministic optimization and target any arbitrarily small suboptimality. To do so, we build on the following two key insights.

• **Fact 1 (statistical accuracy):** When it comes to learning problems, the ERM (5.2) is a surrogate of the population minimization

$$x^\star \in \operatorname*{argmin}_{x\in\mathcal{K}} F_P(x) \triangleq \mathbb{E}_{Z\sim\mathbb{P}}\, \ell(x;Z)\,. \tag{5.3}$$

The ultimate goal is to estimate $x^\star$ via the ERM (5.2). Denoting by $x_\varepsilon \in \mathcal{K}$ the estimate returned by the algorithm, we have the risk decomposition (neglecting the approximation error due to the use of a specific set of models $x \in \mathcal{K}$):

$$\begin{aligned} &F_P(x_\varepsilon) - F_P(x^\star) \\ &= \underbrace{\Big\{F_P(x_\varepsilon) - F(x_\varepsilon)\Big\}}_{\leq \text{statistical error}} + \Big\{F(x_\varepsilon) - F(x^\star)\Big\} + \underbrace{\Big\{F(x^\star) - F_P(x^\star)\Big\}}_{\leq \text{statistical error}} \\ &\leq \mathcal{O}(\text{statistical error}) + \underbrace{\Big\{F(x_\varepsilon) - F(\widehat{x})\Big\}}_{=\text{optimization error}} \end{aligned} \tag{5.4}$$

where the statistical error is usually of the order $\mathcal{O}(1/\sqrt{N})$ or $\mathcal{O}(1/N)$ (cf. Sec. 5.1). It is thus sufficient to reach an optimization accuracy $F(x_\varepsilon) - F(\widehat{x}) = \mathcal{O}(\text{statistical error})$. This can potentially save communications.

• **Fact 2 (statistical similarity):** Under mild assumptions on the loss functions and i.i.d samples across the agents (e.g., [141], [202]), it holds with high probability (and uniformly on $\mathcal{Z}$)

$$\left\|\nabla^2 f_{\mathrm{i}}(x) - \nabla^2 F(x)\right\| \leq \beta = \tilde{\mathcal{O}}(1/\sqrt{n}), \quad \forall x \in \mathcal{K}, \tag{5.5}$$

with $\tilde{\mathcal{O}}$ hiding log-factors and the dependence on d. In words, the local empirical losses f_{i} are statistically similar to each other and the average F, especially for large n.

The key insight of Fact 1 is that one can target suboptimal solutions of (5.2) within the statistical error. This is different from seeking a distributed optimization method that achieves any arbitrarily small empirical suboptimality. Fact 2 suggests a further reduction in the communication complexity via *statistical preconditioning*, that is, subsampling the Hessian of F over the local data sets, so that no Hessian matrix has to be transmitted over the network. In thsi chapter, we show that, if synergically combined, these two facts can improve the communication complexity of distributed second order methods over meshed networks.

5.0.1 Major contributions

We propose and analyze a decentralization of the cubic regularization of the Newton method [134] over meshed networks. The algorithm employs a local computation procedure performed in parallel by the agents coupled with a round of (perturbed) consensus mechanisms that aim to track *locally* the gradient of F (a.k.a. gradient-tracking) as well as enforce an agreement on the local optimization directions. The optimization procedure is an inexact, preconditioned (cubic regularized) Newton step whereby the gradient of F is estimated by gradient tracking while the Hessian of F is subsampled over the local data sets. Neither a line-search nor communication of Hessian matrices over the network are performed.

We established for the first time *global* convergence for different classes of ERM problems, as summarized in Table 5.1. Our results are of two types: i) classical complexity analysis (number of communication rounds) for arbitrary solution accuracy (right panel); ii) and complexity bounds for statistically optimal solutions (left panel, V_N is the statistical error). Postponing to Sec 5.3 a detailed discussion of these results, here we highlight some key novelties of our findings. **Convex ERM:** For convex F, if arbitrary ε-solutions are targeted, the algorithm exhibits a two-speed behavior: 1) a first rate of the order of $\mathcal{O}((1/\sqrt{1-\rho}) \cdot \sqrt{LD^3/\varepsilon^{1+\alpha}})$, as long as $\varepsilon = \Omega(LD^3\beta^2)$; up to the network dependent factor $1/\sqrt{1-\rho}$, this (almost) matches the rate of the centralized Newton method [134]; and 2) the slower rate $\mathcal{O}((1/\sqrt{1-\rho}) \cdot (LD^3\beta^2)/\varepsilon)$, which is due to the local subsampling of the global Hessian $\nabla^2 F$; this term is dominant for smaller values of ε. The interesting fact is

that $\varepsilon = \Omega(LD^3\beta^2)$ is of the order of the statistical error V_N. Therefore, rates of the order of centralized ones are provably achieved up to statistical precision (left panel). **Strongly Convex ERM ($\beta < \mu$):** The communication complexity shows a three-phase rate behaviour (right panel); for arbitrarily small $\varepsilon > 0$, the worst-case communication complexity is linear, of the order of $\widetilde{\mathcal{O}}\Big((1/\sqrt{1-\rho}) \cdot (\beta/\mu) \cdot \log(1/\varepsilon)\Big)$. Faster rates are certified when $\varepsilon = \mathcal{O}(V_N)$ (left panel). Note that the region of superlinear convergence is a false improvement when the first term $m^{1/4}\sqrt{LD/\mu}$ is dominant, e.g., F is ill-conditioned and n is not large. This term is unavoidable [134]–unless more refined function classes are considered, such as self-concordant or quadratic ($L = 0$). The left panel shows improved rates in the latter case or under an initialization within a $\mathcal{O}(1/\sqrt{n})$-neighborhood of the solution. **Strongly Convex ERM ($\beta \geq \mu$):** This is a common setting when F_P is convex and a regularizer is used in the ERM (5.2) for learnability/stability purposes; typically, $\mu = \mathcal{O}(1/\sqrt{N})$, $\beta = \mathcal{O}(1/\sqrt{n})$. We proved linear rate for arbitrary ε-values. Differently from the majority of first-order methods over meshed networks (cf. Sec. 5.0.2), this rate does not depend on the condition number of F but on the generally more favorable ratio β/μ. Furthermore, when $\varepsilon = \mathcal{O}(V_N)$, the rate does not improve over the convex case.

In summary, we propose a second-order method solving convex and strongly convex problems over meshed networks that, for the first time, enjoys global complexity bounds and communication complexity close to oracle complexity of centralized methods up to the statistical precision.

Table 5.1. Communication complexity of DiRegINA to $\varepsilon > 0$ suboptimality for (strongly) convex ERM. **Right column:** arbitrary ε values. **Left column:** $\varepsilon = \Omega(V_N)$, V_N is the statistical error [cf. (5.4)]. The other parameters are: μ and L are the strong convexity constant of F and Lipschitz constant of $\nabla^2 F$, respectively; D and D_p are estimates of the optimality gap at the initial point; β measures the similarity of $\nabla^2 f_{\mathrm{i}}$ [cf. (5.5)]; ρ characterizes the connectivity of the network; and $\alpha > 0$ is an arbitrarily small constant.

Problem		$\varepsilon = \Omega(V_N)$ **(statistical error)**	$\varepsilon > 0$ **(arbitrary)**
Convex $\mu = 0$ $V_N = \mathcal{O}(1/\sqrt{N})$	Thm. 5.3.1 Cor. 5.3.1	$\widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{1-\rho}} \cdot \sqrt{\frac{LD^3}{V_N^{1+\alpha}}}\right)$	$\widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{1-\rho}} \cdot \left\{\sqrt{\frac{LD^3}{\varepsilon^{1+\alpha}}} + \frac{LD^3\beta}{\varepsilon^{1+\alpha/2}}\right\}\right)$
Strongly-convex $0 < \beta < \mu$ $V_N = \mathcal{O}(1/N)$ $\mu = \mathcal{O}(1)$	$L > 0$ Thm. 5.3.2 Cor.5.3.2	$\widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{1-\rho}}\left\{m^{1/4}\sqrt{\frac{LD}{\mu}} + \log\log\left(\frac{\mu^3}{mL^2V_N}\right)\right\}\right)$	$\widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{1-\rho}}\left\{m^{1/4} \cdot \sqrt{\frac{LD}{\mu}} + \log\log\left(\frac{\mu^2}{\beta^2}\min\left\{1, \frac{\beta^2\mu}{mL^2} \cdot \frac{1}{\varepsilon}\right\}\right) + \frac{\beta}{\mu}\log\left(\frac{\beta^2\mu}{mL^2\varepsilon}\right)\right\}\right)$
	$L > 0$ +initialization (Cor. 5.3.3)	$\widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{1-\rho}} \cdot \left\{\log\log\left(\frac{\mu^3}{mL^2} \cdot \frac{1}{V_N}\right)\right\}\right), \quad \beta = \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$	$\widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{1-\rho}}\left\{\log\log\left(\frac{\mu^2}{\beta^2} \cdot \min\left(1, \frac{\beta^2\mu}{mL^2\varepsilon}\right)\right) + \frac{\beta}{\mu}\log\left(\frac{\beta^2\mu}{mL^2\varepsilon}\right)\right\}\right)$
	$L = 0$ (Thm. 5.6.5, appendix 5.6.5.4)	$\widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{1-\rho}} \cdot \log\log\left(\frac{D_p}{V_N}\right)\right), \quad \beta = \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$	$\widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{1-\rho}} \cdot \left\{\log\log\left(\frac{D_p}{\varepsilon}\right) + \frac{\beta}{\mu}\log\left(\frac{D_p\beta^2}{\mu^2\varepsilon}\right)\right\}\right)$
Strongly-convex (regularized) $0 < \mu \le \beta$ $V_N = \mathcal{O}(1/\sqrt{N})$	$L > 0$ (Thm. 5.3.3)	$\widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{1-\rho}} m^{1/2}\sqrt{\frac{LD}{V_N}}\right), \quad \begin{cases}\mu = \mathcal{O}(V_N)\\ \beta = \mathcal{O}(\frac{1}{\sqrt{n}})\end{cases}$	$\widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{1-\rho}}\left\{\sqrt{\frac{LD}{\mu}}\left(1 + m^{1/4}\sqrt{\frac{\beta}{\mu}}\right) + \frac{\beta}{\mu}\log\left(\frac{\beta^2\mu}{mL^2\varepsilon}\right)\right\}\right)$
	$L = 0$ (Thm. 5.6.6, appendix 5.6.7)	$\widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{1-\rho}} \cdot m^{1/2} \cdot \log\left(\frac{1}{V_N}\right)\right), \quad \begin{cases}\mu = \mathcal{O}(V_N)\\ \beta = \mathcal{O}(\frac{1}{\sqrt{n}})\end{cases}$	$\widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{1-\rho}} \cdot \frac{\beta}{\mu} \cdot \log\left(\frac{1}{\varepsilon}\right)\right)$

5.0.2 Related Works

The literature of distributed optimization is vast; here we review relevant methods applicable to *meshed networks*, omitting the less relevant work considering only master-slave systems (a.k.a star networks).

• **Statistically oblivious methods:** Despite being vast and providing different communication and oracle complexity bounds, the literature (e.g., [139], [146], [203]–[208]) on decentralized **first-order methods** for minimizing Q-Lipschitz-smooth and μ-strongly convex global objective F mostly focuses on the particular case where $n = 1$ in (5.1) and $\mathcal{K} = \mathbb{R}^d$, and does not take into account statistical similarity of the risks. The best convergence results for nonaccelerated first-order methods certify linear rate, scaling with the condition number $\kappa = Q/\mu$ (Q is the Lipschitz constant of ∇F); Nesterov-based acceleration improves the dependence to $\sqrt{\kappa}$ [209]. This performance can still be unsatisfactory when $1 + \beta/\mu < \kappa$ (resp. $1 + \beta/\mu < \sqrt{\kappa}$). This is the typical situation of ill-conditioned problems, such as many learning problems where the regularization parameter that is optimal for test predictive performance is very small [202]. For instance, consider the ridge-regression problem with optimal regularization parameter $\mu = 1/\sqrt{mn}$ (Table 1 in [141]), we have: $\kappa = \mathcal{O}(\sqrt{m \cdot n})$ while $\beta/\mu = \mathcal{O}(\sqrt{m})$. Notice that the former grows with the local sample size n, while the latter is independent.

A number of **second-order methods** were proposed for distributed optimization over meshed networks, with typical results being local superlinear convergence [210]–[212] or global linear convergence no better than that of first-order methods [213]–[217]. Improved upon first-order methods global bounds are achieved by exploiting expensive sending local Hessians over the network–such as [201], obtaining communication complexity bound $\mathcal{O}((mL\|\nabla f(x_0)\|/\mu^2)+\log\log(1/\varepsilon))$–or employing double-loop schemes [200] wherein at each iteration, a distributed first-order method is called to find the Newton direction, obtaining iteration complexity $\mathcal{O}(\sqrt[3]{LD^3/\varepsilon})$ at the price of excessive communications per iteration. Furthermore, these schemes cannot handle constraints. To the best of our knowledge, no distributed second-order method over meshed networks has been proved to globally converge with communication complexity bounds even up to a network dependent factor close

to the standard [134] bounds $\mathcal{O}(\sqrt{(LD^3)/\varepsilon})$ for convex and $\mathcal{O}(\sqrt{LD/\mu}+\log\log(\mu^3/L^2\varepsilon))$ for μ-strongly convex problems. Table 5.1 shows the first results of this genre.

• **Methods exploiting statistical similarity:** Starting the works [139], [140] several papers studied the idea of statistical preconditioning to decrease the communication complexity over star networks, for different problem classes; example include [140], [218], [219] (quadratic losses), [141] (self-concordant losses), [220] (under $n > d$), and [142] (composite optimization), with [202], [221] employing acceleration. None of these methods are implementable over meshed networks, because they rely on a centralized (master) node. To our knowledge, Network-DANE [183] and SONATA [222] are the only two methods that leverage statistical similarity to enhance convergence of distributed methods over meshed networks; [183] studies strongly convex quadratic losses while [222] considers general objectives, achieving a communication complexity of $\widetilde{\mathcal{O}}((1/\sqrt{1-\rho})\cdot\beta/\mu\cdot\log(1/\varepsilon))$. Both schemes call at every iteration for an exact solution of local strongly convex problems while our subproblems are based on second-order approximations, computationally thus less demanding. Nevertheless, our algorithm retains same rate dependence on β/μ. Our study covers also non-strongly convex losses.

5.1 Setup and Background

5.1.1 Problem setting

We study convex and strongly convex instances of the ERM (5.2); specifically, we make the following assumptions [note that, although explicitly omitted, each $f_{\mathrm{i}}(x)$ and thus F depend on the sample $z \in \mathcal{Z}$ via $\ell(x,z)$; all the assumptions below are meant to hold uniformly on $\mathcal{Z}$].

Assumption 5.1.1 (convex ERM)**.** *The following hold:*

(i) $\emptyset \neq \mathcal{K} \subseteq \mathbb{R}^d$ *is closed and convex;*

(ii) Each $f_{\mathrm{i}} : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}$ *is twice differentiable and* μ_{i}*- strongly convex on (an open set containing)* $\mathcal{K}$*, with* $\mu_{\mathrm{i}} \geq 0$*;*

(iii) Each ∇f_{i} is Q_{i}-Lipschitz continuous on $\mathcal{K}$, where ∇f_{i} is the gradient with respect to x; let $Q_{\max} \triangleq \max_{\mathrm{i}=1,\ldots,m} Q_{\mathrm{i}}$;

(iv) F has bounded level sets.

Assumption 5.1.2 (strongly convex ERM)**.** *Assumption 5.1.1(i)-(iii) holds and F is μ-strongly convex on $\mathcal{K}$, with $\mu > 0$.*

The following condition is standard when studying second order methods.

Assumption 5.1.3. *$\nabla^2 F : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ is L-Lipschitz continuous on $\mathcal{K}$, i.e., $\left\|\nabla^2 F(x) - \nabla^2 F(y)\right\| \leq L\left\|x - y\right\|$, for some $L > 0$ and all $x, y \in \mathcal{K}$.*

Statistical accuracy: As anticipated earlier, we are interested in computing estimates of the population minimizer (5.3) up to the statistical error using the ERM rule (5.4). To do so, throughout the chapter, we postulate the following standard uniform convergence property, which suffices for learnability by (5.4): there exists a constant V_N, dependent on $N = m\,n$, such that

$$\sup_{x\in\mathcal{K}} |F(x) - F_P(x)| \leq V_N \quad \text{w.h.p.} \tag{5.6}$$

The statistical accuracy V_N has been widely studied in the literature, e.g., [223]–[227]. Consistently with these works, we will assume:

1. $V_N = \mathcal{O}(1/N)$, for μ-strongly convex F and F_P, with $0 < \mu = \mathcal{O}(1)$;
2. $V_N = \mathcal{O}(1/\sqrt{N})$ for convex or μ-strongly convex F, with $\mu = \mathcal{O}(1/\sqrt{N})$.

These cases cover a variety of problems of practical interest. An example of case 1 is a loss in the form $\ell(x; z) = f(x; z) + (\mu/2)\|x\|^2$, with fixed regularization parameter and f convex in x (uniformly on z), as in ERM of linear predictors for supervised learning [228]. Case 2 captures traditional low-dimensional ($n > d$) ERM with convex losses or regularized losses as above with optimal regularization parameter $\mu = \mathcal{O}(1/\sqrt{N})$ [184], [225], [226].

Under (5.6), the suboptimality gap at given $x \in \mathcal{K}$ reads:[1]

$$F_P(x) - F_P(x^\star) \leq \mathcal{O}(V_N) + \Big\{F(x) - F(\widehat{x})\Big\}, \quad \text{w.h.p.} \tag{5.7}$$

[1]↑We point out that our results hold under (5.7), which can also be established using weaker conditions than (5.6), e.g., invoking stability arguments [229].

Therefore, our ultimate goal will be computing ε-solutions x_ε of (5.2) of the order $\varepsilon = \mathcal{O}(V_N)$.

Statistical similarity: Towards above goal, we can exploit statistical similarity which naturally exists in big-data; see Chapter 4, Sec. 4.1.1.2 for more details. Thus we are interested in studying problem (5.2) under statistical similarity of f_i's. We recall Definition 4.1.1 from Chapter 4:

Assumption 5.1.4 (β-related f_i's)**.** *The local functions f_i's are β-related: $\|\nabla^2 F(x) - \nabla^2 f_\mathrm{i}(x)\|_2 \leq \beta$, for all $x \in \mathcal{K}$ and some $\beta \geq 0$.*

The interesting case is when $1 + \beta/\mu \ll \kappa \triangleq Q/\mu$, where Q is the Lipschitz constant of ∇F on $\mathcal{K}$ (uniformly on $\mathcal{Z}$). Under standard assumptions on data distributions and learning model underlying the ERM-see, e.g., [141], [202]–β is of the order $\beta = \mathcal{O}\left(1/\sqrt{n}\right)$, with high probability. In our analysis, when we target convergence to the statistical error, we will tacitly assume such dependence of β on the local sample size. Note that our bounds hold for general situations when Assumption 5.1.4 may hold due to some other reason besides statistical arguments.

5.1.2 Network setting

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

Assumption 5.1.5 (On the network)**.** *The graph $\mathcal{G}$ is connected.*

5.2 Algorithmic Design: DiRegINA

We aim at decentralizing the cubic regularization of the Newton method [134] over undirected graphs. The main challenge in developing such an algorithm is to track and adapt a faithful estimates of the global gradient and Hessian matrix of F at each agent, without incurring in an unaffordable communication overhead while still guaranteeing convergence at fast rates. Our idea is to estimate locally the gradient ∇F via gradient-tracking [177], [230]

while the Hessian $\nabla^2 F$ is replaced by the local subsampled estimates $\nabla^2 f_{\mathrm{i}}$ (statistical preconditioning). The algorithm, termed DiRegINA (Distributed Regularized Inexact Newton Algorithm), is formally introduced in Algorithm 5, and commented next.

Each agent maintains and updates iteratively a local copy $x_{\mathrm{i}} \in \mathbb{R}^d$ of the global optimization variable x along with the auxiliary variable $s_{\mathrm{i}} \in \mathbb{R}^d$, which estimates the gradient of the global objective F; x_{i}^{ν} (resp. s_{i}^{ν}) denotes the value of x_{i} (resp. s_{i}) at iteration $\nu \geq 0$. (S.1) is the optimization step wherein every agent i, given x_{i}^{ν} and s_{i}^{ν}, minimizes an inexact local second-order approximation of F, as defined in (5.8a). In this surrogate function, i) y_{i}^{ν} acts as an approximation of ∇F at x_{i}^{ν}, that is, $s_{\mathrm{i}}^{\nu} \approx \nabla F(x_{\mathrm{i}}^{\nu})$; ii) in the quadratic term, $\nabla^2 f_{\mathrm{i}}(x_{\mathrm{i}}^{\nu})$ plays the role of $\nabla^2 F(x_{\mathrm{i}}^{\nu})$ (due to statistical similarity, cf. Assumption 5.1.4) with $\tau_{\mathrm{i}} I$ ensuring strong convexity of the objective; and iii) the last term is the cubic regularization as in the centralized method [134]. In (S.2), based upon exchange of the two vectors $x_{\mathrm{i}}^{\nu+}$ and s_{i}^{ν} with their immediate neighbors, each agent updates the estimate $x_{\mathrm{i}}^{\nu} \to x_{\mathrm{i}}^{\nu+1}$ via the consensus step (5.8b) and $s_{\mathrm{i}}^{\nu} \to s_{\mathrm{i}}^{\nu+1}$ via the perturbed consensus (5.8c), which in fact tracks $\nabla F(x_{\mathrm{i}}^{\nu})$ [177], [230]. The weights $(W_K)_{\mathrm{i,j}=1}^{m}$ in (5.8b)-(5.8c) are free design quantities and subject to the following conditions, where $\mathcal{P}_K$ denotes the set of polynomials with degree less than or equal than $K = 1, 2, \ldots$.

Assumption 5.2.1 (On the weight matrix W_K)**.** *The matrix $W_K = P^K(\overline{W})$, where $P_K \in \mathcal{P}_K$ with $P_K(1) = 1$, and $\overline{W} \triangleq \left(\bar{w}_{\mathrm{ij}}\right)_{\mathrm{i,j}=1}^{m}$ is a reference matrix satisfying the following conditions:*

(a) *$\overline{W}$ has a sparsity pattern compliant with $\mathcal{G}$, that is*

i) $\bar{w}_{\mathrm{ii}} > 0$, for all $\mathrm{i} = 1, \ldots, m$;

ii) $\bar{w}_{\mathrm{ij}} > 0$, if $(\mathrm{i}, \mathrm{j}) \in \mathcal{E}$; and $\bar{w}_{\mathrm{ij}} = 0$ otherwise.

(b) *$\overline{W}$ is doubly stochastic, i.e., $1^\top \overline{W} = 1^\top$ and $\overline{W} 1 = 1$.*

Let $\rho_K \triangleq \lambda_{\max}(W_K - 11^\top/m)$ [$\lambda_{\max}(\bullet)$ denotes the largest eigenvalue of the matrix argument]

When $K = 1$, $W_K = \overline{W}$, that is, a single round of communication per iteration is performed. Several rules have been proposed in the literature for $\overline{W}$ to be compliant with

Assumption 5.2.1, such as the Laplacian, the Metropolis-Hasting, and the maximum-degree weights rules; see, e.g., [231] and references therein. When $K > 1$, K rounds of communications per iteration ν are employed. For instance, this can be performed using the same reference matrix $\overline{W}$ (satisfying Assumption 5.2.1) in each communication exchange, resulting in $W_K = \overline{W}^K$ and $\rho_K = \rho^K$, with $\rho = \lambda_{\max}(\overline{W} - 11^\top/m) < 1$. Faster information mixing can be obtained using suitably designed polynomials $P_K(\overline{W})$, such as Chebyshev [163], [205] or orthogonal (a.k.a. Jacobi) [232] polynomials (notice that $P_K(1) = 1$ is to ensure the doubly stochasticity of W_K when $\overline{W}$ is doubly stochastic).

Although the minimization (5.8a) may look challenging, it is showed in [134] that its computational complexity is of the same order as of finding the standard Newton step. Importantly, in our algorithm, these are local steps made without any communications between the nodes.

Algorithm 5: Distribute Regularized Inexact Newton Algorithm (DiRegINA)

Data: $x_{\mathrm{i}}^0 \in \mathcal{K}$ and $y_{\mathrm{i}}^0 = \nabla f_{\mathrm{i}}(x_{\mathrm{i}}^0)$, $\tau_{\mathrm{i}} > 0$, $M_{\mathrm{i}} > 0$, $\forall \mathrm{i}$.
Iterate: $\nu = 1, 2, \ldots$
`[S.1] [Local Optimization]` Each agent i computes $x_{\mathrm{i}}^{\nu+}$:

$$\begin{aligned} x_{\mathrm{i}}^{\nu+} = &\operatorname*{argmin}_{y \in \mathcal{K}} F(x_{\mathrm{i}}^\nu) + \langle y_{\mathrm{i}}^\nu, y - x_{\mathrm{i}}^\nu \rangle \\ &+ \frac{1}{2} \left\langle \left[\nabla^2 f_{\mathrm{i}}(x_{\mathrm{i}}^\nu) + \tau_{\mathrm{i}} I \right] (y - x_{\mathrm{i}}^\nu), y - x_{\mathrm{i}}^\nu \right\rangle + \frac{M_{\mathrm{i}}}{6} \left\| y - x_{\mathrm{i}}^\nu \right\|^3. \end{aligned} \tag{5.8a}$$

`[S.2] [Local Communication]` Each agent i updates its local variables according to

$$x_{\mathrm{i}}^{\nu+1} = \sum_{\mathrm{j}=1}^{m} (W_K)_{\mathrm{i,j}}\, x_{\mathrm{j}}^{\nu+}, \tag{5.8b}$$

$$y_{\mathrm{i}}^{\nu+1} = \sum_{\mathrm{j}=1}^{m} (W_K)_{\mathrm{i,j}} \left(y_{\mathrm{j}}^\nu + \nabla f_{\mathrm{j}}(x_{\mathrm{j}}^{\nu+1}) - \nabla f_{\mathrm{j}}(x_{\mathrm{j}}^\nu) \right). \tag{5.8c}$$

end

On the initialization: We will study convergence of Algorithm 5 under two sets of initialization for the x-variables, namely: i) random initialization and ii) statistically informed initialization. The latter is given by

$$x_{\mathrm{i}}^0 = \sum_{\mathrm{j}=1}^m (W_K)_{\mathrm{i},\mathrm{j}} x_{\mathrm{j}}^{-1}, \quad \text{with} \quad x_{\mathrm{i}}^{-1} = \operatorname*{argmin}_{x \in \mathcal{K}} f_{\mathrm{i}}(x). \tag{5.9}$$

This corresponds to a preliminary round of consensus on the local solutions x_{i}^{-1}. This second strategy takes advantage of the statistical similarity of f_{i}'s to guarantee, under (5.6), an initial optimality gap of the order of: $p^0 \triangleq \frac{1}{m}\sum_{\mathrm{i}=1}^m (F(x_{\mathrm{i}}^0) - F(\widehat{x})) = \mathcal{O}(1/\sqrt{n})$. If we further assume $\mu_{\mathrm{i}} > 0$, for all i, one can show that $p^0 = \mathcal{O}(1/n)$. This will be shown to significantly improve the convergence rate of the algorithm, at a negligible extra communication cost (but local computations).

5.3 Convergence Analysis

In this section, we study convergence of DiRegINA applied to convex (cf. Sec. 5.3.1) and strongly convex ERM (5.2), the latter with either $\beta < \mu$ (cf. Sec. 5.3.2) or $\beta \geq \mu > 0$ (cf. Sec. 5.3.3). Our complexity results are of two type: i) classical rate bounds targeting any arbitrary ERM suboptimality $\varepsilon > 0$; and ii) convergence rates to V_N-solutions of (5.2) (statistical error). Our complexity bounds are established in terms of the suboptimality gap:

$$p^\nu \triangleq \frac{1}{m}\sum_{\mathrm{i}=1}^m \left(F(x_{\mathrm{i}}^\nu) - F(\widehat{x})\right), \tag{5.10}$$

where $\{x_{\mathrm{i}}^\nu\}_{\mathrm{i}=1}^m$ is the iterate generated by DiRegINA at iteration ν (iterations are counted as number of optimization steps `(S.1)`). Similarly to the centralized case [134], our bounds also depend on the following distance of initial points x_{i}^0, $\mathrm{i} = 1, \ldots, m$, from a given optimum $\widehat{x}$ of (5.2)

$$D \triangleq \max_{x_{\mathrm{i}} \in \mathcal{K}, \forall \mathrm{i}} \left\{ \max_{\mathrm{i}=1\ldots,m} ||x_{\mathrm{i}} - \widehat{x}|| : \sum_{\mathrm{i}=1}^m F(x_{\mathrm{i}}) \leq \sum_{\mathrm{i}=1}^m F(x_{\mathrm{i}}^0) \right\}.$$

Note that $D < \infty$ (cf. Assumption 5.1.1).

For the sake of simplicity, in the rate bounds we hide universal constants and log factors independent on ε via $\widetilde{\mathcal{O}}$-notation; the exact expressions can be found in the supplementary material along with a detailed characterization of all the rate regions travelled by the algorithm.

5.3.1 Convex ERM (5.2)

Our first result pertains to convex F (and F_P).

Theorem 5.3.1. *Consider the ERM (5.2) under Assumptions 5.1.1, 5.1.3, and 5.1.4 over a graph $\mathcal{G}$ satisfying Assumption 5.1.5; and let $\{x_{\mathrm{i}}^{\nu}\}_{\mathrm{i}=1}^{m}$ be the sequence generated by DiRegINA under the following tuning: $M_{\mathrm{i}} = L > 0$ and $\tau_{\mathrm{i}} = 2\beta$, for all $\mathrm{i} = 1, \dots, m$; $W_K = P_K(\overline{W})$ (and $P_K(1) = 1$), where $\overline{W}$ is a given matrix satisfying Assumption 5.2.1 with $\rho = \lambda_{\max}(\overline{W} - 11^{\top}/m)$, and $K = \widetilde{\mathcal{O}}(\log(1/\varepsilon)/\sqrt{1-\rho})$, with $\varepsilon > 0$ being the target accuracy. Then, the total number of communications for DiRegINA to make $p^{\nu} \leq \varepsilon$ reads*

$$\widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{1-\rho}} \cdot \left\{\sqrt{\frac{LD^3}{\varepsilon^{1+\alpha}}} + \frac{LD^3\beta}{\varepsilon^{1+\alpha/2}}\right\}\right), \tag{5.11}$$

where $\alpha > 0$ is arbitrarily small. In particular, if the $\mathcal{G}$ is a star or fully-connected, $\rho = 0$ and $\alpha = 0$.

Proof. See Appendix 5.6.4 in the supplementary material. □

The rate expression (5.11) has an interesting interpretation. The multiplicative factor $1/\sqrt{1-\rho} > 1$ accounts for the rounds of communications per iteration (optimization steps) while the other two terms quantify the overall number of iterations to reach the desired accuracy ε. Note that the first of these two terms, $\mathcal{O}(\sqrt{LD^3/\varepsilon^{1+\alpha}})$, is "almost" identical to the rate of the centralized Newton method (with a slight difference definition of D; see [134]) while the other one, $\mathcal{O}((LD^3\beta)/\varepsilon^{1+\alpha/2})$, is a byproduct of the discrepancy between local and global Hessian matrices. This shows a two-speed behavior of the algorithm, depending on the target accuracy $\varepsilon > 0$: 1) as long as $\varepsilon = \Omega(LD^3\beta^2)$, $\mathcal{O}((LD^3\beta^2)/\varepsilon)$ can be neglected and the algorithm exhibits almost centralized fast convergence (up to the network effect),

$\mathcal{O}(\frac{1}{\sqrt{1-\rho}}\sqrt{LD^3/\varepsilon^{1+\alpha}})$; 2) on the other hand, for smaller (order of) ε, the rate is determined by the worst-term $\mathcal{O}(\frac{1}{\sqrt{1-\rho}}(LD^3\beta^2)/\varepsilon)$.

The interesting observation is that, in the setting above and under (5.6), (5.7) holds with $V_N = \mathcal{O}(1/\sqrt{N})$ and $\beta = \mathcal{O}(1/\sqrt{n})$. Hence, $\varepsilon = \Omega(LD^3\beta^2)$ is of the order of the statistical error V_N, as long as $m \leq n$, which is a reasonable condition. This together with Theorem 5.3.1 implies that fast rates (of the order of centralized ones) can be certified up to the statistical precision, as formalized next.

Corollary 5.3.1 (V_N-solution)**.** *Instate the setting of Theorem 5.3.1, and let* $V_N = \mathcal{O}(1/\sqrt{N})$, $\beta = \mathcal{O}(1/\sqrt{n})$, *and* $m \leq n$. *Then DiRegINA returns a* V_N*-solution of* (5.2) *in*

$$\widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{1-\rho}} \cdot \sqrt{\frac{LD^3}{V_N^{1+\alpha}}}\right) \tag{5.12}$$

communications.

5.3.2 Strongly-convex ERM (5.2) with $\beta < \mu$

We consider now the case of F μ-strongly convex and $\beta < \mu$. The complementary case $\beta \geq \mu$ is studied in Sec. 5.3.3.

Theorem 5.3.2. *Instate the setting of Theorem 5.3.1 with Assumption 5.1.1 replaced by Assumption 5.1.2 and* $K = \widetilde{\mathcal{O}}(1/\sqrt{1-\rho})$*; and further assume* $\beta < \mu$. *Then, the total number of communications for DiRegINA to make* $p^\nu \leq \varepsilon$ *reads*

$$\widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{1-\rho}}\left\{m^{\frac{1}{4}}\sqrt{\frac{LD}{\mu}} + \log\log\left[\frac{\mu^2}{\beta^2}\cdot\min\left(1, \frac{\beta^2\mu}{mL^2}\cdot\frac{1}{\varepsilon}\right)\right] + \frac{\beta}{\mu}\log\left[\max\left(1, \frac{\beta^2\mu}{mL^2}\cdot\frac{1}{\varepsilon}\right)\right]\right\}\right). \tag{5.13}$$

Proof. See Appendix 5.6.5 in the supplementary material. □

DiRegINA exhibits a different rate behavior, depending on the value of ϵ. We notice three "regions": 1) a first phase of the order of $\widetilde{\mathcal{O}}\left(m^{1/4}\sqrt{LD/\mu}\right)$ number of iterations; 2) the second region is of quadratic convergence, with rate of the order of $\log\log(1/\varepsilon)$; and

finally 3) the region of linear convergence with rate $\widetilde{\mathcal{O}}\left(\beta/\mu \log(1/\varepsilon)\right)$. This last region is not present in the rate of the centralized cubic regularization of the Newton method and is due to the Hessians discrepancy. Clearly, for arbitrarily small $\varepsilon > 0$, (5.13) is dominated by the last term, resulting in a linear convergence. This linear rate is slightly worse than that of SONATA [222] in sight of first two terms in (5.13). This is because DiRegINA is an inexact (and thus more computationally efficient) method than [222]. We remark that more favorable complexity estimates can be obtained when $L = 0$ (i.e., f_{i}'s are quadratic)–we refer the reader to the supplementary material for details.

The algorithm does not enter in the last region if $\varepsilon = \Omega(\beta^2\mu/(mL^2))$. This means that faster rate can be guaranteed up to V_N-solutions, as stated next.

Corollary 5.3.2 (V_N-solution)**.** *Instate the setting of Theorem 5.3.2, and let $V_N = \mathcal{O}(1/N)$, $\beta = \mathcal{O}(1/\sqrt{n})$, $\mu = \mathcal{O}(1)$, and $m \leq n$. DiRegINA returns a V_N-solution of (5.2) in*

$$\widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{1-\rho}}\left\{m^{1/4}\sqrt{\frac{LD}{\mu}} + \log\log\left(\frac{\mu^3}{mL^2V_N}\right)\right\}\right) \tag{5.14}$$

communications.

When the problem is ill-conditioned (i.e. $\mu \ll 1$) the first term $m^{1/4}\sqrt{LD/\mu}$ may dominate the $\log\log$ term in (5.14), unless n is extremely large (and thus V_N very small). This term is unavoidable–it is present also in the centralized instances of Newton-type methods–unless more refined function classes are considered, such as (generalized) self-concordant [233]–[235]. In the supplementary material, we present results for quadratic losses (cf. Appendix 5.6.5.4). Here, we take another direction and show that the initialization strategy (5.9) is enough to get rid of the first phase.

Corollary 5.3.3 (V_N-solution + initialization)**.** *Instate the setting of Theorem 5.3.2 and further assume: $\mu_{\mathrm{i}} = \Omega(1)$, for all $\mathrm{i} = 1, \dots m$, and $n = \Omega(L^2/\mu^3 \cdot m)$. DiRegINA , initialized with (5.9), returns a V_N-solution of (5.2) in*

$$\widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{1-\rho}}\left\{\log\log\left(\frac{\mu^3}{mL^2}\cdot\frac{1}{V_N}\right)\right\}\right) \tag{5.15}$$

communications.

Proof. See Appendix 5.6.5.5 in the supporting material. □

5.3.3 Strongly-convex ERM (5.2) with $\beta \geq \mu$

We now consider the complementary case $\beta \geq \mu$. This is a common setting when F_P is convex and a regularizer is used in the ERM (5.2), making F μ-strongly convex; typically, $\mu = \mathcal{O}(1/\sqrt{N})$ while $\beta = \mathcal{O}(1/\sqrt{n})$.

Theorem 5.3.3. *Instate the setting of Theorem 5.3.2 with now $\mu \leq \beta \leq 1$. Then, the total number of communications for DiRegINA to make $p^\nu \leq \varepsilon$ reads*

$$\widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{1-\rho}}\left\{\sqrt{\frac{LD}{\mu}}\left(1+m^{\frac{1}{4}}\sqrt{\frac{\beta}{\mu}}\right)+\frac{\beta}{\mu}\log\left(\frac{\beta^2\mu}{mL^2}\frac{1}{\varepsilon}\right)\right\}\right). \tag{5.16}$$

Proof. See Appendix 5.6.6 in the supplementary material. □

For arbitrary small $\varepsilon > 0$, the rate (5.16) is dominated by the linear term. When we target V_N-solutions, in this setting $V_N = \mathcal{O}(1/\sqrt{N})$, $\mu = \mathcal{O}(V_N)$ (as for the regularized ERM setting), and $\beta = \mathcal{O}(1/\sqrt{n})$, (5.16) becomes

$$\widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{1-\rho}}\cdot m^{1/2}\cdot\sqrt{\frac{LD}{V_N}}\right). \tag{5.17}$$

Note that this rate is of the same order of the one achieved in the convex setting (with no regularization)–see Corollary 5.3.1. If the functions f_{i} are quadratic, the rate, as expected, improves and reads (see supporting material, Appendix 5.6.7)

$$\widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{1-\rho}}\cdot m^{1/2}\cdot\log\left(\frac{1}{V_N}\right)\right).$$

Note that, on star networks ($\rho = 0$), this rate improves on that of DANE [140].

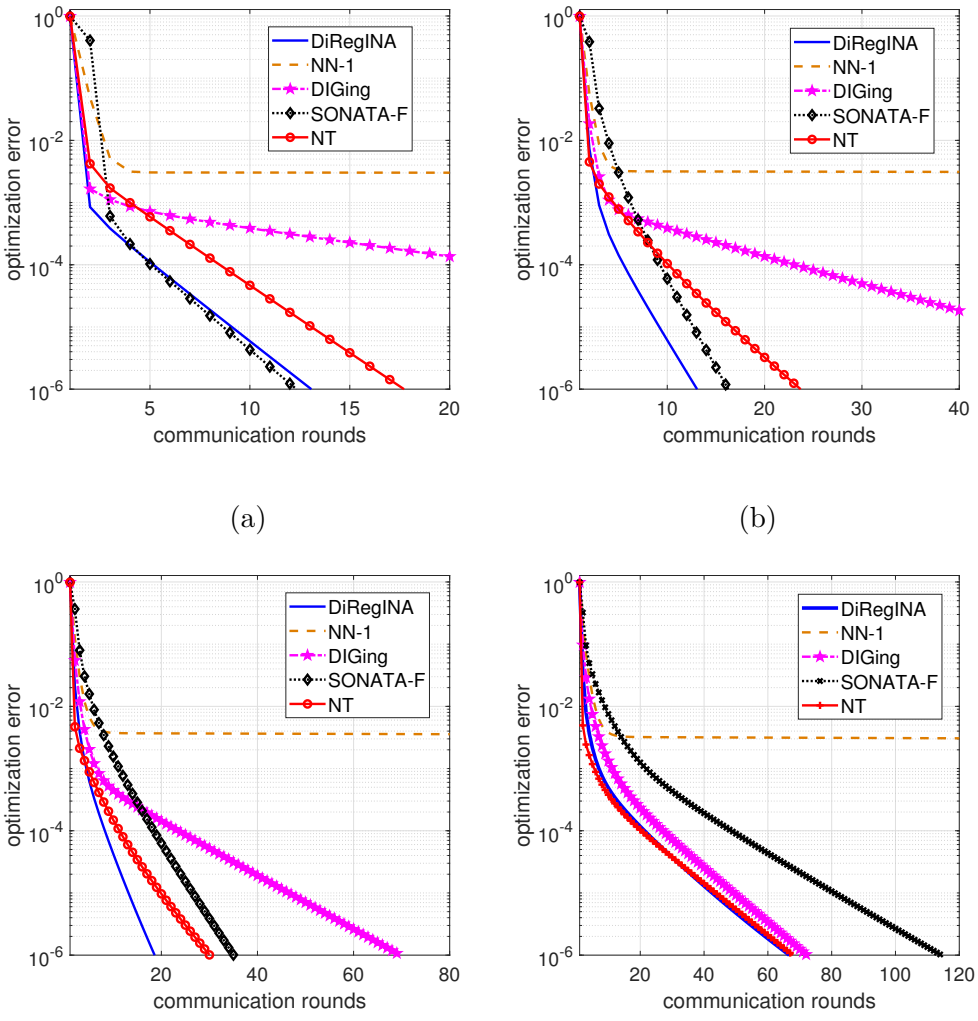


Figure 5.1. Distributed ridge regression: (a) star-topology; and Erdős-Rényi graph with (b) $\rho = 0.20$, (c) $\rho = 0.41$, (d) $\rho = 0.69$.

5.4 Experiments

In this section we test numerically our theoretical findings on two classes of problems over meshed networks: 1) ridge regression and 2) logistic regression. Other experiments can be found in the supplementary material (cf. Sec. 5.6.1).

The network graph is generated using an Erdős-Rényi model $G(m, p)$, with $m = 30$ nodes and different values of p to span different level of connectivity.

We compare DiRegINA with the following methods:

• *Distributed (first-order) method with gradient tracking:* we consider SONATA [222] and DIGing [204]; both build on the idea of gradient tracking, with the former applicable also to constrained problems. For the SONATA algorithm, we will simulate two instances, namely: SONATA-L (L stands for linearization) and SONATA-F (F stands for full); the former uses only first-order information in the agents' local updates (as DGing) while the latter exploits functions' similarity by employing local mirror-descent-based optimization.

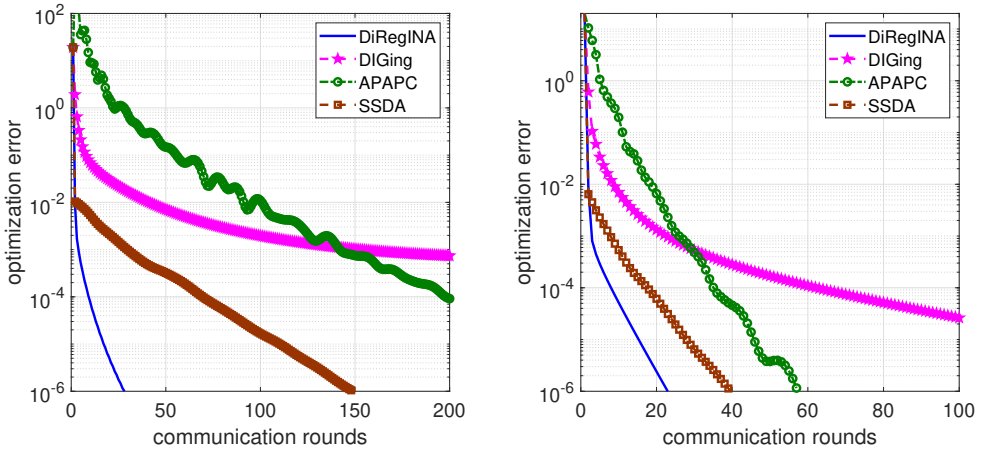
• *Distributed accelerated first-order methods:* we consider APAPC [236] and SSDA [205], which employ Nesterov acceleration on the local optimization steps–with the former using primal gradients while the latter requiring gradients of the conjugate functions–and Chebyshev acceleration on the consensus steps. These schemes do not leverage any similarity among the local agents' functions.

• *Distributed second-order methods:* We implement i) Network Newton-K (NN-K) [237] with $K = 1$ so that it has the same communication cost per iteration of DiRegINA ; ii) SONATA-F [222], which is a mirror descent-type distributed scheme wherein agents need to solve *exactly* a strongly convex optimization problem; and iii) Newton Tracking (NT) [217], which has been shown the outperform the majority of distributed second-order methods.

All the algorithms are coded in MATLAB R2019a, running on a computer with Intel(R) Core(TM) i7-8650U CPU@1.90GHz, 16.0 GB of RAM, and 64-bit Windows 10.

5.4.1 Distributed Ridge Regression

We train ridge regression, LIBSVM, scaled `mg` dataset [238], which is an instance of (5.2) with $f_{\rm i}(x) = (1/2n)\left\|A_{\rm i}x - b_{\rm i}\right\|^2 + \frac{\lambda}{2}\left\|x\right\|^2$ and $\mathcal{K} = \mathbb{R}^d$, with $d = 6$. We set $\lambda = 1/\sqrt{N} = 0.0269$;



(a) (b)

Figure 5.2. Distributed ridge regression. Synthetic data on Erdős-Rényi graph with $\rho = 0.7$: a) $\beta/\mu = 158.1$, $\kappa^{1/2} = 34.55$; b) $\beta/\mu = 11.974$, $\kappa^{1/2} = 11.1$.

we estimate $\beta = 0.1457$ and $\mu = 0.0929$. The graph parameter $p = 0.6, 0.33, 0.28$, resulting in the connectivity values $\rho \approx 0.20, 0.41, 0.70$, respectively. We compared DiRegINA, NN-1, DIGing, SONATA-F and NT, all initialized from the same identical random point. The coefficients of the matrix $\overline{W}$ are chosen according to the Metropolis–Hastings rule [239]. The free parameters of the algorithm are tuned manually; specifically: DiRegINA, $\tau = 2\beta$, $M = 1\mathrm{e}-3$, and $K = 1$; NN-1, $\alpha = 1\mathrm{e}-3$ and $\epsilon = 1$; DIGing, stepsize equal to 0.5; SONATA-F, $\tau = 0.27$; NT, $\epsilon = 0.08$ and $\alpha = 0.1$. This tuning corresponds to the best practical performance we observed.

In Fig. 5.1, we plot the function residual p^ν defined in (5.10) versus the communication rounds in the four aforementioned network settings. DiRegINA demonstrates good performance over first-order methods, and compares favorably also with SONATA-F (which has higher computational cost). Note the change of rate, as predicted by our theory, with linear rate in the last stage. NN-1 is not competitive while NT in some settings is comparable with DiRegINA , but we observed to be more sensitive to the tuning.

The second experiment aims at comparing DiRegINA with the distributed accelerated methods APAPC [236] and SSDA [205] (DIGing is used as benchmark of first-order non-

accelerated schemes). We tested these schemes on two instances of the Ridge regression problem using synthetic data, corresponding to $\beta/\mu \gg \sqrt{\kappa}$ and $\beta/\mu \approx \sqrt{\kappa}$. Recall that SSDA and APAPC converge linearly at a rate proportional to $\sqrt{\kappa}$ while the convergence rate of DiRegINA depends (up to log factors) on β/μ. The problem data are generated as follows: the ground truth $x^* \in \mathbb{R}^d$ is a random vector, $x^* \sim \mathcal{N}(0, I)$, with $d = 40$; samples $b_{\mathrm{i}} \triangleq (b_{\mathrm{i}}^{(\mathrm{j})})_{\mathrm{j}=1}^n$, with $n = 50$, are generated according to the linear model $b_{\mathrm{i}}^{(\mathrm{j})} = a_{\mathrm{i}}^{(\mathrm{j})\top} x^* + \epsilon_{\mathrm{i}}^{(\mathrm{j})}$ where $\epsilon_{\mathrm{i}}^{(\mathrm{j})} \sim \mathcal{N}(0, 1\mathrm{e}-4)$. To obtain controlled values for β, $A_{\mathrm{i}} \triangleq (a_{\mathrm{i}}^{(\mathrm{j})})_{\mathrm{j}=1}^n$ are constructed as follows: we first generate n i.i.d samples $A_1 \triangleq (a_1^{(\mathrm{j})})_{\mathrm{j}=1}^n$, with rows drawn from $\mathcal{N}(0, I)$; then, we set each $A_{\mathrm{i}} = A_1 + E_k$, where E_k in a random matrix with rows drawn from $\mathcal{N}(0, \sigma I)$. The choices of σ are considered resulting in two different values of β, namely: $\sigma = 1/(dn)$ and $\sigma = 7.5/(dn)$, resulting in $\beta = 0.31$ and $\beta = 4.08$, respectively. The values of the condition number read $\kappa = 123.21$ and $\kappa = 1.19\mathrm{e}3$, respectively. The network is simulated as the Erdős-Rényi graph with $p = 0.28$, resulting in $\rho \approx 0.7$; the number of agents is $m = 30$. The tuning of DiRegINA and DIGing is the same as in Fig. 5.1 while APAPC and SSDA are manually tuned for best practical performance.

In Fig. 5.2, we plot the function residual p^ν defined in (5.10) versus the communication rounds; the two panels refer to two different values of $(\beta/\mu, \sqrt{\kappa})$. The figures show that even when β/μ is larger than $\sqrt{\kappa}$, DiRegINA outperforms the accelerated first order methods; roughly, it is from two to five time faster than the best simulated first order method.

5.4.2 Distributed Logistic Regression

We train logistic regression models, regularized by the ℓ_2-ball constraint (with radius 1). The problem is an instance of (5.2), with each $f_{\mathrm{i}}(x) = -(1/n)\sum_{\mathrm{j}=1}^n [\xi_{\mathrm{i}}^{(\mathrm{j})} \ln(z_{\mathrm{i}}^{(\mathrm{j})}) + (1 - \xi_{\mathrm{i}}^{(\mathrm{j})}) \ln(1 - z_{\mathrm{i}}^{(\mathrm{j})})]$, where $z_{\mathrm{i}}^{(\mathrm{j})} \triangleq 1/(1 + \mathrm{e}^{-\langle a_{\mathrm{i}}^{(\mathrm{j})}, x\rangle})$ and binary class labels $\xi_{\mathrm{i}}^{(\mathrm{j})} \in \{0, 1\}$ and vectors $a_{\mathrm{i}}^{(\mathrm{j})}$, $\mathrm{i} = 1, \dots m$ and $\mathrm{j} = 1, \dots, n$ are determined by the data set. We considered the LIBSVM `a4a` ($N = 4,781$, $d = 123$) and synthetic data ($N = 900$, $d = 150$). The latter are generated as follows: a random ground truth $x^* \sim \mathcal{N}(0, I)$, i.i.d. sample $\{a_{\mathrm{i}}^{(\mathrm{j})}\}_{\mathrm{i,j}}$, and $\{\xi_{\mathrm{i}}^{(\mathrm{j})}\}_{\mathrm{i,j}}$ are generated according to the binary model $\xi_{\mathrm{i}}^{(\mathrm{j})} = 1$ if $\langle a_{\mathrm{i}}^{(\mathrm{j})}, x^*\rangle \geq 0$ and $\xi_{\mathrm{i}}^{(\mathrm{j})} = 0$ otherwise. We consider Erdős-Rényi network models with connectivity $\rho = 0.367$ and $\rho = 0.757$.

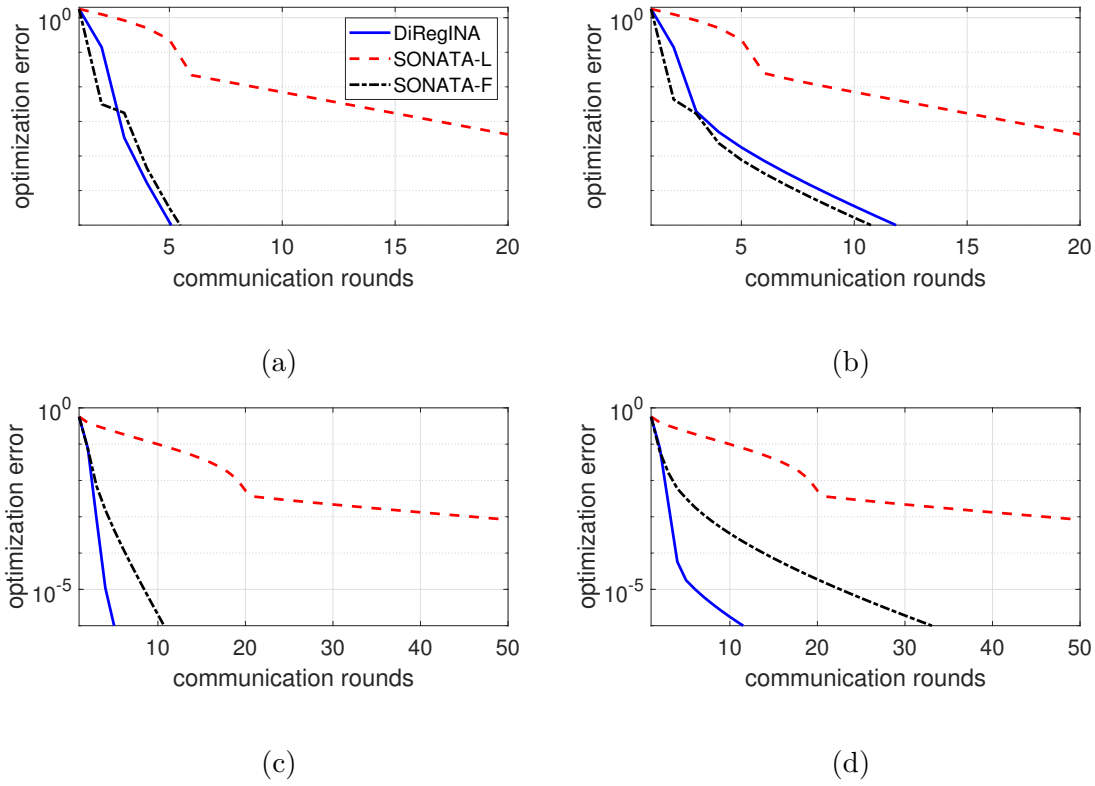


Figure 5.3. Distributed logistic regression: 1) `a4a` dataset on Erdős-Rényi graph with (a) $\rho = 0.367$ (b) $\rho = 0.757$; 2) Synthetic data on Erdős-Rényi graph with (c) $\rho = 0.367$ (d) $\rho = 0.757$.

We compare DiRegINA with SONATA-F and SONATA-L, since they are the only two algorithms in the list that can handle constrained problems. We report results obtained under the following tuning: (i) both SONATA variants, $\alpha = 0.1$; and (ii) DiRegINA , $M = 1$ and $\tau_{\rm i} = 1{\rm e}-3$. The coefficients of the matrix $\overline{W}$ are chosen according to the Metropolis–Hastings rule [239].

In Fig. 5.3, we plot the function residual p^ν defined in (5.10) versus the communication rounds, in the different mentioned network settings. On real data [panels (a)-(b)], DiRegINA and SONATA-F performs equally well, outperforming SONATA-L (first-order method). When tested on the synthetic problem [panel (c)-(d)] with less local samples n and larger dimension d, DiRegINA shows a consistently faster rate, while SONATA-F slows down on less connected networks. Notice also the two-phase rate of DiRegINA , as predicted

by our theory: an initial superlinear rate up to (approximately) the statistical precision, followed by a linear one for high accuracy.

5.5 Conclusions

We proposed the first second-order distributed algorithm for convex and strongly convex problems over meshed networks with *global* communication complexity bounds which, up to the network dependent factor $\widetilde{\mathcal{O}}(1/\sqrt{1-\rho})$, (almost) match the iteration complexity of centralized second-order method [134] in the regime when the desired accuracy is moderate. We showed that this regime is reasonable when one considers ERM problems for which there is no need to optimize beyond the statistical error. Importantly, our method avoids expensive communications of Hessians over the network and keeps the amount of information sent in each communication round similar to first-order methods.

This work is just a starting point towards a theory of second-order methods with performance guarantees on meshed networks under statistical similarity; many questions remain open. An obvious one is incorporating acceleration to improve communication complexity bounds under statistical similarity. A first attempt towards this goal is the follow-up work [135], where an accelerated second-order method exploiting statistical similarity has been analyzed for master/workers architectures. The extension to arbitrary graphs remains an open problem. Second, our main goal here has been decreasing communications, which does not guarantee optimal oracle (computational) complexity–this is because we did not take advantage of the finite-sum structure of the *local* optimization problems. Stochastic optimization algorithms equipped with Variance Reduction (VR) techniques have been proved to be quite effective to obtain cheaper iterations while preserving fast convergence [136], [137]. However, these methods do not exploit any statistical similarity, resulting in less favorable communication complexity whenever $\beta/\mu \ll Q/\mu$. It would be then interesting to investigate whether VR techniques can improve both communication and oracle complexity when statistical similarity is explicitly employed in the algorithmic design.

5.6 Appendix

The appendix is organized as follows. Sec. 5.6.1 provides additional numerical experiments, complementing those in Sec. 5.4. In Sec. 5.6.3, we establish asymptotic convergence of DiRegINA and prove some intermediate results that are instrumental for our rate analysis. Sec. 5.6.4-5.6.7 are devoted to prove the results in Sec. 5.3, namely: Theorem 5.3.1 is proved in Sec. 5.6.4; Theorem 5.3.2 and Corollary 5.3.3 are proved in Sec. 5.6.5; and finally, Theorem 5.3.3 is proved in Sec. 5.6.6.

Furthermore, there are some convergence results stated in Table 5.1 for sake of brevity; they are reported here in the following sections: i) the case of quadratic functions f_i in the setting of Theorem 5.3.2 is stated in Theorem 5.6.5 in Sec. 5.6.5.4 while the case of quadratic f_i's in the setting of Theorem 5.3.3 is stated in Theorem 5.6.6, Sec. 5.6.7.

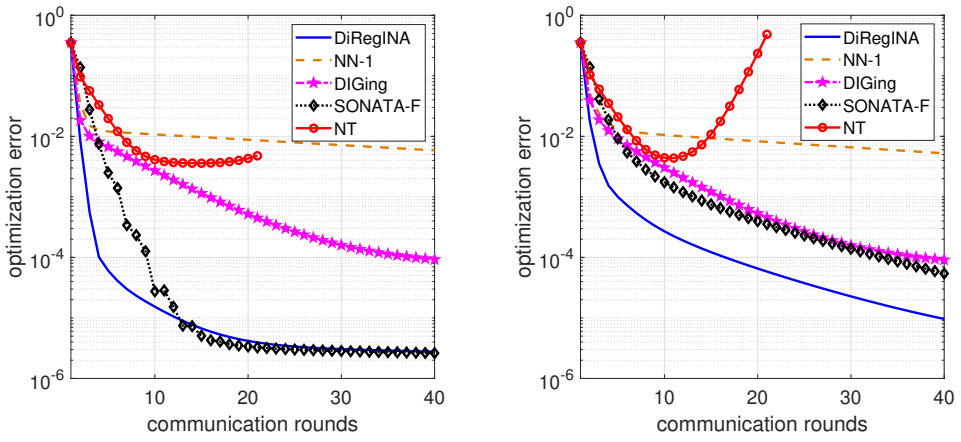


Figure 5.4. Distributed ridge regression on `space-ga` dataset and Erdős-Rényi graph with (a) $\rho = 0.3843$ (b) $\rho = 0.8032$.

5.6.1 Additional Numerical Experiments

5.6.1.1 Distributed ridge regression problem

We consider a (non-strongly) convex instance of the regression problem. Specifically, we have: $f_i(x) = (1/2n)\left\|A_i x - b_i\right\|^2$ and $\mathcal{K} = \mathbb{R}^d$, where A_i and b_i are determined by the scaled

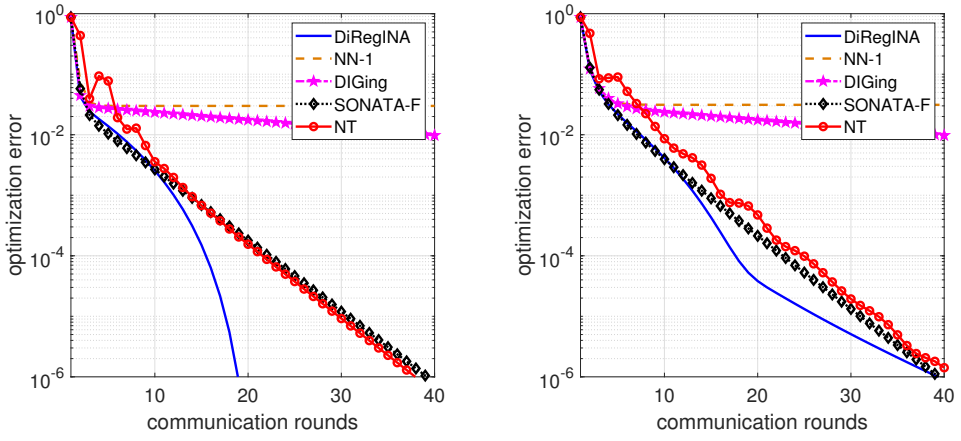
LIBSVM dataset `space-ga` ($N = 3107$, $d = 6$, and $\beta = 0.6353$). The network is simulated as the Erdős-Rényi network model, with $m = 30$ and two connectivity values, $\rho = 0.3843$ and $\rho = 0.8032$. We compared DiRegINA with the algorithms described in Sec. 5.3, namely: NN-1, NT, DIGing and SONATA-F. Note that NN-1 and NT are not guaranteed to converge when applied to convex (non-strongly convex) functions. The tuning of the algorithm is the same as the one described in Sec. 5.4.1. In Fig. 5.4, we plot the optimization error versus the communication rounds achieved by the aforementioned algorithms in the two network settings, $\rho = 0.3843$ and $\rho = 0.8032$. As already observed for the other simulated problems (cf. Sec. 5.4.1), SONATA-F shows similar performance of DiRegINA when running on well-connected networks while its performance deteriorates in poorly connected network. NT seems to be non-convergent while NN1 and DIGing converge, yet slow, to acceptable accuracy.

5.6.1.2 Regularized logistic regression

We train logistic regression models, regularized by an additive ℓ_2-norm (with coefficient $\lambda > 0$). The problem is an instance of (5.2), with each $f_{\mathrm{i}}(x) = -(1/n)\sum_{\mathrm{j}=1}^{n}[\xi_{\mathrm{i}}^{(\mathrm{j})}\ln(z_{\mathrm{i}}^{(\mathrm{j})}) + (1-\xi_{\mathrm{i}}^{(\mathrm{j})})\ln(1-z_{\mathrm{i}}^{(\mathrm{j})})] + (\lambda/2)||x||^2$ and $\mathcal{K} = \mathbb{R}^d$, where $z_{\mathrm{i}}^{(\mathrm{j})} \triangleq 1/(1+\mathrm{e}^{-\langle a_{\mathrm{i}}^{(\mathrm{j})}, x\rangle})$ and binary class labels $\xi_{\mathrm{i}}^{(\mathrm{j})} \in \{0,1\}$ and vectors $a_{\mathrm{i}}^{(\mathrm{j})}$, $\mathrm{i} = 1,\dots m$ and $\mathrm{j} = 1,\dots,n$ are determined by the data set. We considered the LIBSVM `a4a` ($N = 4,781$, $d = 123$) and we set $\lambda = 1/\sqrt{mn}$. The Network is simulated according to the Erdős-Rényi model with $m = 30$ and connectivity $\rho = 0.3372$ and $\rho = 0.7387$.

We compare DiRegINA , NN-1, DIGing, SONATA-F and NT, all initialized from the same random point. The free parameters of the algorithms are tuned manually; the best practical performance are observed with the following tuning: DiRegINA is tuned as described in Sec. 5.4.2, i.e., $\tau = 1$, $M = 1\mathrm{e}-3$, and $K = 1$; NN-1, $\alpha = 1\mathrm{e}-3$ and $\epsilon = 1$; DIGing, stepsize equal to 1; SONATA-F, $\tau = 0.1$; NT, $\epsilon = 0.2$ and $\alpha = 0.05$.

In Fig. 5.5, we plot the optimization error versus the communication rounds achieved by the aforementioned algorithms in two network settings corresponding to $\rho = 0.3372$ and $\rho = 0.7387$. In both settings (panels (a)-(b)), NN-1 and DIGing still exhibits slow



(a) (b)

Figure 5.5. Distributed logistic regression on `a4a` dataset and Erdős-Rényi graph with (a) $\rho = 0.3372$ (b) $\rho = 0.7387$.

convergence, with a slight advantage of DIGing over NN-1. DiRegINA , NT and SONATA-F, perform similarly, with DiRegINA showing some improvements when the network is better connected [panel (a)].

5.6.2 Notations and Preliminary Results

We begin introducing some notation which will be used in all the proofs, along with some preliminary results.

Define

$$\delta_{\mathrm{i}}^{\nu} \triangleq y_{\mathrm{i}}^{\nu} - \nabla F(x_{\mathrm{i}}^{\nu}) \quad \text{and} \quad B_{\mathrm{i}}^{\nu} \triangleq \nabla^2 f_{\mathrm{i}}(x_{\mathrm{i}}^{\nu}) - \nabla^2 F(x_{\mathrm{i}}^{\nu}), \tag{5.18}$$

The local surrogate function $\widetilde{F}_{\mathrm{i}}(y; x_{\mathrm{i}}^{\nu})$ in (5.8a) can be rewritten as

$$\widetilde{F}_{\mathrm{i}}(y; x_{\mathrm{i}}^{\nu}) \triangleq F(x_{\mathrm{i}}^{\nu}) + \langle \nabla F(x_{\mathrm{i}}^{\nu}) + \delta_{\mathrm{i}}^{\nu}, y - x_{\mathrm{i}}^{\nu} \rangle + \frac{1}{2} \left\langle \left[\nabla^2 F(x_{\mathrm{i}}^{\nu}) + B_{\mathrm{i}}^{\nu} + \tau_{\mathrm{i}} I \right] (y - x_{\mathrm{i}}^{\nu}), y - x_{\mathrm{i}}^{\nu} \right\rangle + \frac{M_{\mathrm{i}}}{6} \left\| y - x_{\mathrm{i}}^{\nu} \right\|^3. \tag{5.19}$$

Let us recall the following basic result, which is a consequence of Assumption 5.1.3.

Lemma 5.6.1 (Lemma 1.2.4 in [234]). *Let $F : \mathbb{R}^d \to \mathbb{R}$ be a twice-differentiable function satisfying Assumption 5.1.3. Then, for all $x, y \in \mathbb{R}^d$,*

$$\left|F(y) - F(x) - \langle \nabla F(x), y - x\rangle - \frac{1}{2}\left\langle \nabla^2 F(x)(y-x), y-x\right\rangle\right| \leq \frac{L}{6}\left\|y-x\right\|^3. \tag{5.20}$$

$$\left\|\nabla F(y) - \nabla F(x) - \nabla^2 F(x)(y-x)\right\| \leq \frac{L}{2}\left\|y-x\right\|^2. \tag{5.21}$$

Setting $x = x_{\mathrm{i}}^{\nu}$ in (5.20) implies

$$F(x_{\mathrm{i}}^{\nu}) + \langle \nabla F(x_{\mathrm{i}}^{\nu}), y - x_{\mathrm{i}}^{\nu}\rangle + \frac{1}{2}\left\langle \nabla^2 F(x_{\mathrm{i}}^{\nu})(y-x_{\mathrm{i}}^{\nu}), y-x_{\mathrm{i}}^{\nu}\right\rangle \leq F(y) + \frac{L}{6}\left\|y-x_{\mathrm{i}}^{\nu}\right\|^3, \quad \forall y \in \mathbb{R}^d,$$

which, together with (5.19), gives the following upper bound for the surrogate function $\widetilde{F}_{\mathrm{i}}$ defined in (5.19):

$$\widetilde{F}_{\mathrm{i}}(y; x_{\mathrm{i}}^{\nu}) \leq F(y) + \frac{1}{2}\left\|y-x_{\mathrm{i}}^{\nu}\right\|^2_{(\beta+\tau_{\mathrm{i}})I} + \frac{M_{\mathrm{i}}+L}{6}\left\|y-x_{\mathrm{i}}^{\nu}\right\|^3 + \left\langle \delta_{\mathrm{i}}^{\nu}, y - x_{\mathrm{i}}^{\nu}\right\rangle, \quad \forall y \in \mathbb{R}^d, \tag{5.22}$$

where for a positive semidefinite matrix A, $\|x\|_A^2 \triangleq \langle Ax, x\rangle$. We also denote

$$\Delta x_{\mathrm{i}}^{\nu} \triangleq x_{\mathrm{i}}^{\nu+} - x_{\mathrm{i}}^{\nu}, \quad \delta^{\nu} \triangleq (\delta_{\mathrm{i}}^{\nu})_{\mathrm{i}=1}^m, \quad J \triangleq 11^{\top}/m, \tag{5.23}$$

where we remind that $x_{\mathrm{i}}^{\nu+}$ is obtained by the minimization of the local surrogate function $\widetilde{F}_{\mathrm{i}}(y; x_{\mathrm{i}}^{\nu})$. The rest of the symbols and notations are as defined in the main manuscript.

5.6.3 Asymptotic convergence of DiRegINA

In this section we prove the following theorem stating asymptotic convergence of DiRegINA .

Theorem 5.6.1. *Let Assumptions 5.1.1 and 5.1.3-5.1.4 and 5.1.5 hold, $M_{\mathrm{i}} \geq L$ and $\tau_{\mathrm{i}} = 2\beta$ for all $\mathrm{i} = 1, \ldots, m$. If a reference matrix $\overline{W}$ satisfying Assumption 5.2.1 is used in steps (5.8b)-(5.8c), with $\rho \triangleq \lambda_{\max}(\overline{W} - J) < 1$ and $K = \tilde{\mathcal{O}}(1/\sqrt{1-\rho})$ (explicit condition is provided in eq. (5.42)), then $p^{\nu} \to 0$ and $||x_{\mathrm{i}}^{\nu} - x_{\mathrm{j}}^{\nu}|| \to 0$, as $\nu \to \infty$ for all $\mathrm{i}, \mathrm{j} = 1, \ldots, m$.*

We prove the theorem in three main steps:

Step 1 (Sec. 5.6.3.1): Deriving optimization bounds on the per-iteration decrease of p^ν optimization error;

Step 2 (Sec. 5.6.3.2): Bounding the gradient tracking error δ^ν, which in turn affects the per-iteration decrease of p^ν;

Step 3 (Sec. 5.6.3.3): Constructing a proper Lyapunov function based on the error terms in the previous two steps, whose dynamics imply asymptotic convergence of DiRegINA . To simplify the derivations, we study the case of strongly convex or nonstronlgy convex F together, by setting $\mu = 0$ in the latter case.

5.6.3.1 Optimization error bounds

In this subsection we establish an upper bound for $p^{\nu+1} - p^\nu$ [cf. (5.33)]. We begin with two technical intermediate results–Lemma 5.6.2 and Lemma 5.6.3.

Lemma 5.6.2. *Under Assumption 5.1.1, there holds*

$$\widetilde{F}_{\mathrm{i}}(x_{\mathrm{i}}^{\nu+};x_{\mathrm{i}}^{\nu}) \leq \widetilde{F}_{\mathrm{i}}(x_{\mathrm{i}}^{\nu};x_{\mathrm{i}}^{\nu}) - \frac{M_{\mathrm{i}}}{3}\left\|\Delta x_{\mathrm{i}}^{\nu}\right\|^3 - \frac{\mu_{\mathrm{i}}+\tau_{\mathrm{i}}}{2}\left\|\Delta x_{\mathrm{i}}^{\nu}\right\|^2. \tag{5.24}$$

Proof. By the optimality of $x_{\mathrm{i}}^{\nu+}$ in (5.19), we infer

$$\left\langle y_{\mathrm{i}}^{\nu} + \left[\nabla^2 f_{\mathrm{i}}(x_{\mathrm{i}}^{\nu}) + \tau_{\mathrm{i}} I\right]\Delta x_{\mathrm{i}}^{\nu}, \Delta x_{\mathrm{i}}^{\nu}\right\rangle \leq -\frac{M_{\mathrm{i}}}{2}\left\|\Delta x_{\mathrm{i}}^{\nu}\right\|^3. \tag{5.25}$$

Since $\widetilde{F}_{\mathrm{i}}(x_{\mathrm{i}}^{\nu};x_{\mathrm{i}}^{\nu}) = F(x_{\mathrm{i}}^{\nu})$, we have

$$\begin{aligned}
&\widetilde{F}_{\mathrm{i}}(x_{\mathrm{i}}^{\nu+};x_{\mathrm{i}}^{\nu}) - \widetilde{F}_{\mathrm{i}}(x_{\mathrm{i}}^{\nu};x_{\mathrm{i}}^{\nu}) \\
&\overset{(5.19)}{=} \left\langle y_{\mathrm{i}}^{\nu}, x_{\mathrm{i}}^{\nu+} - x_{\mathrm{i}}^{\nu}\right\rangle + \frac{1}{2}\left\langle \left[\nabla^2 f_{\mathrm{i}}(x_{\mathrm{i}}^{\nu}) + \tau_{\mathrm{i}} I\right]\Delta x_{\mathrm{i}}^{\nu}, \Delta x_{\mathrm{i}}^{\nu}\right\rangle + \frac{M_{\mathrm{i}}}{6}\left\|x_{\mathrm{i}}^{\nu+} - x_{\mathrm{i}}^{\nu}\right\|^3 \\
&\overset{(5.25)}{\leq} -\frac{1}{2}\left\langle \left[\nabla^2 f_{\mathrm{i}}(x_{\mathrm{i}}^{\nu}) + \tau_{\mathrm{i}} I\right]\Delta x_{\mathrm{i}}^{\nu}, \Delta x_{\mathrm{i}}^{\nu}\right\rangle - \frac{M_{\mathrm{i}}}{3}\left\|\Delta x_{\mathrm{i}}^{\nu}\right\|^3 \\
&\leq -\frac{M_{\mathrm{i}}}{3}\left\|x_{\mathrm{i}}^{\nu+} - x_{\mathrm{i}}^{\nu}\right\|^3 - \frac{\mu_{\mathrm{i}}+\tau_{\mathrm{i}}}{2}\left\|x_{\mathrm{i}}^{\nu+} - x_{\mathrm{i}}^{\nu}\right\|^2.
\end{aligned}$$

□

Lemma 5.6.3. *Let Assumptions 5.1.1 and 5.1.3-5.1.4 hold. Then, any arbitrary $\epsilon > 0$, we have*

$$F(x_{\mathrm{i}}^{\nu+}) - \widetilde{F}_{\mathrm{i}}(x_{\mathrm{i}}^{\nu+}; x_{\mathrm{i}}^{\nu}) \leq -\frac{M_{\mathrm{i}} - L}{6}||\Delta x_{\mathrm{i}}^{\nu}||^3 - \frac{\tau_{\mathrm{i}} - \beta - \epsilon}{2}||\Delta x_{\mathrm{i}}^{\nu}||^2 + \frac{1}{2\epsilon}\left\|\delta_{\mathrm{i}}^{\nu}\right\|^2. \tag{5.26}$$

Proof. Taylor's theorem applied to functions $\widetilde{F}_{\mathrm{i}}(\cdot; x_{\mathrm{i}}^{\nu})$ and $F(\cdot)$ around x_{i}^{ν} yields

$$F(x_{\mathrm{i}}^{\nu+}) = F(x_{\mathrm{i}}^{\nu}) + \langle \nabla F(x_{\mathrm{i}}^{\nu}), \Delta x_{\mathrm{i}}^{\nu} \rangle + \Delta x_{\mathrm{i}}^{\nu\top} H_{\mathrm{i}}^{\nu} \Delta x_{\mathrm{i}}^{\nu}, \tag{5.27a}$$

$$\widetilde{F}_{\mathrm{i}}(x_{\mathrm{i}}^{\nu+}; x_{\mathrm{i}}^{\nu}) = \widetilde{F}_{\mathrm{i}}(x_{\mathrm{i}}^{\nu}; x_{\mathrm{i}}^{\nu}) + \left\langle \nabla \widetilde{F}_{\mathrm{i}}(x_{\mathrm{i}}^{\nu}; x_{\mathrm{i}}^{\nu}), \Delta x_{\mathrm{i}}^{\nu} \right\rangle + \Delta x_{\mathrm{i}}^{\nu\top} \widetilde{H}_{\mathrm{i}}^{\nu} \Delta x_{\mathrm{i}}^{\nu}, \tag{5.27b}$$

where

$$H_{\mathrm{i}}^{\nu} = \int_0^1 (1-\theta)\nabla^2 F(\theta x_{\mathrm{i}}^{\nu+} + (1-\theta)x_{\mathrm{i}}^{\nu}) d\theta,$$
$$\widetilde{H}_{\mathrm{i}}^{\nu} = \int_0^1 (1-\theta)\nabla^2 \widetilde{F}_{\mathrm{i}}(\theta x_{\mathrm{i}}^{\nu+} + (1-\theta)x_{\mathrm{i}}^{\nu}; x_{\mathrm{i}}^{\nu}) d\theta.$$

Since $\widetilde{F}_{\mathrm{i}}(x_{\mathrm{i}}^{\nu}; x_{\mathrm{i}}^{\nu}) = F(x_{\mathrm{i}}^{\nu})$ and $\nabla \widetilde{F}_{\mathrm{i}}(x_{\mathrm{i}}^{\nu}; x_{\mathrm{i}}^{\nu}) = \nabla F(x_{\mathrm{i}}^{\nu}) + \delta_{\mathrm{i}}^{\nu}$, subtracting (5.27a)-(5.27b) gives

$$F(x_{\mathrm{i}}^{\nu+}) - \widetilde{F}_{\mathrm{i}}(x_{\mathrm{i}}^{\nu+}; x_{\mathrm{i}}^{\nu}) = \left\langle \left(H_{\mathrm{i}}^{\nu} - \widetilde{H}_{\mathrm{i}}^{\nu}\right)\Delta x_{\mathrm{i}}^{\nu}, \Delta x_{\mathrm{i}}^{\nu} \right\rangle - \langle \delta_{\mathrm{i}}^{\nu}, \Delta x_{\mathrm{i}}^{\nu} \rangle . \tag{5.28}$$

Now let us simplify (5.28). Note that the hessian of $\widetilde{F}_{\mathrm{i}}(\cdot; x_{\mathrm{i}}^{\nu})$ is

$$\nabla^2 \widetilde{F}_{\mathrm{i}}(x_{\mathrm{i}}; x_{\mathrm{i}}^{\nu}) = \nabla^2 F(x_{\mathrm{i}}^{\nu}) + B_{\mathrm{i}}^{\nu} + \tau_{\mathrm{i}} I + M_{\mathrm{i}} G(x_{\mathrm{i}}; x_{\mathrm{i}}^{\nu}), \tag{5.29}$$

where

$$G\left(x_{\mathrm{i}}; x_{\mathrm{i}}^{\nu}\right) \triangleq \frac{1}{2}\left(\left\|x_{\mathrm{i}} - x_{\mathrm{i}}^{\nu}\right\| I + \frac{(x_{\mathrm{i}} - x_{\mathrm{i}}^{\nu})(x_{\mathrm{i}} - x_{\mathrm{i}}^{\nu})^{\top}}{\left\|x_{\mathrm{i}} - x_{\mathrm{i}}^{\nu}\right\|}\right).$$

Hence,

$$
\begin{aligned}
&H_{\mathrm{i}}^{\nu}-\widetilde{H}_{\mathrm{i}}^{\nu}\\
=&\int_0^1(1-\theta)\nabla^2F\left(\theta x_{\mathrm{i}}^{\nu+}+(1-\theta)x_{\mathrm{i}}^{\nu}\right)d\theta-\int_0^1(1-\theta)\nabla^2\widetilde{F}_{\mathrm{i}}\left(\theta x_{\mathrm{i}}^{\nu+}+(1-\theta)x_{\mathrm{i}}^{\nu};x_{\mathrm{i}}^{\nu}\right)d\theta\\
\overset{(5.29)}{=}&\int_0^1(1-\theta)\nabla^2F\left(\theta x_{\mathrm{i}}^{\nu+}+(1-\theta)x_{\mathrm{i}}^{\nu}\right)d\theta-\int_0^1(1-\theta)\left[\nabla^2F(x_{\mathrm{i}}^{\nu})+B_{\mathrm{i}}^{\nu}\right]d\theta-\int_0^1(1-\theta)\tau_{\mathrm{i}}Id\theta\\
&-\int_0^1(1-\theta)M_{\mathrm{i}}\theta G(x_{\mathrm{i}}^{\nu+};x_{\mathrm{i}}^{\nu})d\theta\\
=&\int_0^1(1-\theta)\left(\nabla^2F\left(\theta x_{\mathrm{i}}^{\nu+}+(1-\theta)x_{\mathrm{i}}^{\nu}\right)-\nabla^2F(x_{\mathrm{i}}^{\nu})\right)d\theta\\
&-\int_0^1(1-\theta)B_{\mathrm{i}}^{\nu}d\theta-\int_0^1(1-\theta)\tau_{\mathrm{i}}Id\theta-\int_0^1(1-\theta)M_{\mathrm{i}}\theta G(x_{\mathrm{i}}^{\nu+};x_{\mathrm{i}}^{\nu})d\theta\\
\overset{(a)}{\preceq}&\int_0^1(1-\theta)L\theta||x_{\mathrm{i}}^{\nu+}-x_{\mathrm{i}}^{\nu}||Id\theta\\
&-\int_0^1(1-\theta)B_{\mathrm{i}}^{\nu}d\theta-\int_0^1(1-\theta)\tau_{\mathrm{i}}Id\theta-\int_0^1(1-\theta)M_{\mathrm{i}}\theta G(x_{\mathrm{i}}^{\nu+};x_{\mathrm{i}}^{\nu})d\theta\\
=&-\frac{M_{\mathrm{i}}}{6}G(x_{\mathrm{i}}^{\nu+};x_{\mathrm{i}}^{\nu})+\frac{L}{6}||x_{\mathrm{i}}^{\nu+}-x_{\mathrm{i}}^{\nu}||I-\frac{\tau_{\mathrm{i}}}{2}I-\frac{B_{\mathrm{i}}^{\nu}}{2}
\end{aligned}
\tag{5.30}
$$

where (a) holds since ∇^2F is L-Lipschitz continuous. Combining (5.28) and (5.30), we conclude

$$
\begin{aligned}
F(x_{\mathrm{i}}^{\nu+})-\widetilde{F}_{\mathrm{i}}(x_{\mathrm{i}}^{\nu+};x_{\mathrm{i}}^{\nu})&\leq-\frac{M_{\mathrm{i}}-L}{6}||\Delta x_{\mathrm{i}}^{\nu}||^3-\frac{\tau_{\mathrm{i}}}{2}||\Delta x_{\mathrm{i}}^{\nu}||^2-\frac{1}{2}\left\langle B_{\mathrm{i}}^{\nu}\Delta x_{\mathrm{i}}^{\nu},\Delta x_{\mathrm{i}}^{\nu}\right\rangle-\left\langle\delta_{\mathrm{i}}^{\nu},\Delta x_{\mathrm{i}}^{\nu}\right\rangle\\
&\leq-\frac{M_{\mathrm{i}}-L}{6}||\Delta x_{\mathrm{i}}^{\nu}||^3-\frac{\tau_{\mathrm{i}}-\beta-\epsilon}{2}||\Delta x_{\mathrm{i}}^{\nu}||^2+\frac{1}{2\epsilon}\left\|\delta_{\mathrm{i}}^{\nu}\right\|^2,
\end{aligned}
$$

for arbitrary $\epsilon>0$, where the last inequality is due to the Cauchy-Schwarz inequality and $|\left\langle B_{\mathrm{i}}^{\nu}\Delta x_{\mathrm{i}}^{\nu},\Delta x_{\mathrm{i}}^{\nu}\right\rangle|\leq\beta||\Delta x_{\mathrm{i}}^{\nu}||^2$, which is a consequence of (5.18) and Assumption 5.1.4. □

We are now in a position to prove the main result of this subsection.

Combining (5.24) in Lemma 5.6.3 with (5.26) in Lemma 5.6.2, and using $\widetilde{F}_{\mathrm{i}}(x_{\mathrm{i}}^{\nu};x_{\mathrm{i}}^{\nu})=F(x_{\mathrm{i}}^{\nu})$, yields

$$
F(x_{\mathrm{i}}^{\nu+})-F(x_{\mathrm{i}}^{\nu})\leq-\left(\frac{M_{\mathrm{i}}}{2}-\frac{L}{6}\right)||\Delta x_{\mathrm{i}}^{\nu}||^3-\left(\frac{\mu_{\mathrm{i}}}{2}+\tau_{\mathrm{i}}-\frac{\beta+\epsilon}{2}\right)||\Delta x_{\mathrm{i}}^{\nu}||^2+\frac{1}{2\epsilon}\left\|\delta_{\mathrm{i}}^{\nu}\right\|^2.
$$

Since under either Assumption 5.1.1 or Assumption 5.1.2 combined with Assumption 5.1.4 it holds that $\mu_{\mathrm{i}} \geq \max\{0, \mu - \beta\}$, we obtain

$$F(x_{\mathrm{i}}^{\nu+}) - F(x_{\mathrm{i}}^{\nu}) \leq -\left(\frac{M_{\mathrm{i}}}{2} - \frac{L}{6}\right)||\Delta x_{\mathrm{i}}^{\nu}||^3 - \left(\frac{\max(0, \mu - \beta)}{2} + \tau_{\mathrm{i}} - \frac{\beta + \epsilon}{2}\right)||\Delta x_{\mathrm{i}}^{\nu}||^2 + \frac{1}{2\epsilon}\left\|\delta_{\mathrm{i}}^{\nu}\right\|^2. \tag{5.31}$$

Denoting $p^{\nu+} \triangleq (1/m)\sum_{\mathrm{i}=1}^{m}\left\{F(x_{\mathrm{i}}^{\nu+}) - F(\widehat{x})\right\}$, we derive a simple relation with $p^{\nu+1}$:

$$\begin{aligned} p^{\nu+1} + F(\widehat{x}) = \frac{1}{m}\sum_{\mathrm{i}=1}^{m} F\left(x_{\mathrm{i}}^{\nu+1}\right) &\overset{(5.8b)}{=} \frac{1}{m}\sum_{\mathrm{i}=1}^{m} F\left(\sum_{\mathrm{j}=1}^{m}(W_K)_{\mathrm{i,j}}\, x_{\mathrm{j}}^{\nu+}\right) \\ &\overset{(a)}{\leq} \frac{1}{m}\sum_{\mathrm{i,j}=1}^{m}(W_K)_{\mathrm{i,j}} F\left(x_{\mathrm{j}}^{\nu+}\right) \overset{(b)}{=} \frac{1}{m}\sum_{\mathrm{j}=1}^{m} F\left(x_{\mathrm{j}}^{\nu+}\right) = p^{\nu+} + F(\widehat{x}), \end{aligned} \tag{5.32}$$

where (a) is due to convexity of F (cf. Assumptions 5.1.1 and 5.1.2) and $\sum_{\mathrm{j}=1}^{m}(W_K)_{\mathrm{ij}} = 1$ (cf. Assumption 5.2.1); and in (b) we used $\sum_{\mathrm{i}=1}^{m}(W_K)_{\mathrm{ij}} = 1$ (cf. Assumption 5.2.1). Summing (5.31) over i while setting $\epsilon = \beta$, $\tau_{\mathrm{i}} = 2\beta$ and $M_{\mathrm{i}} \geq L/3$ (recall that it is assumed $M_{\mathrm{i}} \geq L$), gives the desired per-iteration decrease of p^{ν} when $\left\|\delta^{\nu}\right\|$ is sufficiently small:

$$p^{\nu+1} - p^{\nu} \overset{(5.32)}{\leq} p^{\nu+} - p^{\nu} \leq -\frac{\max(\mu, \beta)}{2} \cdot \frac{1}{m}\sum_{\mathrm{i}=1}^{m}\|\Delta x_{\mathrm{i}}^{\nu}\|^2 + \frac{1}{2m\beta}\left\|\delta^{\nu}\right\|^2. \tag{5.33}$$

5.6.3.2 Network error bounds

The goal of this subsection is to prove an upper bound for $\left\|\delta^{\nu}\right\|$ in terms of the number of communication steps K, implying that this error can be made sufficiently small by choosing sufficiently large K. For notation simplicity and without loss of generality, we assume $d = 1$; the case $d > 1$ follows trivially.

Recall that $x^{\nu} \triangleq (x_{\mathrm{i}}^{\nu})_{\mathrm{i}=1}^{m}$, $y^{\nu} \triangleq (y_{\mathrm{i}}^{\nu})_{\mathrm{i}=1}^{m}$, $J \triangleq (1/m)1_m 1_m^{\top}$, and

$$x_{\perp}^{\nu} \triangleq (I - J)x^{\nu} = x^{\nu} - 1_m\frac{1_m^{\top}x^{\nu}}{m}, \quad y_{\perp}^{\nu} \triangleq (I - J)y^{\nu} = y^{\nu} - 1_m\frac{1_m^{\top}y^{\nu}}{m}, \quad \Delta x^{\nu} \triangleq (\Delta x_{\mathrm{i}}^{\nu})_{\mathrm{i}=1}^{m}.$$

Note that the vectors $x_\perp^\nu$ and $y_\perp^\nu$ are the consensus and gradient-tracking errors; when $\|x_\perp^\nu\| = \|y_\perp^\nu\| = 0$, we have $x_{\mathrm{i}}^\nu = x_{\mathrm{j}}^\nu$ and $y_{\mathrm{i}}^\nu = y_{\mathrm{j}}^\nu$ for all $\mathrm{i}, \mathrm{j} = 1, \ldots, m$. The following holds for $x_\perp^\nu$ and $y_\perp^\nu$.

Lemma 5.6.4 (Proposition 3.5 in [222]). *Under Assumptions 5.1.1 and 5.1.5 and 5.2.1, for all $\nu \geq 0$,*

$$\|x_\perp^{\nu+1}\| \leq \rho_K \|x_\perp^\nu\| + \rho_K \|\Delta x^\nu\|, \tag{5.34a}$$

$$\|y_\perp^{\nu+1}\| \leq \rho_K \|y_\perp^\nu\| + 2Q_{\max}\rho_K \|x_\perp^\nu\| + Q_{\max}\rho_K \|\Delta x^\nu\|, \tag{5.34b}$$

where $\rho_K = \lambda_{\max}(W_K - J) < 1$. Note that in case of K-rounds of communications using a reference matrix $\overline{W}$ with $\rho \triangleq \lambda_{\max}(\overline{W} - J) < 1$, we have $\rho_K = \rho^K$; if Chebyshev acceleration is employed, we have $\rho_K = (1 - \sqrt{1-\rho})^K$.

Now let us bound δ_{i}^ν defined in (5.18). Note that by column-stochasticity of W_K and initialization rule $s_{\mathrm{i}}^0 = \nabla f_{\mathrm{i}}(x_{\mathrm{i}}^0)$, it can be trivially concluded from (5.8c) that

$$1_m^\top y^\nu = \sum_{\mathrm{j}=1}^m \nabla f_{\mathrm{j}}(x_{\mathrm{j}}^\nu).$$

Hence,

$$\begin{aligned}
\left\|\delta_{\mathrm{i}}^\nu\right\|^2 =& \left\| y_{\mathrm{i}}^\nu - \frac{1_m^\top y^\nu}{m} + \frac{1}{m}\sum_{\mathrm{j}=1}^m \nabla f_{\mathrm{j}}(x_{\mathrm{j}}^\nu) - \nabla F(x_{\mathrm{i}}^\nu) \right\|^2 \\
\overset{(a)}{\leq}& 2\left\| y_{\mathrm{i}}^\nu - \frac{1_m^\top y^\nu}{m} \right\|^2 + \frac{2Q_{\max}^2}{m}\left(\sum_{\mathrm{j}=1}^m \left\| x_{\mathrm{i}}^\nu \pm \frac{1_m^\top x^\nu}{m} - x_{\mathrm{j}}^\nu \right\|^2 \right) \\
\leq& 2\left\| y_{\mathrm{i}}^\nu - \frac{1_m^\top y^\nu}{m} \right\|^2 + \frac{4Q_{\max}^2}{m}\left(\|x_\perp^\nu\|^2 + m \left\| x_{\mathrm{i}}^\nu - \frac{1_m^\top x^\nu}{m} \right\|^2 \right),
\end{aligned} \tag{5.35}$$

where (a) is due to $Q_{\max}$-Lipschitz continuity of ∇f_{i}. Summing (5.35) over i and taking the square root, gives

$$\left\|\delta^\nu\right\| \leq \tilde{\delta}^\nu \triangleq \sqrt{2}\left(\|y_\perp^\nu\| + 2Q_{\max}\|x_\perp^\nu\|\right). \tag{5.36}$$

It remains to bound $\tilde{\delta}^{\nu}$ defined above:

$$\begin{aligned}\tilde{\delta}^{\nu+1} = \sqrt{2}\left(\|y_{\perp}^{\nu+1}\| + 2Q_{\max}\|x_{\perp}^{\nu+1}\|\right) \overset{(a)}{\leq} &\rho_K\sqrt{2}\left(\|y_{\perp}^{\nu}\| + 4Q_{\max}\|x_{\perp}^{\nu}\|\right) + 3\sqrt{2}Q_{\max}\rho_K\|\Delta x^{\nu}\| \\ \leq &2\rho_K\tilde{\delta}^{\nu} + 3\sqrt{2}Q_{\max}\rho_K\|\Delta x^{\nu}\|,\end{aligned}$$

where in (a) we used Lemma 5.6.4 [cf. (5.34a)-(5.34b)]. Consequently,

$$(\tilde{\delta}^{\nu+1})^2 \leq 8\rho_K^2(\tilde{\delta}^{\nu})^2 + 36Q_{\max}^2\rho_K^2\|\Delta x^{\nu}\|^2. \tag{5.37}$$

Since ρ_K decreases as K increases, the latter inequality provides a leverage to make $\tilde{\delta}^{\nu+1}$ sufficiently small by choosing K sufficiently large.

5.6.3.3 Asymptotic convergence

We combine the results of the previous two subsections to finally prove Theorem 5.6.1. Combining (5.33) and (5.36), we obtain

$$p^{\nu+1} \leq p^{\nu} - \frac{\max(\beta,\mu)}{2m}\|\Delta x^{\nu}\|^2 + \frac{1}{2m\beta}(\tilde{\delta}^{\nu})^2. \tag{5.38}$$

Next, we combine (5.37) with (5.38) multiplied by some weight $w > 0$ to obtain

$$wp^{\nu+1} + (\tilde{\delta}^{\nu+1})^2 \leq wp^{\nu} + \left(8\rho_K^2 + \frac{w}{2m\beta}\right)(\tilde{\delta}^{\nu})^2 - w\left(\frac{\max(\beta,\mu)}{2m} - \frac{36}{w}Q_{\max}^2\rho_K^2\right)||\Delta x^{\nu}||^2. \tag{5.39}$$

Let $w = c_w\beta$, for some $0 < c_w \leq 1$. Then, if

$$8\rho_K^2 + \frac{w}{2m\beta} \leq c_w, \quad \frac{\max(\beta,\mu)}{4m} \geq \frac{36}{w}Q_{\max}^2\rho_K^2, \tag{5.40}$$

(5.39) becomes

$$wp^{\nu+1} + (\tilde{\delta}^{\nu+1})^2 \leq wp^{\nu} + c_w(\tilde{\delta}^{\nu})^2 - \frac{w\max(\beta,\mu)}{4m}||\Delta x^{\nu}||^2. \tag{5.41}$$

Note that by Lemma 5.6.4, condition (5.40) holds if

$$K \geq \frac{1}{\sqrt{1-\rho}} \log \left(\max \left\{ \frac{2\sqrt{2}}{\sqrt{c_w(1-\frac{1}{2m})}}, \frac{12\sqrt{m}Q_{\max}}{\sqrt{c_w \beta \max(\beta,\mu)}} \right\} \right). \tag{5.42}$$

Denoting

$$\xi^\nu \triangleq wp^\nu + (\tilde{\delta}^\nu)^2, \tag{5.43}$$

let us show that $\xi^\nu \to 0$ as $\nu \to \infty$, which implies that the optimization error p^ν and network error $\tilde{\delta}^\nu$ asymptotically vanish. Since $\xi^\nu \geq 0$, inequality (5.41) implies $\sum_{\nu=0}^{\infty} ||\Delta x^\nu||^2 < \infty$. Thus, $||\Delta x^\nu|| \to 0$; and $||\Delta x^\nu|| \leq D_1$, for some $D_1 > 0$ and all $\nu \geq 0$. Further, $\{\xi^\nu\}_\nu$ is non-increasing and $||\xi^\nu|| \leq D_2$ for some $D_2 > 0$ and all $\nu \geq 0$. Thus, $p^\nu \leq D_2/w$, which together with Assumption 5.1.1(iv) and Assumption 5.1.2, also implies $||x_{\mathrm{i}}^\nu|| \leq D_3$ for some D_3, all i and $\nu \geq 0$. Using $||\Delta x^\nu|| \to 0$ and (5.37), if $8\rho_K^2 < 1$ (which holds under (5.42)), we obtain that $\tilde{\delta}^\nu \to 0$. Finally, it remains to show that $p^\nu \to 0$. Using optimality condition of $x_{\mathrm{i}}^{\nu+}$ defined in (5.8a), we get

$$\left\langle \nabla F(x_{\mathrm{i}}^\nu) + \delta_{\mathrm{i}}^\nu + \left[\nabla^2 F(x_{\mathrm{i}}^\nu) + B_{\mathrm{i}}^\nu + \tau_{\mathrm{i}} I\right] \Delta x_{\mathrm{i}}^\nu + \frac{M_{\mathrm{i}}}{2} ||\Delta x_{\mathrm{i}}^\nu|| \Delta x_{\mathrm{i}}^\nu, \widehat{x} - x_{\mathrm{i}}^{\nu+} \right\rangle \geq 0.$$

Rearranging terms gives

$$\begin{aligned} &\left\langle \nabla F(x_{\mathrm{i}}^\nu) + \nabla^2 F(x_{\mathrm{i}}^\nu) \Delta x_{\mathrm{i}}^\nu, \widehat{x} - x_{\mathrm{i}}^{\nu+} \right\rangle \\ \geq &\left\langle \frac{M_{\mathrm{i}}}{2} ||\Delta x_{\mathrm{i}}^\nu|| \Delta x_{\mathrm{i}}^\nu, x_{\mathrm{i}}^{\nu+} - \widehat{x} \right\rangle + \left\langle \delta_{\mathrm{i}}^\nu, x_{\mathrm{i}}^{\nu+} - \widehat{x} \right\rangle + \left\langle \tilde{B}_{\mathrm{i}}^\nu \Delta x_{\mathrm{i}}^\nu, x_{\mathrm{i}}^{\nu+} - \widehat{x} \right\rangle, \end{aligned} \tag{5.44}$$

where $\tilde{B}_{\mathrm{i}}^\nu \triangleq B_{\mathrm{i}}^\nu + \tau_{\mathrm{i}} I$. By convexity of F, we can write

$$\begin{aligned} 0 \geq &F(\widehat{x}) - F(x_{\mathrm{i}}^{\nu+}) \geq \left\langle \nabla F(x_{\mathrm{i}}^{\nu+}), \widehat{x} - x_{\mathrm{i}}^{\nu+} \right\rangle \\ = &\left\langle \nabla F(x_{\mathrm{i}}^{\nu+}) - \nabla F(x_{\mathrm{i}}^\nu) - \nabla^2 F(x_{\mathrm{i}}^\nu) \Delta x_{\mathrm{i}}^\nu, \widehat{x} - x_{\mathrm{i}}^{\nu+} \right\rangle + \left\langle \nabla F(x_{\mathrm{i}}^\nu) + \nabla^2 F(x_{\mathrm{i}}^\nu) \Delta x_{\mathrm{i}}^\nu, \widehat{x} - x_{\mathrm{i}}^{\nu+} \right\rangle \\ \overset{(5.44)}{\geq} &\left\langle \nabla F(x_{\mathrm{i}}^{\nu+}) - \nabla F(x_{\mathrm{i}}^\nu) - \nabla^2 F(x_{\mathrm{i}}^\nu) \Delta x_{\mathrm{i}}^\nu, \widehat{x} - x_{\mathrm{i}}^{\nu+} \right\rangle + \left\langle \frac{M_{\mathrm{i}}}{2} ||\Delta x_{\mathrm{i}}^\nu|| \Delta x_{\mathrm{i}}^\nu, x_{\mathrm{i}}^{\nu+} - \widehat{x} \right\rangle \\ &+ \left\langle \delta_{\mathrm{i}}^\nu, x_{\mathrm{i}}^{\nu+} - \widehat{x} \right\rangle + \left\langle \tilde{B}_{\mathrm{i}}^\nu \Delta x_{\mathrm{i}}^\nu, x_{\mathrm{i}}^{\nu+} - \widehat{x} \right\rangle. \end{aligned} \tag{5.45}$$

Using Lipschitz continuity of ∇F, $||\Delta x_{\rm i}^{\nu}|| \to 0$ and $\tilde{\delta}^{\nu} \to 0$ (hence $||\delta_{\rm i}^{\nu}|| \to 0$), we conclude that the RHS of (5.45) asymptotically vanishes, for all ${\rm i} = 1, \ldots, m$. Hence, $F(x_{\rm i}^{\nu+}) - F(\widehat{x}) \to 0$, for all ${\rm i} = 1, \ldots, m$. Using (5.32), we finally obtain $p^{\nu} \to 0$.

Finally, by (5.36) and $\tilde{\delta}^{\nu} \to 0$, we obtain $\|y_{\perp}^{\nu}\| \to 0$ and $\|x_{\perp}^{\nu}\| \to 0$, implying $||x_{\rm i}^{\nu} - x_{\rm j}^{\nu}|| \to 0$, for all ${\rm i}, {\rm j} = 1, \ldots, m$ as $\nu \to \infty$. This concludes the proof of Theorem 5.6.1.

Remark 5.6.1. *Note that* (5.37) *implies*

$$(\tilde{\delta}^{\nu})^2 \leq \rho_K^2 \bar{D}_{\delta}, \quad \bar{D}_{\delta} \triangleq 8D_2 + 36Q_{\max}^2 D_1^2, \quad \forall \nu \geq 0, \tag{5.46}$$

since $(\tilde{\delta}^{\nu})^2 \leq \xi^{\nu} \leq D_2$ *and* $||\Delta x^{\nu}|| \leq D_1$, *for all* $\nu \geq 0$.

5.6.4 Proof of Theorem 5.3.1

We first prove a detailed "region-based" complexity of DiRegINA (cf. Theorem 5.6.2, Subsec. 5.6.4.1) for the prevalent scenario $0 < \beta \leq 1$ [recall that typically $\beta = \mathcal{O}(1/\sqrt{n})$]. For the sake of completeness, the case $\beta \geq 1$ is studied in Theorem 5.6.3 (cf. Subsec. 5.6.4.2). Building on Theorems 5.6.2-5.6.3, we can finally prove the main result, Theorem 5.3.1 (cf. Subsec. 5.6.4.3).

5.6.4.1 Complexity Analysis when $0 < \beta \leq 1$

Theorem 5.6.2 ($0 < \beta \leq 1$ and $L > 0$)**.** *Let Assumptions 5.1.1 and 5.1.3-5.1.4 and 5.1.5 hold along with* $0 < \beta \leq 1$. *Let* $M_{\rm i} = L > 0$, $\tau_{\rm i} = 2\beta$, *and recall the definition of* $D > 0$ *implying* $||x_{\rm i}^0 - \widehat{x}|| \leq D$, *for all* ${\rm i} = 1, \ldots, m$. *W.l.o.g. assume* $D \geq 2/L$. *Pick an accuracy* $\varepsilon > 0$. *If a reference matrix* $\overline{W}$ *satisfying Assumption 5.2.1 is used in steps* (5.8b)-(5.8c), *with* $\rho \triangleq \lambda_{\max}(\overline{W} - J) < 1$ *and* $K = \tilde{\mathcal{O}}(\log(1/\varepsilon)/\sqrt{1-\rho})$ *(the explicit expression of* K *can be found in* (5.64)*), then the sequence* $\{p^{\nu}\}$ *generated by DiRegINA satisfies the following:*

(a) if $p^{\nu} \geq 2LD^3$,

$$p^{\nu+1} \leq \frac{5}{6}\, p^{\nu},$$

(b) if $\beta^2 \cdot (2LD^3) \leq p^\nu \leq 2LD^3$,

$$p^\nu \leq \frac{244 \cdot LD^3}{\nu^2},$$

(c) if $\varepsilon \leq p^\nu \leq \beta^2 \cdot (2LD^3)$,

$$p^\nu \leq 24^2 \cdot (LD^3)^2 \cdot \frac{\beta^2}{\varepsilon} \cdot \frac{1}{\nu^2}.$$

Proof. Recalling Lemma 5.6.3 from the proof of Theorem 5.6.1, we can write

$$F(x_{\mathrm{i}}^{\nu+}) \leq \widetilde{F}_{\mathrm{i}}(x_{\mathrm{i}}^{\nu+}; x_{\mathrm{i}}^\nu) + \frac{1}{2\epsilon}\left\|\delta_{\mathrm{i}}^\nu\right\|^2, \tag{5.47}$$

for arbitrary $\epsilon > 0$, $M_{\mathrm{i}} \geq L$, and $\tau_{\mathrm{i}} \geq \beta + \epsilon$. In addition, by the upperbound approximation of $\widetilde{F}_{\mathrm{i}}(\cdot; x_{\mathrm{i}}^\nu)$ in (5.22), there holds

$$\widetilde{F}_{\mathrm{i}}(y; x_{\mathrm{i}}^\nu) \leq F(y) + \frac{1}{2}\left\|y - x_{\mathrm{i}}^\nu\right\|^2_{(\beta+\tau_{\mathrm{i}}+\epsilon)I} + \frac{M_{\mathrm{i}} + L}{6}\left\|y - x_{\mathrm{i}}^\nu\right\|^3 + \frac{1}{2\epsilon}\left\|\delta_{\mathrm{i}}^\nu\right\|^2, \quad \forall y \in \mathcal{K}. \tag{5.48}$$

Let $\alpha_0 \in (0, 1]$. Set $\epsilon = \beta$ and $\tau_{\mathrm{i}} = 2\beta$. By (5.47)-(5.48) and $x_{\mathrm{i}}^{\nu+}$ being the minimizer of $\widetilde{F}(\cdot; x_{\mathrm{i}}^\nu)$ [see (5.8a)], we obtain

$$\begin{aligned}
&F(x_{\mathrm{i}}^{\nu+}) - F(\widehat{x}) \\
&\leq \min_{y \in \mathcal{K}} \left\{ F(y) - F(\widehat{x}) + 2\beta\left\|y - x_{\mathrm{i}}^\nu\right\|^2 + \frac{M_{\mathrm{i}} + L}{6}\left\|y - x_{\mathrm{i}}^\nu\right\|^3 + \frac{1}{\beta}\left\|\delta_{\mathrm{i}}^\nu\right\|^2 \right\} \\
&\leq \min_{\alpha \in [0, \alpha_0]} \left\{ F(y) - F(\widehat{x}) + 2\beta\left\|y - x_{\mathrm{i}}^\nu\right\|^2 + \frac{M_{\mathrm{i}} + L}{6}\left\|y - x_{\mathrm{i}}^\nu\right\|^3 + \frac{1}{\beta}\left\|\delta_{\mathrm{i}}^\nu\right\|^2 \right. \\
&\qquad\qquad \left. : y = \alpha\widehat{x} + (1 - \alpha)x_{\mathrm{i}}^\nu \right\} \\
&\leq \min_{\alpha \in [0, \alpha_0]} \left\{ (1 - \alpha)\left(F(x_{\mathrm{i}}^\nu) - F(\widehat{x})\right) \right. \\
&\qquad\qquad \left. + 2\beta\alpha^2\left\|\widehat{x} - x_{\mathrm{i}}^\nu\right\|^2 + \frac{M_{\mathrm{i}} + L}{6}\alpha^3\left\|\widehat{x} - x_{\mathrm{i}}^\nu\right\|^3 + \frac{1}{\beta}\left\|\delta_{\mathrm{i}}^\nu\right\|^2 \right\},
\end{aligned} \tag{5.49}$$

where the last inequality holds by the convexity of F. Note that, by definition, $||x_{\mathrm{i}}^0 - \widehat{x}|| \leq D$, for all $\mathrm{i} = 1, \ldots, m$. Assuming $||x_{\mathrm{i}}^\nu - \widehat{x}|| \leq D$, for all $\mathrm{i} = 1, \ldots, m$, we prove descent at iteration $\nu + 1$, i.e. $p^{\nu+1} < p^\nu$, unless $p^\nu = 0$. Note that by Assumption 5.1.1(iv), if $\{p^\nu\}_\nu$ is

non-increasing, then $||x_{\mathrm{i}}^{\nu} - \widehat{x}|| \leq D$ for all $\nu \geq 0$ and $\mathrm{i} = 1, \ldots, m$. Now set $M_{\mathrm{i}} = L$ in (5.49) and compute the mean over $\mathrm{i} = 1, \ldots, m$, which yields

$$p^{\nu+1} \overset{(5.32)}{\leq} p^{\nu+} \leq \min_{\alpha \in [0, \alpha_0]} \left\{ (1-\alpha) p^{\nu} + 2\beta\alpha^2 D^2 + \frac{LD^3}{3}\alpha^3 + \frac{1}{m\beta}\left\|\delta^{\nu}\right\|^2 \right\}. \tag{5.50}$$

Denote

$$C_1 \triangleq \frac{LD^3}{3}. \tag{5.51}$$

Since $D \geq \frac{2}{L}$, it holds $2\beta D^2 \leq 3\beta C_1$. Then, setting $\alpha_0 = \min\{1, p^{\nu}/(6\beta C_1)\}$ in (5.50) yields

$$\begin{aligned} p^{\nu+1} &\leq \min_{\alpha \in [0, \min\{1, \frac{p^{\nu}}{6\beta C_1}\}]} \left\{ (1-\alpha) p^{\nu} + 3\beta C_1 \alpha^2 + C_1 \alpha^3 + \frac{1}{m\beta}\left\|\delta^{\nu}\right\|^2 \right\} \\ &\leq \min_{\alpha \in [0, \min\{1, \frac{p^{\nu}}{6\beta C_1}\}]} \left\{ (1-\alpha/2) p^{\nu} + C_1 \alpha^3 + \frac{1}{m\beta}\left\|\delta^{\nu}\right\|^2 \right\}. \end{aligned} \tag{5.52}$$

Let us assess (5.52) over the following "regions". Denoting by α^* the minimizer of the optimization problem at the RHS of (5.52), we have the following:

(a) If $p^{\nu} \geq 6C_1$, then $\alpha^* = 1$ and

$$p^{\nu+1} \leq \frac{1}{2} p^{\nu} + C_1 + \frac{1}{m\beta}\left\|\delta^{\nu}\right\|^2 \leq \left(\frac{1}{2} + \frac{1}{6}\right) p^{\nu} + \frac{1}{m\beta}\left\|\delta^{\nu}\right\|^2, \tag{5.53}$$

and under the condition

$$\frac{1}{m\beta}\left\|\delta^{\nu}\right\|^2 \leq \frac{1}{6} p^{\nu} \impliedby \frac{1}{m\beta}\left\|\delta^{\nu}\right\|^2 \leq C_1, \tag{5.54}$$

(5.53) yields

$$p^{\nu+1} \leq \frac{5}{6}\, p^{\nu}.$$

Note that, by (5.46) and Lemma 5.6.4, condition (5.54) holds if

$$K \geq \frac{1}{\sqrt{1-\rho}} \cdot \frac{1}{2} \log\left(\frac{\bar{D}_{\delta}}{m\beta C_1}\right). \tag{5.55}$$

(b) If $6\beta^2 C_1 \le p^\nu \le 6C_1$, then $\alpha^* = \sqrt{\frac{p^\nu}{6C_1}}$ and

$$p^{\nu+1} \le p^\nu - \frac{(p^\nu)^{3/2}}{3\sqrt{6C_1}} + \frac{1}{m\beta}\left\|\delta^\nu\right\|^2, \tag{5.56}$$

and if (similar to derivation of (5.55))

$$K \ge \frac{1}{\sqrt{1-\rho}} \cdot \frac{1}{2}\log\left(\frac{\bar{D}_\delta}{m\beta^4 C_1}\right) \implies \frac{1}{m\beta}\left\|\delta^\nu\right\|^2 \le \beta^3 C_1 \implies \frac{1}{m\beta}\left\|\delta^\nu\right\|^2 \le \frac{(p^\nu)^{3/2}}{6\sqrt{6C_1}}, \tag{5.57}$$

(5.56) implies

$$p^{\nu+1} \le p^\nu - \frac{(p^\nu)^{3/2}}{6\sqrt{6C_1}}. \tag{5.58}$$

Finally, since p^ν is non-increasing,

$$\begin{aligned}\frac{1}{\sqrt{p^{\nu+1}}} - \frac{1}{\sqrt{p^\nu}} =& \frac{p^\nu - p^{\nu+1}}{\left(\sqrt{p^\nu}+\sqrt{p^{\nu+1}}\right)\sqrt{p^\nu p^{\nu+1}}} \overset{(5.58)}{\ge} \frac{\frac{1}{6\sqrt{6C_1}}(p^\nu)^{3/2}}{\left(\sqrt{p^\nu}+\sqrt{p^{\nu+1}}\right)\sqrt{p^\nu p^{\nu+1}}} \\ \ge& c_0 \triangleq \frac{1}{12}\sqrt{\frac{1}{6C_1}},\end{aligned}$$

and consequently,

$$p^\nu \le \frac{1}{c_0^2\left(\nu + \frac{1}{c_0\sqrt{p^0}}\right)^2} \le \frac{1}{c_0^2\nu^2}.$$

(c) If $\varepsilon \le p^\nu \le 6\beta^2 C_1$, then $\alpha^* = \frac{p^\nu}{6\beta C_1}$ and

$$p^{\nu+1} \le p^\nu - \frac{(p^\nu)^2}{18\beta C_1} + \frac{1}{m\beta}\left\|\delta^\nu\right\|^2, \tag{5.59}$$

and if (similar to derivation of (5.55))

$$K \ge \frac{1}{\sqrt{1-\rho}} \cdot \frac{1}{2}\log\left(\frac{36C_1\bar{D}_\delta}{m\varepsilon^2}\right) \implies \frac{1}{m\beta}\left\|\delta^\nu\right\|^2 \le \frac{\varepsilon^2}{36\beta C_1} \implies \frac{1}{m\beta}\left\|\delta^\nu\right\|^2 \le \frac{(p^\nu)^2}{36\beta C_1}, \tag{5.60}$$

we deduce from (5.59)

$$p^{\nu+1} \le p^\nu - \frac{(p^\nu)^2}{36\beta C_1}. \tag{5.61}$$

Since p^ν is non-increasing,

$$\begin{aligned}\frac{1}{\sqrt{p^{\nu+1}}}-\frac{1}{\sqrt{p^\nu}}=&\frac{p^\nu-p^{\nu+1}}{\left(\sqrt{p^\nu}+\sqrt{p^{\nu+1}}\right)\sqrt{p^\nu p^{\nu+1}}}\overset{(5.61)}{\geq}\frac{\frac{1}{36\beta C_1}(p^\nu)^2}{\left(\sqrt{p^\nu}+\sqrt{p^{\nu+1}}\right)\sqrt{p^\nu p^{\nu+1}}}\\ \geq&\tilde{c}_0\triangleq\frac{\sqrt{\varepsilon}}{72\beta C_1},\end{aligned}\tag{5.62}$$

and consequently,

$$p^\nu\leq\frac{1}{\tilde{c}_0^2\left(\nu+\frac{1}{\tilde{c}_0\sqrt{p^0}}\right)^2}\leq\frac{1}{\tilde{c}_0^2\nu^2}=72^2\cdot C_1^2\cdot\frac{\beta^2}{\varepsilon}\cdot\frac{1}{\nu^2}.\tag{5.63}$$

Finally, combining all the conditions (5.42), (5.55),(5.57), and (5.60), the requirement on K reads

$$K\geq\frac{1}{\sqrt{1-\rho}}\cdot\frac{1}{2}\log\left(\max\left\{\frac{16}{c_w},\frac{12^2mQ_{\max}^2}{c_w\beta\max(\beta,\mu)},\frac{\bar{D}_\delta}{\min\left\{m\beta C_1,m\beta^4C_1,\frac{m}{36C_1}\varepsilon^2\right\}}\right\}\right),\tag{5.64}$$

where $\bar{D}_\delta$ and C_1 are defined in (5.46) and (5.51), respectively. □

5.6.4.2 Complexity Analysis when $\beta\geq 1$

Theorem 5.6.3 ($\beta\geq 1$ and $L>0$)**.** *Let Assumptions 5.1.1 and 5.1.3-5.1.4and 5.1.5 hold and $\beta\geq 1$. Let $M_{\mathrm{i}}=L>0$, $\tau_{\mathrm{i}}=2\beta$, and recall the definition of $D>0$ implying $\max_{\mathrm{i}\in[m]}||x_{\mathrm{i}}^0-\widehat{x}||\leq D$. W.l.o.g. assume $D\geq 2/L$. Pick an arbitrary $\varepsilon>0$. If a reference matrix $\overline{W}$ satisfying Assumption 5.2.1 is used in steps (5.8b)-(5.8c), with $\rho\triangleq\lambda_{\max}(\overline{W}-J)<1$ and $K=\widetilde{\mathcal{O}}(\log(1/\varepsilon)/\sqrt{1-\rho})$ (the explicit expression is given in (5.64)), then the sequence $\{p^\nu\}$ generated by DiRegINA satisfies the following:*

(a) if $p^\nu\geq\beta\cdot(2LD^3)$,

$$p^{\nu+1}\leq\frac{5}{6}\,p^\nu,$$

(b) if $\varepsilon\leq p^\nu\leq\beta\cdot(2LD^3)$,

$$p^\nu\leq 24^2\cdot(LD^3)^2\cdot\frac{\beta^2}{\varepsilon}\cdot\frac{1}{\nu^2}.$$

Proof. Excluding β, the parameter setting is identical to Theorem 5.6.2. Recall (5.52), i.e.,

$$p^{\nu+1} \leq \min_{\alpha \in [0, \min\{1, \frac{p^\nu}{6\beta C_1}\}]} \left\{ (1-\alpha/2)p^\nu + C_1\alpha^3 + \frac{1}{m\beta}\left\|\delta^\nu\right\|^2 \right\}, \tag{5.65}$$

where C_1 is defined in (5.51). Denoting by α^* the minimizer of the optimization problem at the RHS of (5.52), we have:

(a) If $p^\nu \geq 6\beta C_1$, then $\alpha^* = 1$ and under (5.64), (5.65) yields

$$p^{\nu+1} \leq \frac{4 + 1/\beta}{6} p^\nu \leq \frac{5}{6}\, p^\nu.$$

(b) If $\varepsilon \leq p^\nu \leq 6\beta C_1$, then $\alpha^* = \frac{p^\nu}{6\beta C_1}$. Under (5.64), (5.65) yields

$$p^{\nu+1} \leq p^\nu - \frac{(p^\nu)^2}{36\beta C_1},$$

and following similar steps as in derivation of (5.63), we obtain

$$p^\nu \leq \frac{1}{\tilde{c}_0^2 \nu^2} = 72^2 \cdot C_1^2 \cdot \frac{\beta^2}{\varepsilon} \cdot \frac{1}{\nu^2}.$$

□

5.6.4.3 Proof of main theorem

We proceed to prove Theorem 5.3.1. Given an accuracy $0 < \varepsilon \ll 1$, when $0 < \beta \leq 1$, Theorem 5.6.2 gives the following expression of rate: to achieve $p^\nu \leq \varepsilon$, DiRegINA requires

$$O\left(\log\left(\frac{1}{6C_1}\right) + \sqrt{\frac{LD^3}{\varepsilon}} + \frac{\beta\,(LD^3)}{\varepsilon}\right) = \tilde{O}\left(\sqrt{\frac{LD^3}{\varepsilon}} + \frac{\beta\,(LD^3)}{\varepsilon}\right), \tag{5.66}$$

iterations, while if $\beta \geq 1$, by Theorem 5.6.3, DiRegINA requires

$$O\left(\log\left(\frac{1}{2\beta LD^3}\right) + \frac{\beta\,(LD^3)}{\varepsilon}\right) = \tilde{O}\left(\frac{\beta\,(LD^3)}{\varepsilon}\right)$$

iterations. Therefore, (5.66) is a valid rate complexity expression (in terms of iterations) in both discussed cases (i.e. $0 < \beta \leq 1$ and $\beta \geq 1$). Now, recall that every iteration requires K rounds of communications, with K satisfying (5.42) and (5.64); hence $K = \tilde{O}\left(1/\sqrt{1-\rho} \cdot \log(1/\varepsilon)\right) = \tilde{O}\left(1/\sqrt{1-\rho} \cdot \varepsilon^{-\alpha/2}\right)$, for any arbitrary small $\alpha > 0$. Therefore the final communication complexity reads

$$\tilde{O}\left(\frac{1}{\sqrt{1-\rho}} \cdot \left\{\sqrt{\frac{LD^3}{\varepsilon^{1+\alpha}}} + \frac{\beta\left(LD^3\right)}{\varepsilon^{1+\frac{\alpha}{2}}}\right\}\right).$$

5.6.5 Proof of Theorem 5.3.2 and Corollary 5.3.3

We begin introducing some intermediate technical results, instrumental to proving the main theorems, namely: i) Lemmata 5.6.6-5.6.5 in Sec. 5.6.5.1; and ii) a detailed "region-based" complexity of DiRegINA as in in Theorem 5.6.4 (cf. Sec. 5.6.5.2). We prove Theorem 5.3.2 and the improved rates in case of quadratic functions in Sec. 5.6.5.3 and Sec. 5.6.5.4, respectively. Finally, Corollary 5.3.3 is proved in Sec. 5.6.5.5.

5.6.5.1 Connections between the optimization error, network error and $||\Delta x^\nu||$

We establish necessary connections between the optimization error p^ν, the network error $\left\|\delta^\nu\right\|$ and $||\Delta x^\nu||$ in Lemmata 5.6.5-5.6.6:

Lemma 5.6.5. *Let Assumptions 5.1.2-5.1.4 hold, $\tau_{\mathrm{i}} = 2\beta$, and $M_{\mathrm{i}} \geq L/3$. Then*

$$\frac{1}{m}\sum_{\mathrm{i}=1}^{m}\left\|\Delta x_{\mathrm{i}}^\nu\right\|^2 \leq \frac{8}{\mu}\, p^\nu + \frac{2}{m\beta\mu}\left\|\delta^\nu\right\|^2, \tag{5.67}$$

where p^ν is defined in (5.10).

Proof. By μ-strongly convexity of F and optimality of $\widehat{x}$,

$$\begin{aligned} F(x_{\mathrm{i}}^{\nu+}) - F(\widehat{x}) \geq & \frac{\mu}{2}\left\|x_{\mathrm{i}}^{\nu+} - \widehat{x}\right\|^2 \geq \frac{\mu}{4}\left\|x_{\mathrm{i}}^{\nu+} - x_{\mathrm{i}}^{\nu}\right\|^2 - \frac{\mu}{2}\left\|x_{\mathrm{i}}^{\nu} - \widehat{x}\right\|^2 \\ \geq & \frac{\mu}{4}\left\|x_{\mathrm{i}}^{\nu+} - x_{\mathrm{i}}^{\nu}\right\|^2 - \left(F(x_{\mathrm{i}}^{\nu}) - F(\widehat{x})\right). \end{aligned}$$

Averaging the above inequalities over $\mathrm{i} = 1, \ldots, m$, yields

$$\frac{1}{m}\sum_{\mathrm{i}=1}^{m}\left\|\Delta x_{\mathrm{i}}^{\nu}\right\|^2 \leq \frac{4}{\mu}\left(p^{\nu+}+p^{\nu}\right),$$

where $p^{\nu+} = (1/m)\sum_{\mathrm{i}=1}^{m}\left\{F(x_{\mathrm{i}}^{\nu+}) - F(\widehat{x})\right\}$. Using (5.33) proves (5.67). □

Lemma 5.6.6. *Let Assumptions 5.1.2-5.1.4 hold and set $\tau_{\mathrm{i}} = 2\beta$. Define*

$$\omega_0 \triangleq \frac{12\beta}{\sqrt{L^2+4M_{\max}^2}}, \quad M_{\max} \triangleq \max_{\mathrm{i}\in[m]} M_{\mathrm{i}}.$$

Then

$$\frac{1}{m}\sum_{\mathrm{i}=1}^{m}\left\{F(x_{\mathrm{i}}^{\nu+}) - F(\widehat{x})\right\} \leq \varphi\left(\{x_{\mathrm{i}}^{\nu+}\}_{\mathrm{i}}, \{x_{\mathrm{i}}^{\nu}\}_{\mathrm{i}}\right) + \frac{8}{m\mu}\left\|\delta^{\nu}\right\|^2, \tag{5.68}$$

where

$$\varphi\left(\{x_{\mathrm{i}}^{\nu+}\}_{\mathrm{i}}, \{x_{\mathrm{i}}^{\nu}\}_{\mathrm{i}}\right) = \begin{cases} \frac{L^2+4M_{\max}^2}{m\mu}\left(\sum_{\mathrm{i}=1}^{m}\left\|x_{\mathrm{i}}^{\nu+}-x_{\mathrm{i}}^{\nu}\right\|^2\right)^2, & \text{if} \quad \mathcal{C}\text{:}\ \sqrt{\sum_{\mathrm{i}=1}^{m}\left\|x_{\mathrm{i}}^{\nu+}-x_{\mathrm{i}}^{\nu}\right\|^2} \geq \omega_0; \\ \frac{144\beta^2}{m\mu}\sum_{\mathrm{i}=1}^{m}\left\|x_{\mathrm{i}}^{\nu+}-x_{\mathrm{i}}^{\nu}\right\|^2, & \text{if} \quad \overline{\mathcal{C}}\text{:}\ \sqrt{\sum_{\mathrm{i}=1}^{m}\left\|x_{\mathrm{i}}^{\nu+}-x_{\mathrm{i}}^{\nu}\right\|^2} < \omega_0. \end{cases}$$

Proof. Recall (5.44), a consequence of optimality of $x_{\mathrm{i}}^{\nu+}$ (defined in (5.8a)), reads

$$\begin{aligned} &\left\langle \nabla F(x_{\mathrm{i}}^{\nu}) + \nabla^2 F(x_{\mathrm{i}}^{\nu})\Delta x_{\mathrm{i}}^{\nu}, \widehat{x} - x_{\mathrm{i}}^{\nu+}\right\rangle \\ \geq &\left\langle \frac{M_{\mathrm{i}}}{2}||\Delta x_{\mathrm{i}}^{\nu}||\Delta x_{\mathrm{i}}^{\nu}, x_{\mathrm{i}}^{\nu+} - \widehat{x}\right\rangle + \left\langle \delta_{\mathrm{i}}^{\nu}, x_{\mathrm{i}}^{\nu+} - \widehat{x}\right\rangle + \left\langle \tilde{B}_{\mathrm{i}}^{\nu}\Delta x_{\mathrm{i}}^{\nu}, x_{\mathrm{i}}^{\nu+} - \widehat{x}\right\rangle, \end{aligned} \tag{5.69}$$

where $\tilde{B}_{\mathrm{i}}^{\nu} = B_{\mathrm{i}}^{\nu} + \tau_{\mathrm{i}} I$ and recall $\Delta x_{\mathrm{i}}^{\nu} = x_{\mathrm{i}}^{\nu+} - x_{\mathrm{i}}^{\nu}$ [cf. (5.23)]. By μ-strongly convexity of F,

$$
\begin{aligned}
& F(\widehat{x}) - F(x_{\mathrm{i}}^{\nu+}) \\
\geq & \left\langle \nabla F(x_{\mathrm{i}}^{\nu+}), \widehat{x} - x_{\mathrm{i}}^{\nu+} \right\rangle + \frac{\mu}{2} \left\| x_{\mathrm{i}}^{\nu+} - \widehat{x} \right\|^2 \\
= & \left\langle \nabla F(x_{\mathrm{i}}^{\nu+}) - \nabla F(x_{\mathrm{i}}^{\nu}) - \nabla^2 F(x_{\mathrm{i}}^{\nu}) \Delta x_{\mathrm{i}}^{\nu}, \widehat{x} - x_{\mathrm{i}}^{\nu+} \right\rangle + \frac{\mu}{2} \left\| x_{\mathrm{i}}^{\nu+} - \widehat{x} \right\|^2 \\
& + \left\langle \nabla F(x_{\mathrm{i}}^{\nu}) + \nabla^2 F(x_{\mathrm{i}}^{\nu}) \Delta x_{\mathrm{i}}^{\nu}, \widehat{x} - x_{\mathrm{i}}^{\nu+} \right\rangle \\
\overset{(5.69)}{\geq} & \left\langle \nabla F(x_{\mathrm{i}}^{\nu+}) - \nabla F(x_{\mathrm{i}}^{\nu}) - \nabla^2 F(x_{\mathrm{i}}^{\nu}) \Delta x_{\mathrm{i}}^{\nu}, \widehat{x} - x_{\mathrm{i}}^{\nu+} \right\rangle + \frac{\mu}{2} \left\| x_{\mathrm{i}}^{\nu+} - \widehat{x} \right\|^2 \\
& + \left\langle \frac{M_{\mathrm{i}}}{2} ||\Delta x_{\mathrm{i}}^{\nu}|| \Delta x_{\mathrm{i}}^{\nu}, x_{\mathrm{i}}^{\nu+} - \widehat{x} \right\rangle + \left\langle \delta_{\mathrm{i}}^{\nu}, x_{\mathrm{i}}^{\nu+} - \widehat{x} \right\rangle + \left\langle \tilde{B}_{\mathrm{i}}^{\nu} \Delta x_{\mathrm{i}}^{\nu}, x_{\mathrm{i}}^{\nu+} - \widehat{x} \right\rangle \\
\geq & - \frac{1}{2\mu} \left\| \nabla F(x_{\mathrm{i}}^{\nu+}) - \nabla F(x_{\mathrm{i}}^{\nu}) - \nabla^2 F(x_{\mathrm{i}}^{\nu}) \Delta x_{\mathrm{i}}^{\nu} \right\|^2 \\
& + \left\langle \frac{M_{\mathrm{i}}}{2} ||\Delta x_{\mathrm{i}}^{\nu}|| \Delta x_{\mathrm{i}}^{\nu}, x_{\mathrm{i}}^{\nu+} - \widehat{x} \right\rangle + \left\langle \delta_{\mathrm{i}}^{\nu}, x_{\mathrm{i}}^{\nu+} - \widehat{x} \right\rangle + \left\langle \tilde{B}_{\mathrm{i}}^{\nu} \Delta x_{\mathrm{i}}^{\nu}, x_{\mathrm{i}}^{\nu+} - \widehat{x} \right\rangle,
\end{aligned}
\tag{5.70}
$$

and by applying Lemma 5.6.1 (cf. inequality (5.21)) to the first term on the RHS of (5.70) along with Cauchy-schwarz inequality, yield

$$
\begin{aligned}
& F(\widehat{x}) - F(x_{\mathrm{i}}^{\nu+}) \\
\geq & - \left(\frac{L^2}{8\mu} + \frac{M_{\mathrm{i}}}{4\epsilon_0} \right) \|\Delta x_{\mathrm{i}}^{\nu}\|^4 - \frac{M_{\mathrm{i}} \epsilon_0}{4} \left\| x_{\mathrm{i}}^{\nu+} - \widehat{x} \right\|^2 - \frac{1}{2\epsilon_1} \left\| \delta_{\mathrm{i}}^{\nu} \right\|^2 - \frac{\epsilon_1}{2} \left\| x_{\mathrm{i}}^{\nu+} - \widehat{x} \right\|^2 + \left\langle \tilde{B}_{\mathrm{i}}^{\nu} \Delta x_{\mathrm{i}}^{\nu}, x_{\mathrm{i}}^{\nu+} - \widehat{x} \right\rangle \\
\overset{(a)}{\geq} & - \left(\frac{L^2}{8\mu} + \frac{M_{\mathrm{i}}}{4\epsilon_0} \right) \|\Delta x_{\mathrm{i}}^{\nu}\|^4 - \left(\frac{M_{\mathrm{i}} \epsilon_0}{2\mu} + \frac{\epsilon_1}{\mu} \right) \left(F(x_{\mathrm{i}}^{\nu+}) - F(\widehat{x}) \right) - \frac{1}{2\epsilon_1} \left\| \delta_{\mathrm{i}}^{\nu} \right\|^2 + \left\langle \tilde{B}_{\mathrm{i}}^{\nu} \Delta x_{\mathrm{i}}^{\nu}, x_{\mathrm{i}}^{\nu+} - \widehat{x} \right\rangle,
\end{aligned}
\tag{5.71}
$$

for arbitrary $\epsilon_0, \epsilon_1 > 0$, where (a) is due to the μ-strongly convexity of F and optimality of $\widehat{x}$. By Assumption 5.1.4 and some algebraic manipulations, the last term on the RHS of (5.71) is lower-bounded as

$$
\begin{aligned}
\left\langle \Delta x_{\mathrm{i}}^{\nu}, x_{\mathrm{i}}^{\nu+} - \widehat{x} \right\rangle_{\tilde{B}_{\mathrm{i}}^{\nu}} \geq & - \frac{\beta + \tau_{\mathrm{i}}}{2\epsilon_2} \left\| \Delta x_{\mathrm{i}}^{\nu} \right\|^2 - \frac{\epsilon_2 (\beta + \tau_{\mathrm{i}})}{2} \left\| x_{\mathrm{i}}^{\nu+} - \widehat{x} \right\|^2 \\
\overset{(a)}{\geq} & - \frac{\beta + \tau_{\mathrm{i}}}{2\epsilon_2} \left\| \Delta x_{\mathrm{i}}^{\nu} \right\|^2 - \frac{\epsilon_2 (\beta + \tau_{\mathrm{i}})}{\mu} \left(F(x_{\mathrm{i}}^{\nu+}) - F(\widehat{x}) \right),
\end{aligned}
\tag{5.72}
$$

with arbitrary $\epsilon_2 > 0$, where (a) follows from the μ-strong convexity of F and optimality of $\widehat{x}$. Set

$$\epsilon_0 = \frac{\mu}{2M_{\max}}, \quad \epsilon_1 = \frac{\mu}{4}, \quad \epsilon_2 = \frac{\mu}{4(\beta+\tau_{\max})},$$

where $\tau_{\max} \triangleq \max_{\mathrm{i}\in[m]} \tau_{\mathrm{i}}$; then combining (5.71)-(5.72) and averaging over $\mathrm{i} = 1, \ldots, m$, lead to

$$\frac{1}{m}\sum_{\mathrm{i}=1}^{m}\left(F(x_{\mathrm{i}}^{\nu+}) - F(\widehat{x})\right) \leq \frac{L^2+4M_{\max}^2}{2m\mu}\sum_{\mathrm{i}=1}^{m}\|\Delta x_{\mathrm{i}}^{\nu}\|^4 + \frac{8\left(\beta+\tau_{\max}\right)^2}{m\mu}\sum_{\mathrm{i}=1}^{m}\left\|\Delta x_{\mathrm{i}}^{\nu}\right\|^2 + \frac{8}{m\mu}\left\|\delta^{\nu}\right\|^2. \tag{5.73}$$

The bound (5.68) is a direct consequence of (5.73), with $\tau_{\mathrm{i}} = 2\beta$, for all $\mathrm{i} = 1, \ldots, m$. □

5.6.5.2 Preliminary complexity results

Theorem 5.6.4. *Let Assumptions 5.1.2-5.1.4 and 5.1.5 hold. Let also $M_{\mathrm{i}} \geq L$ and $\tau_{\mathrm{i}} = 2\beta$, for all $\mathrm{i} = 1, \ldots, m$, and denote*

$$C_2 \triangleq \xi \cdot \frac{(M_{\max}+L)\sqrt{2m}}{3\mu^{3/2}}, \quad M_{\max} \triangleq \max_{\mathrm{i}\in[m]} M_{\mathrm{i}},$$

for some arbitrary $\xi \geq 1$. If a reference matrix $\overline{W}$ satisfying Assumption 5.2.1 is used in steps (5.8b)-(5.8c), with $\rho \triangleq \lambda_{\max}(\overline{W} - J) < 1$ and $K = \widetilde{\mathcal{O}}(1/\sqrt{1-\rho})$ (the explicit expression of K is given in (5.98)), then the sequence $\{p^{\nu}\}$ generated by DiRegINA satisfies the following:

(a) If

$$p^{\nu} \geq \underline{p}_1 \triangleq \frac{\mu^3}{2m(M_{\max}+L)^2\xi^2}\left(1+\frac{4\beta}{\mu}\right)^4,$$

then

$$(p^{\nu})^{1/4} \leq (p^0)^{1/4} - \frac{\nu}{12\sqrt{3C_2}}.$$

(b) Assume [exclusively in this case (b)] $\beta \leq \mu$ and denote

$$\tilde{p}^{\nu} \triangleq p^{\nu}/c^2, \quad c \triangleq \frac{\mu\sqrt{\mu}}{8\sqrt{m(L^2+4M_{\max}^2)}}, \quad \underline{p}_2 \triangleq \frac{2\cdot 12^4}{L^2+4M_{\max}^2}\cdot\frac{\beta^2\mu}{m}.$$

If $p^\nu \geq \underline{p}_2$ *and* $p^{\nu-1} \leq c^2$, *then* $\tilde{p}^\nu \leq (\tilde{p}^{\nu-1})^2$.

(c) If

$$p^\nu < \underline{p}_3 \triangleq \frac{9}{L^2 + 4M_{\max}^2} \cdot \frac{\beta^2 \mu}{m}, \tag{5.74}$$

then $\{p^\nu\}$ *converges Q-linearly to zero with rate*

$$\left(1 + \frac{\max(\beta, \mu)}{4mb_2}\right)^{-1} = \left(1 + \frac{1}{576} \cdot \frac{\mu \max(\beta, \mu)}{\beta^2}\right)^{-1}. \tag{5.75}$$

Proof. We organize the proof into three parts, **(a)-(c)**, in accordance with the three cases in the statement of the theorem.

(a) Recall Lemma 5.6.3 from the proof of Theorem 5.6.1:

$$F(x_{\mathrm{i}}^{\nu+}) \leq \widetilde{F}_{\mathrm{i}}(x_{\mathrm{i}}^{\nu+}; x_{\mathrm{i}}^\nu) + \frac{1}{2\epsilon}\left\|\delta_{\mathrm{i}}^\nu\right\|^2, \tag{5.76}$$

for arbitrary $\epsilon > 0$, where $M_{\mathrm{i}} \geq L$ and $\tau_{\mathrm{i}} \geq \beta + \epsilon$. In addition, by the upperbound approximation of $\widetilde{F}_{\mathrm{i}}(\cdot; x_{\mathrm{i}}^\nu)$ in (5.22), there holds

$$\widetilde{F}_{\mathrm{i}}(y; x_{\mathrm{i}}^\nu) \leq F(y) + \frac{1}{2}\left\|y - x_{\mathrm{i}}^\nu\right\|^2_{(\beta+\tau_{\mathrm{i}}+\epsilon)I} + \frac{M_{\mathrm{i}} + L}{6}\left\|y - x_{\mathrm{i}}^\nu\right\|^3 + \frac{1}{2\epsilon}\left\|\delta_{\mathrm{i}}^\nu\right\|^2, \quad \forall y \in \mathcal{K}. \tag{5.77}$$

Set $\tau_{\mathrm{i}} = 2\beta$ and $\epsilon = \beta$, then by (5.76)-(5.77) and $x_{\mathrm{i}}^{\nu+}$ being the minimizer of $\widetilde{F}(\cdot; x_{\mathrm{i}}^\nu)$,

$$\begin{aligned}
&F(x_{\mathrm{i}}^{\nu+}) - F(\widehat{x}) \\
\leq &\min_{y \in \mathcal{K}} \left\{ F(y) - F(\widehat{x}) + 2\beta\left\|y - x_{\mathrm{i}}^\nu\right\|^2 + \frac{M_{\mathrm{i}} + L}{6}\left\|y - x_{\mathrm{i}}^\nu\right\|^3 + \frac{1}{\beta}\left\|\delta_{\mathrm{i}}^\nu\right\|^2 \right\} \\
\leq &\min_{\alpha \in [0, \alpha_0]} \left\{ F(y) - F(\widehat{x}) + 2\beta\left\|y - x_{\mathrm{i}}^\nu\right\|^2 + \frac{M_{\mathrm{i}} + L}{6}\left\|y - x_{\mathrm{i}}^\nu\right\|^3 + \frac{1}{\beta}\left\|\delta_{\mathrm{i}}^\nu\right\|^2 : y = \alpha\widehat{x} + (1-\alpha)x_{\mathrm{i}}^\nu \right\} \\
\overset{(a)}{\leq} &\min_{\alpha \in [0, \alpha_0]} \left\{ (1-\alpha)\left(F(x_{\mathrm{i}}^\nu) - F(\widehat{x})\right) - \frac{\alpha(1-\alpha)\mu}{2}\left\|x_{\mathrm{i}}^\nu - \widehat{x}\right\|^2 \right. \\
&\qquad \left. + 2\beta\alpha^2\left\|\widehat{x} - x_{\mathrm{i}}^\nu\right\|^2 + \frac{M_{\mathrm{i}} + L}{6}\alpha^3\left\|\widehat{x} - x_{\mathrm{i}}^\nu\right\|^3 + \frac{1}{\beta}\left\|\delta_{\mathrm{i}}^\nu\right\|^2 \right\},
\end{aligned} \tag{5.78}$$

where (a) is due to the μ-strong convexity of F. If $\alpha_0 = 1/(1+4\beta/\mu)$, (5.78) implies

$$F(x_{\mathrm{i}}^{\nu+}) - F(\widehat{x}) \leq \min_{\alpha\in[0,\alpha_0]} \left\{(1-\alpha)\left(F(x_{\mathrm{i}}^{\nu}) - F(\widehat{x})\right) + \frac{M_{\mathrm{i}}+L}{6}\alpha^3\left\|\widehat{x} - x_{\mathrm{i}}^{\nu}\right\|^3 + \frac{1}{\beta}\left\|\delta_{\mathrm{i}}^{\nu}\right\|^2\right\},$$

where by the μ-strongly convexity of F and optimality of $\widehat{x}$, we also deduce

$$\begin{aligned} &F(x_{\mathrm{i}}^{\nu+}) - F(\widehat{x}) \\ &\leq \min_{\alpha\in[0,\alpha_0]} \left\{(1-\alpha)\left(F(x_{\mathrm{i}}^{\nu}) - F(\widehat{x})\right) + \frac{M_{\mathrm{i}}+L}{6}\alpha^3\left(\frac{2}{\mu}\left(F(x_{\mathrm{i}}^{\nu}) - F(\widehat{x})\right)\right)^{3/2} + \frac{1}{\beta}\left\|\delta_{\mathrm{i}}^{\nu}\right\|^2\right\}. \end{aligned} \tag{5.79}$$

Averaging (5.79) over $\mathrm{i} = 1, 2, \ldots, m$ while using (5.32), yields

$$p^{\nu+1} \leq \min_{\alpha\in[0,\alpha_0]} \left\{(1-\alpha)p^{\nu} + C_2\alpha^3\left(p^{\nu}\right)^{3/2} + \frac{1}{m\beta}\left\|\delta^{\nu}\right\|^2\right\}, \quad C_2 \triangleq \xi\cdot\frac{(M_{\max}+L)\sqrt{2m}}{3\mu^{3/2}}, \tag{5.80}$$

where $M_{\max} = \max_{\mathrm{i}\in[m]} M_{\mathrm{i}}$ and $\xi \geq 1$ is arbitrary.

Denote by α^* the minimizer of the RHS of (5.80); then if $p^{\nu} \geq \underline{p}_1 \triangleq 1/(9C_2^2\alpha_0^4)$, we have $\alpha^* = 1/\sqrt{3C_2\sqrt{p^{\nu}}}$, and

$$p^{\nu+1} \leq p^{\nu} - \frac{2(p^{\nu})^{3/4}}{3\sqrt{3C_2}} + \frac{1}{m\beta}\left\|\delta^{\nu}\right\|^2. \tag{5.81}$$

If

$$\frac{1}{m\beta}\left\|\delta^{\nu}\right\|^2 \leq \frac{1}{3\sqrt{3C_2}}(\underline{p}_1)^{3/4} \implies \frac{1}{m\beta}\left\|\delta^{\nu}\right\|^2 \leq \frac{1}{3\sqrt{3C_2}}(p^{\nu})^{3/4}, \tag{5.82}$$

(5.81) yields

$$p^{\nu+1} \leq p^{\nu} - \tilde{c}\,(p^{\nu})^{3/4}, \quad \forall\nu \geq 0, \quad \tilde{c} \triangleq \frac{1}{3\sqrt{3C_2}}. \tag{5.83}$$

Note that, by (5.46) and Lemma 5.6.4, condition (5.82) holds if

$$K \geq \frac{1}{\sqrt{1-\rho}}\cdot\frac{1}{2}\log\left(\frac{3\bar{D}_{\delta}\sqrt{3C_2}}{m\beta\underline{p}_1^{3/4}}\right). \tag{5.84}$$

We now prove by induction that (5.83) implies

$$(p^{\nu})^{1/4} \leq l_{\nu} \triangleq (p^0)^{1/4} - \frac{\tilde{c}}{4}\nu, \quad \forall\nu \geq 0. \tag{5.85}$$

Clearly, (5.85) holds for $\nu = 0$. Since the RHS of (5.83) is increasing (as a function of p^ν) when $p^\nu \geq (3\tilde{c}/4)^4 = 1/(9 \cdot 2^8 C_2^2)$ (which holds since $p^\nu \geq \underline{p}_1$), then $p^\nu \leq l_\nu^4$ implies

$$p^{\nu+1} \leq l_\nu^4 - \tilde{c} l_\nu^3,$$

which also implies $p^{\nu+1} \leq l_{\nu+1}^4$, as by definition of l^ν in (5.85),

$$l_\nu^4 - l_{\nu+1}^4 = (l_\nu - l_{\nu+1})(l_\nu + l_{\nu+1})\left(l_\nu^2 + l_{\nu+1}^2\right) = \frac{\tilde{c}}{4}(l_\nu + l_{\nu+1})\left(l_\nu^2 + l_{\nu+1}^2\right) \leq \tilde{c}\, l_\nu^3.$$

(b) Recall (5.41) (from the proof of Theorem 5.6.1), which under Assumptions 5.1.2-5.2.1 and condition (5.42), reads

$$wp^{\nu+1} + (\tilde{\delta}^{\nu+1})^2 \leq wp^\nu + c_w(\tilde{\delta}^\nu)^2 - \frac{w\mu}{4m}||\Delta x^\nu||^2. \tag{5.86}$$

Recall also Lemma 5.6.6 when condition C is satisfied, which together with (5.32), implies

$$p^{\nu+1} \leq b_1 \left(\sum_{\mathrm{i}=1}^m \left\|x_{\mathrm{i}}^{\nu+} - x_{\mathrm{i}}^\nu\right\|^2\right)^2 + \frac{8}{m\mu}\left\|\delta^\nu\right\|^2, \quad b_1 \triangleq \frac{L^2 + 4M_{\max}^2}{m\mu}. \tag{5.87}$$

Note that $p^{\nu+1} \geq \underline{p}_2$ implies that condition C in Lemma 5.6.6 holds, as proved next by contradiction. Suppose $p^{\nu+1} \geq \underline{p}_2$ but $||\Delta x^\nu|| < \omega_0$. Then Lemma 5.6.6 yields

$$\underline{p}_2 \leq p^{\nu+1} \overset{(5.32)}{\leq} p^{\nu+} < \frac{144\beta^2}{m\mu} \cdot \omega_0^2 + \frac{8}{m\mu}\left\|\delta^\nu\right\|^2 \overset{(a)}{\leq} \frac{2 \cdot 12^4}{L^2 + 4M_{\max}^2} \cdot \frac{\beta^4}{m\mu},$$

implying $\beta > \mu$, which is in contradiction with the assumption; note that (a) holds under (similar to derivation of (5.84))

$$K \geq \frac{1}{\sqrt{1-\rho}} \cdot \frac{1}{2} \log\left(\frac{\bar{D}_\delta}{18\beta^2\omega_0^2}\right) \implies \frac{8}{m\mu}\left\|\delta^\nu\right\|^2 \leq \frac{144\beta^2\omega_0^2}{m\mu}. \tag{5.88}$$

Now since $x \mapsto x^h$ is subadditive for $0 \leq h \leq 1$, i.e. $(a+b)^h \leq a^h + b^h$ for any $a, b \geq 0$, (5.87) together with (5.36) imply

$$-\sum_{\mathrm{i}=1}^{m} \|\Delta x_{\mathrm{i}}^{\nu}\|^2 \leq -\, b_1^{-\frac{1}{2}} \left(p^{\nu+1}\right)^{\frac{1}{2}} + \sqrt{\frac{8}{m\mu b_1}}\, \tilde{\delta}^{\nu}. \tag{5.89}$$

Combining (5.86) with (5.89) yields

$$wp^{\nu+1} + (\tilde{\delta}^{\nu+1})^2 \leq wp^{\nu} + c_w(\tilde{\delta}^{\nu})^2 - \frac{w\mu}{4m\sqrt{b_1}}\sqrt{p^{\nu+1}} + \frac{w\mu}{4m}\sqrt{\frac{8}{m\mu b_1}}\tilde{\delta}^{\nu},$$

and since $\tilde{\delta}^{\nu} \leq \sqrt{\xi^{\nu}} \leq \sqrt{D_2}, \forall \nu \geq 0$ (see the discussion in Subsec. 5.6.3.3, proof of Theorem 5.6.1), we get

$$wp^{\nu+1} + (\tilde{\delta}^{\nu+1})^2 \leq wp^{\nu} - \frac{w\mu}{4m\sqrt{b_1}}\sqrt{p^{\nu+1}} + C_3\tilde{\delta}^{\nu}, \quad C_3 \triangleq \left(c_w\sqrt{D_2} + \frac{c_w\beta\mu}{4m}\sqrt{\frac{8}{m\mu b_1}}\right). \tag{5.90}$$

Since $p^{\nu+1} \geq \underline{p}_2$, under (similar to derivation of (5.84))

$$K \geq \frac{1}{\sqrt{1-\rho}} \cdot \frac{1}{2}\log\left(\frac{64\bar{D}_{\delta}m^2 b_1 C_3^2}{c_w^2\beta^2\mu^2\underline{p}_2}\right) \implies C_3\tilde{\delta}^{\nu} \leq \frac{w\mu\sqrt{\underline{p}_2}}{8m\sqrt{b_1}}, \tag{5.91}$$

(5.90) yields

$$p^{\nu+1} + c\sqrt{p^{\nu+1}} \leq p^{\nu}, \quad c \triangleq \frac{\mu}{8m\sqrt{b_1}}.$$

Denote by $\tilde{p}^{\nu} \triangleq p^{\nu}/c^2$, then we get $\tilde{p}^{\nu+1} + \sqrt{\tilde{p}^{\nu+1}} \leq \tilde{p}^{\nu}$ which implies quadratic convergence when $p^{\nu+1} \geq \underline{p}_2$ and $\tilde{p}^{\nu} \leq 1 \equiv p^{\nu} \leq c^2$.

(c) Again recall (5.41):

$$wp^{\nu+1} + (\tilde{\delta}^{\nu+1})^2 \leq wp^{\nu} + c_w(\tilde{\delta}^{\nu})^2 - \frac{w\max(\beta,\mu)}{4m}||\Delta x^{\nu}||^2. \tag{5.92}$$

Invoking Lemma 5.6.6 under condition $\bar{\mathtt{C}}$ and $\tau_{\mathrm{i}} = 2\beta$, along with (5.32) and (5.36), we have

$$p^{\nu+1} \leq b_2 \sum_{\mathrm{i}=1}^{m} \left\|x_{\mathrm{i}}^{\nu+} - x_{\mathrm{i}}^{\nu}\right\|^2 + \frac{8}{m\mu}(\tilde{\delta}^{\nu})^2, \quad b_2 \triangleq \frac{144\beta^2}{m\mu}. \tag{5.93}$$

Combining (5.92) and (5.93) yields

$$w\left(1+\frac{\max(\beta,\mu)}{4mb_2}\right)p^{\nu+1}+(\tilde{\delta}^{\nu+1})^2 \leq wp^{\nu}+\left(c_w+\frac{2w\max(\beta,\mu)}{m^2\mu b_2}\right)(\tilde{\delta}^{\nu})^2, \tag{5.94}$$

where by choosing c_w to satisfy

$$\left(c_w+\frac{2w\max(\beta,\mu)}{m^2\mu b_2}\right)\leq\left(1+\frac{\max(\beta,\mu)}{4mb_2}\right)^{-1} \overset{(a)}{\equiv} c_w\leq\left(1+\frac{2\beta\max(\beta,\mu)}{m^2\mu b_2}\right)^{-1}\left(1+\frac{\max(\beta,\mu)}{4mb_2}\right)^{-1}, \tag{5.95}$$

[where (a) is due to $w=c_w\beta$ defined in Sec. 5.6.3.3], (5.94) becomes

$$w\left(1+\frac{\max(\beta,\mu)}{4mb_2}\right)p^{\nu+1}+(\tilde{\delta}^{\nu+1})^2 \leq wp^{\nu}+\left(1+\frac{\max(\beta,\mu)}{4mb_2}\right)^{-1}(\tilde{\delta}^{\nu})^2,$$

implying linear convergence of $\{\xi^\nu\}_\nu$ where

$$\zeta^\nu \triangleq w\left(1+\frac{\max(\beta,\mu)}{4mb_2}\right)p^\nu+(\tilde{\delta}^\nu)^2,$$

and decay rate

$$\left(1+\frac{\max(\beta,\mu)}{4mb_2}\right)^{-1}=\left(1+\frac{1}{576}\cdot\frac{\mu\max(\beta,\mu)}{\beta^2}\right)^{-1}. \tag{5.96}$$

Therefore, $\{p^\nu\}_\nu$ converges Q-linearly with rate (5.96).

Now let us derive (5.74) that defines this region. The goal is to identify the region where $\bar{\mathtt{C}}$ (cf. Lemma 5.6.6) holds. Under the condition (similar to derivation of (5.84))

$$K\geq\frac{1}{\sqrt{1-\rho}}\cdot\frac{1}{2}\log\left(\frac{4\bar{D}_\delta}{\beta\mu\omega_0^2}\right)\implies\frac{2(\tilde{\delta}^\nu)^2}{\beta\mu}\leq\frac{\omega_o^2}{2}, \tag{5.97}$$

and Lemma 5.6.5, there holds

$$\frac{1}{m}\sum_{\mathrm{i}=1}^{m}\left\|\Delta x_{\mathrm{i}}^\nu\right\|^2\leq\frac{8}{\mu}p^\nu+\frac{\omega_0^2}{2m},$$

which implies that $\bar{\mathtt{C}}$ is necessarily satisfied when

$$p^\nu<\frac{\omega_0^2\mu}{16m}=\frac{9}{L^2+4M_{\max}^2}\cdot\frac{\beta^2\mu}{m}.$$

Finally, unifying the conditions on K derived in (5.42). (5.84), (5.88), (5.91), (5.97), K must satisfy

$$K \geq \frac{1}{\sqrt{1-\rho}} \cdot \frac{1}{2} \log \left(\bar{D}_\delta \cdot \max \left\{ \frac{16}{\bar{D}_\delta c_w}, \frac{12^2 m Q_{\max}^2}{\bar{D}_\delta c_w \beta \max(\beta, \mu)}, \frac{3\sqrt{3C_2}}{m\beta \underline{p}_1^{3/4}}, \frac{1}{18\beta^2\omega_0^2}, \frac{64m^2 b_1 C_3^2}{c_w^2 \beta^2 \mu^2 \underline{p}_2}, \frac{4}{\beta\mu\omega_0^2} \right\} \right), \tag{5.98}$$

where recall that $c_w > 0$ must satisfy (5.95). □

5.6.5.3 Proof of Theorem 5.3.2

Let $M_{\mathrm{i}} = L$ for all $\mathrm{i} = 1, \ldots, m$, and set the free parameter $\xi \geq 1$ (defined in Theorem 5.6.4) to $\xi = 100\sqrt{5}$, and define the regions of convergence,

$$\begin{aligned}
&(\mathrm{R0}): \quad \Omega_0 \leq p^\nu, \\
&(\mathrm{R1}): \quad \Omega_1 \leq p^\nu < \Omega_0, \\
&(\mathrm{R2}): \quad \max(\varepsilon, \Omega_2) \leq p^\nu < \Omega_1, \\
&(\mathrm{R3}): \quad \varepsilon \leq p^\nu < \max(\varepsilon, \Omega_2),
\end{aligned}$$

where

$$\Omega_0 = 244 \cdot D^2 \mu, \quad \Omega_1 = c^2/2 = \frac{1}{640L^2} \cdot \frac{\mu^3}{m}, \quad \Omega_2 = \underline{p}_2 = \frac{2 \cdot 12^4}{5L^2} \cdot \frac{\beta^2 \mu}{m},$$

and c and $\underline{p}_2$ are defined in Theorem 5.6.4.

Using Theorem 5.6.2, region (R0) takes at most $\sqrt{\frac{LD}{\mu}}$ iterations. Now using Theorem 5.6.4, region (R1) lasts at most ν_1 iterations satisfying

$$(\Omega_1)^{1/4} \geq (\Omega_0)^{1/4} - \frac{\nu_1}{12\sqrt{3C_2}} \quad \Longleftarrow \quad \nu_1 \geq 480\sqrt{3\sqrt{5}} \cdot m^{1/4} \cdot \sqrt{\frac{LD}{\mu}}.$$

Let us conservatively consider scenarios $\Omega_1 \geq \varepsilon \geq \Omega_2$ and $\varepsilon < \Omega_2$, then the region of quadratic convergence (R2) lasts for at most

$$2\log\left(2\log\left(\min\left\{\frac{c^2}{\Omega_2}, \frac{c^2}{\varepsilon}\right\}\right)\right) \leq 2\log\left[2\log\left[\min\left\{\frac{1}{128 \cdot 12^4} \cdot \frac{\mu^2}{\beta^2}, \frac{\mu^3}{320mL^2} \cdot \frac{1}{\varepsilon}\right\}\right]\right] : \quad c^2 \geq \Omega_2, \varepsilon \leq c^2,$$

iterations. Note that conditions $p^\nu \geq \underline{p}_2$ and $p^\nu < \underline{p}_3$ in Theorem 5.6.4 are sufficient conditions identifying the region of quadratic and linear rate (or more specifically $\mathtt{C}$ and $\bar{\mathtt{C}}$ in Lemma 5.6.6); note that $\underline{p}_2$ and $\underline{p}_3$ are identical up to multiplying constants. Hence, to obtain a valid complexity of overall performance, we pessimistically associate the region of linear rate (R3) with $\varepsilon < p^\nu \leq \max(\varepsilon, \Omega_2)$ rather than $\varepsilon < p^\nu \leq \max(\varepsilon, \underline{p}_3)$; therefore, this region at most lasts for $O(\beta/\mu \cdot \log(\max(\varepsilon, \Omega_2)/\varepsilon))$ iterations. Thus, since the number of communications per iteration is $\tilde{O}\left(1/\sqrt{1-\rho}\right)$ [cf. (5.42), (5.64), (5.98) and note that $\varepsilon = \Omega_0$ in (5.64)], the overall complexity reads

$$\tilde{O}\left(\frac{1}{\sqrt{1-\rho}}\left\{\sqrt{\frac{LD}{\mu}}\left(1+m^{1/4}\right)+\log\left[\log\left[\frac{\mu^2}{\beta^2}\cdot\min\left\{1,\frac{\beta^2\mu}{mL^2}\cdot\frac{1}{\varepsilon}\right\}\right]\right]+\frac{\beta}{\mu}\log\left[\max\left(1,\frac{\beta^2\mu}{mL^2}\cdot\frac{1}{\varepsilon}\right)\right]\right\}\right)$$

communications.

5.6.5.4 The case of quadratic f_i in Theorem 5.3.2

Here we refine the proof of Theorem 5.3.2 to enhance the rate when $L = 0$:

Theorem 5.6.5. *Let Assumptions 5.1.2-5.1.5 hold with $L = 0$ and $\beta < \mu$. Denote by D_p an upperbound of p^0, i.e. $p^0 \leq D_p$ for all $\nu \geq 0$. Also choose $M_\mathrm{i} = \Theta(\mu^{3/2}/\sqrt{mD_p})$ sufficiently small (explicit condition is provided in (5.99)) and $\tau_\mathrm{i} = 2\beta$ for all $\mathrm{i} = 1, \ldots, m$. If a reference matrix $\overline{W}$ satisfying Assumption 5.2.1 is used in steps (5.8b)-(5.8c), with $\rho \triangleq \lambda_{\max}(\overline{W} - J) < 1$ and $K = \tilde{O}\left(1/\sqrt{1-\rho}\right)$ (explicit condition is provided in (5.98)), then for any given $\varepsilon > 0$, DiRegINA returns a solution with $p^\nu \leq \varepsilon$ after total*

$$\tilde{O}\left(\frac{1}{\sqrt{1-\rho}}\cdot\left\{\log\log\left(\frac{D_p}{\varepsilon}\right)+\frac{\beta}{\mu}\log\left(\frac{D_p\beta^2}{\mu^2\varepsilon}\right)\right\}\right)$$

communications. Note that when $\beta = O(1/\sqrt{n})$, $\epsilon = \Omega(V_N)$ and $n \geq m$, the above communication complexity reduces to

$$\tilde{O}\left(\frac{1}{\sqrt{1-\rho}}\cdot\left\{\log\log\left(\frac{D_p}{V_N}\right)\right\}\right).$$

Proof. Let us specialize the results established in Theorem 5.6.4 (in particular case (b)-(c)). Note that, since $L = 0$, we can impose $p^0 \leq c^2/2$ by a proper choice of M_{i}, allowing DiRegINA to circumvent the first region (associated with case (a) in Theorem 5.6.4) and start off in the quadratic rate region. Hence we only need to derive a sufficient condition for $p^0 \leq c^2/2$. Let us first consider case (b): if $M_{\mathrm{i}} = \Theta(\mu^{3/2}/\sqrt{mD_p}), \forall \mathrm{i}$, sufficiently small,

$$M_{\mathrm{i}} \leq \frac{\mu^{3/2}}{16\sqrt{2mD_p}}, \ \forall \mathrm{i} \implies p^0 \leq \frac{\mu^3}{512mM_{\max}^2} \implies p^0 \leq c^2/2, \tag{5.99}$$

where $M_{\max} \triangleq \max_{\mathrm{i}\in[m]} M_{\mathrm{i}}$. Let us also evaluate the precision achieved in case (b), i.e. $\underline{p}_2$: denote by $\underline{C}_M$ such that $M_{\mathrm{i}} \geq \underline{C}_M \mu^{3/2}/\sqrt{mD_p}, \forall \mathrm{i}$, then

$$\underline{p}_2 \triangleq \frac{12^4}{2M_{\max}^2} \cdot \frac{\beta^2\mu}{m} \leq \frac{12^4}{2\underline{C}_M^2} \cdot \frac{\beta^2 D_p}{\mu^2}.$$

Therefore the number of iterations to reach $\varepsilon = \Omega(\underline{p}_2)$ is $O(\log\log(c^2/\underline{p}_2)) = \log\log(D_p/\varepsilon)$, and since $K = \tilde{O}\left(1/\sqrt{1-\rho}\right)$, the total number of communication will be $\tilde{O}\left(1/\sqrt{1-\rho} \cdot \log\log(D_p/\varepsilon)\right)$.

Now let us derive the complexity when $\varepsilon = O(\underline{p}_2)$ (i.e. case (c) in Theorem 5.6.4). Setting $L = 0$ and following similar arguments, for arbitrary precision $\varepsilon > 0$, we obtain a communication complexity $\tilde{O}\left(1/\sqrt{1-\rho} \cdot \{\log\log(D_p/\varepsilon) + \beta/\mu \log(\beta^2 D_p/(\mu^2\varepsilon))\}\right)$. □

5.6.5.5 Proof of Corollary 5.3.3

Let us customize the rate established in Theorem 5.6.4 (in particular case (b)-(c)). We derive a sufficient condition for $p^0 \leq c^2/2$ which guarantees that the initial point is in the region of quadratic convergence. Using initialization policy (5.9), there holds $p^0 \leq C_\Delta/n$ for some $C_\Delta > 0$. Hence, under

$$n \geq \frac{640C_\Delta L^2}{\mu^3} \cdot m \implies p^0 \leq \frac{\mu^3}{640mL^2} \implies p^0 \leq c^2/2,$$

DiRegINA converges quadratically to the precision

$$\underline{p}_2 \triangleq \frac{2 \cdot 12^4}{5L^2} \cdot \frac{\beta^2\mu}{m}.$$

By $\beta = O(1/\sqrt{n})$, $\underline{p}_2 = O(V_N)$. Hence, since $K = \tilde{O}\left(1/\sqrt{1-\rho}\right)$, the total number of communication will be $\tilde{O}\left(1/\sqrt{1-\rho} \cdot \log\log(\mu^3/(mL^2V_N))\right)$.

5.6.6 Proof of Theorem 5.3.3

Let $M_{\mathrm{i}} = L$ for all $\mathrm{i} = 1, \ldots, m$, and set the free parameter $\xi = 50\beta/(3\mu)$ (defined in Theorem 5.6.4) and define the regions of convergence,

$$\begin{aligned}
(\overline{\mathrm{R0}}): &\quad \overline{\Omega}_0 \le p^\nu, \\
(\overline{\mathrm{R1}}): &\quad \overline{\Omega}_1 \le p^\nu < \overline{\Omega}_0, \\
(\overline{\mathrm{R2}}): &\quad \varepsilon \le p^\nu < \overline{\Omega}_1,
\end{aligned}$$

where

$$\overline{\Omega}_0 = 244 \cdot D^2\mu, \quad \overline{\Omega}_1 = \frac{0.9}{L^2} \cdot \frac{\beta^2\mu}{m}.$$

Using Theorem 5.6.2, region $(\overline{\mathrm{R0}})$ takes at most $\sqrt{\frac{LD}{\mu}}$ iteration; note that $\mu = \Omega(\beta^2)$ by assumption $n \ge m$, thus $\overline{\Omega}_0 = \Omega(\beta^2 \cdot 2LD^3)$. Now using Theorem 5.6.4, region $(\overline{\mathrm{R1}})$ lasts at most ν_1 iteration satisfying

$$\left(\overline{\Omega}_1\right)^{1/4} \ge (\overline{\Omega}_0)^{1/4} - \frac{\nu_1}{12\sqrt{3C_2}} \quad \Longleftarrow \quad \nu_1 \ge 240\sqrt{2} \cdot \frac{\sqrt{\beta LD\sqrt{m}}}{\mu}.$$

Finally, by case (c) in Theorem 5.6.4, region $(\overline{\mathrm{R2}})$ lasts for $O(\beta/\mu \cdot \log(\overline{\Omega}_1/\varepsilon))$. Thus, since communication cost per iteration is $\tilde{O}\left(1/\sqrt{1-\rho}\right)$ [cf. (5.42), (5.98)], the overall complexity is

$$\tilde{O}\left(\frac{1}{\sqrt{1-\rho}}\left\{\sqrt{\frac{LD}{\mu}}\left(1 + m^{1/4} \cdot \sqrt{\frac{\beta}{\mu}}\right) + \frac{\beta}{\mu}\log\left(\frac{\beta^2\mu}{mL^2} \cdot \frac{1}{\varepsilon}\right)\right\}\right).$$

5.6.7 The case of quadratic f_{i} in Theorem 5.3.3

Theorem 5.6.6. *Instate the setting of Theorem 5.3.3 where $L = 0$. Then, the total number of communications for DiRegINA to make $p^\nu \le \varepsilon$ reads*

$$\widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{1-\rho}} \cdot \frac{\beta}{\mu}\log\left(\frac{1}{\varepsilon}\right)\right).$$

When $\beta = O(1/\sqrt{n})$, $\epsilon = \Omega(V_N)$ *and* $n \geq m$, *the above communication complexity reduces to*

$$\widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{1-\rho}} \cdot m^{1/2} \cdot \log\left(\frac{1}{V_N}\right)\right).$$

Proof. We customize case **(c)** in Theorem 5.6.4, when $L = 0$. Note that $\bar{\mathtt{C}}$ in Lemma 5.6.6 holds for all $\nu \geq 0$ and condition (5.97) is no longer required. Therefore, the algorithm converges linearly with rate (5.75) and returns a solution within ε precision within $O\left(\beta/\mu \cdot \log(1/\varepsilon)\right)$ iterations and since $K = \tilde{O}\left(1/\sqrt{1-\rho}\right)$ [cf. (5.42)] , the total number of required communications is $\tilde{O}\left(1/\sqrt{1-\rho} \cdot \beta/\mu \cdot \log(1/\varepsilon)\right)$. □

REFERENCES

[1] F. Dörfler, J. W. Simpson-Porco, and F. Bullo, "Breaking the hierarchy: Distributed control and economic optimality in microgrids," *IEEE Transactions on Control of Network Systems*, vol. 3, no. 3, pp. 241–253, 2015.

[2] H. Qi, S. Iyengar, and K. Chakrabarty, "Distributed sensor networks—a review of recent research," *Journal of the Franklin Institute*, vol. 338, no. 6, pp. 655–668, 2001, Distributed Sensor Networks for Real-time Systems with Adaptive C onfiguration, ISSN: 0016-0032.

[3] H. Asama, T. Fukuda, T. Arai, and I. Endo, *Distributed autonomous robotic systems*. Springer Science & Business Media, 2013.

[4] T. Yang, X. Yi, J. Wu, Y. Yuan, D. Wu, Z. Meng, Y. Hong, H. Wang, Z. Lin, and K. H. Johansson, "A survey of distributed optimization," *Annual Reviews in Control*, vol. 47, pp. 278–305, 2019, ISSN: 1367-5788.

[5] M. Zhu and S. Martinez, *Distributed optimization-based control of multi-agent networks in complex environments*. Springer, 2015.

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

[7] J. Sun, Q. Qu, and J. Wright, "When are nonconvex problems not scary?" *arXiv preprint arXiv:1510.06096*, 2015.

[8] P. Jain and P. Kar, "Non-convex optimization for machine learning," *arXiv preprint arXiv:1712.07897*, 2017.

[9] R. Ge, F. Huang, C. Jin, and Y. Yuan, "Escaping from saddle points - online stochastic gradient for tensor decomposition," in *Proc. of the 28th Conf. on Learn. Theory*, P. Grünwald, E. Hazan, and S. Kale, Eds., ser. Proc. Mach. Learn. Res. Vol. 40, Paris, France: PMLR, Jul. 2015, pp. 797–842.

[10] Y. Zhang, "Distributed machine learning with communication constraints," Ph.D. dissertation, UC Berkeley, 2016.

[11] 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.

[12] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[13] Z. Qin, K. Scheinberg, and D. Goldfarb, "Efficient block-coordinate descent algorithms for the group lasso," *Mathematical Programming Computation*, vol. 5, pp. 143–169, 2 Jun. 2013.

[14] G.-X. Yuan, K.-W. Chang, C.-J. Hsieh, and C.-J. Lin, "A comparison of optimization methods and software for large-scale l1-regularized linear classification," *The Journal of Machine Learning Research*, vol. 11, pp. 3183–3234, 2010.

[15] K. Fountoulakis and J. Gondzio, "A Second-Order Method for Strongly Convex L1-Regularization Problems," *arXiv preprint arXiv:1306.5386*, 2013.

[16] I. Necoara and D. Clipici, "Efficient parallel coordinate descent algorithm for convex optimization problems with separable constraints: application to distributed MPC," *Journal of Process Control*, vol. 23, no. 3, pp. 243–253, Mar. 2013.

[17] Y. Nesterov, "Gradient methods for minimizing composite functions," *Mathematical Programming*, vol. 140, pp. 125–161, 1 Aug. 2013.

[18] P. Tseng and S. Yun, "A coordinate gradient descent method for nonsmooth separable minimization," *Mathematical Programming*, vol. 117, no. 1-2, pp. 387–423, Mar. 2009.

[19] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, Jan. 2009.

[20] S. J. Wright, R. D. Nowak, and M. A. Figueiredo, "Sparse reconstruction by separable approximation," *IEEE Trans. on Signal Processing*, vol. 57, no. 7, pp. 2479–2493, Jul. 2009.

[21] Z. Peng, M. Yan, and W. Yin, "Parallel and distributed sparse optimization," in *Signals, Systems and Computers, 2013 Asilomar Conference on*, IEEE, 2013, pp. 659–646.

[22] K. Slavakis and G. B. Giannakis, "Online dictionary learning from big data using accelerated stochastic approximation algorithms," in *Proc. of the IEEE 2014 International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2014)*, Florence, Italy, May 2014.

[23] K. Slavakis, G. B. Giannakis, and G. Mateos, "Modeling and optimization for big data analytics," *IEEE Signal Process. Mag.*, vol. 31, no. 5, pp. 18–31, Sep. 2014.

[24] M. De Santis, S. Lucidi, and F. Rinaldi, "A fast active set block coordinate descent algorithm for l_1-regularized least squares," *eprint arXiv:1403.1738*, Mar. 2014.

[25] S. Sra, S. Nowozin, and S. J. Wright, Eds., *Optimization for Machine Learning*, ser. Neural Information Processing. Cambridge, Massachusetts: The MIT Press, Sep. 2011.

[26] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, *Optimization with Sparsity-inducing Penalties*. Foundations and Trends® in Machine Learning, Now Publishers Inc, Dec. 2011.

[27] J. K. Bradley, A. Kyrola, D. Bickson, and C. Guestrin, "Parallel coordinate descent for l1-regularized loss minimization," in *Proc. of the 28th International Conference on Machine Learning*, Bellevue, WA, USA, Jun. 2011.

[28] M. Patriksson, "Cost approximation: A unified framework of descent algorithms for nonlinear programs," *SIAM Journal on Optimization*, vol. 8, no. 2, pp. 561–582, 1998.

[29] F. Facchinei, S. Sagratella, and G. Scutari, "Flexible parallel algorithms for big data optimization," in *Proc. of the IEEE 2014 International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2014)*, Florence, Italy, May 2014.

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

[31] O. Fercoq, Z. Qu, P. Richtárik, and M. Takáč, "Fast distributed coordinate descent for non-strongly convex losses," *arXiv preprint arXiv:1405.5300*, 2014.

[32] O. Fercoq and P. Richtárik, "Accelerated, parallel and proximal coordinate descent," *arXiv preprint arXiv:1312.5799*, 2013.

[33] Z. Lu and L. Xiao, "Randomized Block Coordinate Non-Monotone Gradient Method for a Class of Nonlinear Programming," *arXiv preprint arXiv:1306.5918v1*, 2013.

[34] I. Necoara and D. Clipici, "Distributed random coordinate descent method for composite minimization," *Technical Report*, pp. 1–41, Nov. 2013.

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

[36] P. Richtárik and M. Takáč, "Parallel coordinate descent methods for big data optimization," *arXiv preprint arXiv:1212.0873*, 2012.

[37] S. Shalev-Shwartz and A. Tewari, "Stochastic methods for $l1$-regularized loss minimization," *The Journal of Machine Learning Research*, pp. 1865–1892, 2011.

[38] Z. Lu and L. Xiao, "On the complexity analysis of randomized block-coordinate descent methods," *arXiv preprint arXiv:1305.4723*, 2013.

[39] I. Necoara and A. Patrascu, "A random coordinate descent algorithm for optimization problems with composite objective function and linear coupled constraints," *Computational Optimization and Applications*, vol. 57, no. 2, pp. 307–337, 2014.

[40] A. Patrascu and I.Necoara, "Efficient random coordinate descent algorithms for large-scale structured nonconvex optimization," *J. of Global Optimization*, pp. 1–23, Feb. 2014.

[41] P. Richtárik and M. Takáč, "Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function," *Mathematical Programming*, vol. 144, no. 1-2, pp. 1–38, 2014.

[42] I. Dassios, K. Fountoulakis, and J. Gondzio, "A second-order method for compressed sensing problems with coherent and redundant dictionaries," *arXiv preprint arXiv:1405.4146*, 2014.

[43] G.-X. Yuan, C.-H. Ho, and C.-J. Lin, "An improved glmnet for l_1-regularized logistic regression," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 1999–2030, 2012.

[44] Y. Yang, M. Pesavento, Z.-Q. Luo, and B. Ottersten, "Block successive convex approximation algorithms for nonsmooth nonconvex optimization," in *2019 53rd Asilomar Conference on Signals, Systems, and Computers*, IEEE, 2019, pp. 660–664.

[45] L. Cannelli, G. Scutari, F. Facchinei, and V. Kungurtsev, "Parallel asynchronous lock-free algorithms for nonconvex big-data optimization," in *2016 50th Asilomar Conference on Signals, Systems and Computers*, IEEE, 2016, pp. 1009–1013.

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

[47] C. Scherrer, A. Tewari, M. Halappanavar, and D. Haglin, "Feature clustering for accelerating parallel coordinate descent," in *Advances in Neural Information Processing Systems (NIPS2012)*, Curran Associates, Inc., 2012, pp. 28–36.

[48] A. Auslender, *Optimisation: méthodes numériques*. Masson, 1976.

[49] P. Tseng, "Convergence of a block coordinate descent method for nondifferentiable minimization," *Journal of optimization theory and applications*, vol. 109, no. 3, pp. 475–494, 2001.

[50] M. Razaviyayn, M. Hong, and Z.-Q. Luo, "A unified convergence analysis of block successive minimization methods for nonsmooth optimization," *SIAM J. on Opt.*, vol. 23, no. 2, pp. 1126–1153, 2013.

[51] M. Razaviyayn, M. Hong, Z.-Q. Luo, and J.-S. Pang, "Parallel successive convex approximation for nonsmooth nonconvex optimization," in *Proc. of the Annual Conference on Neural Information Processing Systems (NIPS)*, Montreal, Quebec, CA, to appear, Dec. 2014.

[52] M. Hong, M. Razaviyayn, and Z.-Q. Luo, "Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems," *arXiv preprint, arXiv:1410.1390*, Oct. 2014.

[53] J. T. Goodman, *Exponential priors for maximum entropy models*, US Patent 7,340,376, Mar. 2008.

[54] K.-W. Chang, C.-J. Hsieh, and C.-J. Lin, "Coordinate descent method for large-scale l2-loss linear support vector machines," *The Journal of Machine Learning Research*, vol. 9, pp. 1369–1398, 2008.

[55] R. Tappenden, P. Richtárik, and J. Gondzio, "Inexact coordinate descent: Complexity and preconditioning," *arXiv preprint arXiv:1304.5530*, 2013.

[56] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning*. Springer Series in Statistics. Springer, New York, 2009.

[57] H. A. Eiselt and V. Marianov, "Pioneering developments in location analysis," in *Foundations of Location Analysis, International Series in Operations Research & Management Science*, A. Eiselt and V. Marianov, Eds., Springer, 2011, ch. 11, pp. 3–22.

[58] A. Chambolle, "An algorithm for total variation minimization and applications," *Journal of Mathematical Imaging and Vision*, vol. 20, no. 1–2, pp. 89–97, Jan. 2004.

[59] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online Dictionary Learning for Sparse Coding," in *Proc. of the 26th International Conference on Machine Learning*, Montreal, Quebec, Canada, Jun. 2009.

[60] D. Goldfarb, S. Ma, and K. Scheinberg, "Fast alternating linearization methods for minimizing the sum of two convex functions," *Mathematical Programming*, vol. 141, pp. 349–382, 1-2 Oct. 2013.

[61] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*, 2nd. Athena Scientific Press, 1989.

[62] G. Scutari, F. Facchinei, P. Song, D. Palomar, and J.-S. Pang, "Decomposition by Partial linearization: Parallel optimization of multi-agent systems," *IEEE Trans. Signal Process.*, vol. 62, pp. 641–656, Feb. 2014.

[63] J.-S. Pang, "Error bounds in mathematical programming," *Mathematical Programming*, vol. 79, no. 1-3, pp. 299–332, 1997.

[64] P. Richtárik and M. Takáč, "On optimal probabilities in stochastic coordinate descent methods," *arXiv preprint arXiv:1310.3438*, 2013.

[65] P. Richtárik and M. Takáč, "Distributed coordinate descent method for learning with big data," *arXiv preprint arXiv:1310.2059*, 2013.

[66] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-Dynamic Programming*. Cambridge, Massachusetts: Athena Scientific Press, May 2011.

[67] R. Ge, J. D. Lee, and T. Ma, "Matrix completion has no spurious local minimum," in *Proc. of the 30th Intern. Conf. on Neural Inf. Process. Syst.*, 2016, pp. 2981–2989.

[68] K. Kawaguchi, "Deep learning without poor local minima," in *Adv. Neural Inf. Process. Syst.*, 2016, pp. 586–594.

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

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

[71] P. Di Lorenzo and G. Scutari, "Distributed nonconvex optimization over networks," in *IEEE Intern. Conf. on Comput. Adv. in Multi-Sensor Adapt. Process.*, 2015, pp. 229–232.

[72] P. Di Lorenzo and G. Scutari, "NEXT: In-network nonconvex optimization," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 2, no. 2, pp. 120–136, Jun. 2016.

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

[74] Y. Chi, Y. M. Lu, and Y. Chen, "Nonconvex optimization meets low-rank matrix factorization: An overview," *IEEE Trans. Signal Process.*, vol. 67, no. 20, pp. 5239–5269, 2019.

[75] A. Griewank, "The modification of newtons method for unconstrained optimization by bounding cubic terms," *Technical report NA/12*, 1981.

[76] Y. Nesterov and B. Polyak, "Cubic regularization of newton method and its global performance," *Math. Program.*, vol. 108, no. 1, pp. 177–205, Aug. 2006, ISSN: 1436-4646.

[77] C. Cartis, N. I. M. Gould, and P. L. Toint, "Adaptive cubic regularisation methods for unconstrained optimization. part i: Motivation, convergence and numerical results," *Math. Program.*, vol. 127, no. 2, pp. 245–295, Apr. 2011, ISSN: 1436-4646.

[78] C. Cartis, N. I. M. Gould, and P. L. Toint, "Adaptive cubic regularisation methods for unconstrained optimization. part ii: Worst-case function- and derivative-evaluation complexity," *Math. Program.*, vol. 130, no. 2, pp. 295–319, Dec. 2011, ISSN: 1436-4646.

[79] N. Agarwal, Z. Allen-Zhu, B. Bullins, E. Hazan, and T. Ma, "Finding approximate local minima faster than gradient descent," in *Proc. of the 49th Annual ACM SIGACT Symp. on Theory of Comp.*, ser. STOC 2017, Montreal, Canada: ACM, 2017, pp. 1195–1199, ISBN: 978-1-4503-4528-6.

[80] J. Moré and D. Sorensen, "Computing a trust region step," *SIAM J. Sci. Stat. Comp.*, vol. 4, no. 3, pp. 553–572, 1983.

[81] M. J. D. Powell, "On the global convergence of trust region algorithms for unconstrained minimization," *Math. Program.*, vol. 29, no. 3, pp. 297–303, Jul. 1984, ISSN: 1436-4646.

[82] F. E. Curtis, D. P. Robinson, and M. Samadi, "A trust region algorithm with a worst-case iteration complexity of $o(1/epsilon3/2)$ for nonconvex optimization," *Math. Program.*, vol. 162, no. 1, pp. 1–32, Mar. 2017, ISSN: 1436-4646.

[83] Y. N. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, and Y. Bengio, "Identifying and attacking the saddle point problem in high-dimensional non-convex optimization," in *Proceedings of the 27th Intern. Conf. on Neural Inf. Process. Syst.*, ser. NIPS'14, vol. 2, Montreal, Canada: MIT Press, 2014, pp. 2933–2941.

[84] Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford, "Accelerated methods for non-convex optimization," *SIAM J. Optim.*, vol. 2, no. 28, pp. 1751–1772, 2008.

[85] N. Agarwal, Z. Allen-Zhu, B. Bullins, E. Hazan, and T. Ma, "Finding approximate local minima faster than gradient descent," in *the 49th Annual ACM SIGACT Symp. on Theory of Comp. (STOC)*, Jun. 2017, pp. 1195–1199.

[86] Y. Carmon and J. C. Duchi, "Gradient descent efficiently finds the cubic-regularized non-convex newton step," *SIAM J. Optim.*, vol. 3, no. 29, pp. 2146–2178, 2019.

[87] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course (Applied Optimization)*. Springer, 2004, p. 254.

[88] J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht, "Gradient descent only converges to minimizers," in *Proc. of 29th Conf. on Learn. Theory*, vol. 49, 2016, pp. 1246–1257.

[89] J. D. Lee, I. Panageas, G. Piliouras, M. Simchowitz, M. I. Jordan, and B. Recht, "First-order methods almost always avoid strict saddle points," *Math. program.*, vol. 176, no. 1-2, pp. 311–337, 2019.

[90] M. O'Neill and S. J. Wright, "Behavior of accelerated gradient methods near critical points of nonconvex functions," *Math. Program.*, vol. 176, no. 1-2, pp. 403–427, 2019.

[91] Q. Li, Z. Zhu, and G. Tang, "Alternating minimizations converge to second-order optimal solutions," in *Proc. of the Int. Conf. on Mach. Learn. (ICML)*, 2019.

[92] M. Hong, J. D. Lee, and M. Razaviyayn, "Gradient primal-dual algorithm converges to second-order stationary solutions for nonconvex distributed optimization," *arXiv:1802.08941*, Feb. 2018.

[93] S. S. Du, C. Jin, J. D. Lee, M. I. Jordan, A. Singh, and B. Poczos, "Gradient descent can take exponential time to escape saddle points," in *Adv. Neural Inf. Process. Syst.*, 2017, pp. 1067–1077.

[94] R. Pemantle, "Nonconvergence to unstable points in urn models and stochastic approximations," *Ann. Probab.*, vol. 18, no. 2, pp. 698–712, 1990.

[95] C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan, "How to escape saddle points efficiently," in *Proc. of the 34th Intern. Conf. on Mach. Learn.*, ser. Proc. Mach. Learn. Res. Vol. 70, International Convention Centre, Sydney, Australia: PMLR, Jun. 2017, pp. 1724–1732.

[96] C. Jin, P. Netrapalli, and M. I. Jordan, "Accelerated gradient descent escapes saddle points faster than gradient descent," in *Proc. of Annual Conf. on Learn. Theory (COLT)*, 2018.

[97] S. Lu, M. Hong, and Z. Wang, "Accelerated gradient descent escapes saddle points faster than gradient descent," in *Proc. of the 36th Intern. Conf. on Mach. Learn.*, vol. 97, 2018, pp. 4134–4143.

[98] J. Zeng and W. Yin, "On nonconvex decentralized gradient descent," *IEEE Trans. on Signal Process.*, vol. 66, no. 11, pp. 2834–2848, Jun. 2018, ISSN: 1053-587X.

[99] P. Bianchi and J. Jakubowicz, "Convergence of a multi-agent projected stochastic gradient algorithm for non-convex optimization," *IEEE Trans. Autom. Control*, vol. 58, no. 2, pp. 391–405, Feb. 2013, ISSN: 0018-9286.

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

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

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

[103] Y. Sun, G. Scutari, and D. Palomar, "Distributed nonconvex multiagent optimization over time-varying networks," in *Proc. of the 50th Asilomar Conf. on Signals, Systems, and Computers*, Nov. 2016, pp. 788–794.

[104] G. Scutari and Y. Sun, "Distributed nonconvex constrained optimization over time-varying digraphs," *Math. Program.*, vol. 176, no. 1-2, pp. 497–544, 2019.

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

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

[107] S. Pu, W. Shi, J. Xu, and A. Nedich, "A push-pull gradient method for distributed optimization in networks," *arXiv:1803.07588v1*, Mar. 2018.

[108] M. Zhu and S. Martinez, "An Approximate Dual Subgradient Algorithm for Multi-Agent Non-Convex Optimization," *IEEE Trans. Autom. Control*, vol. 58, no. 6, pp. 1534–1539, 2013.

[109] M. Hong, D. Hajinezhad, and M. Zhao, "Prox-PDA: The proximal primal-dual algorithm for fast distributed nonconvex optimization and learning over networks," in *Proc. of the 34th Int. Conf. on Mach. Learn. (ICML 2017)*, vol. 70, 2017, pp. 1529–1538.

[110] D. Hajinezhad and M. Hong, "Perturbed proximal primal-dual algorithm for nonconvex nonsmooth optimization," *Math. Program., Series B*, pp. 1–38, 2019.

[111] K. Yuan, Q. Ling, and W. Yin, "On the convergence of decentralized gradient descent," *SIAM J. Optim.*, vol. 26, no. 3, pp. 1835–1854, 2016.

[112] S. B. Gelfand and S. K. Mitter, "Recursive stochastic algorithms for global optimization in r^d," *SIAM J. Control Opt.*, vol. 29, no. 5, pp. 999–1018, 1991.

[113] S. Łojasiewicz, "Une propriété topologique des sous-ensembles analytiques réels," in *Colloques internationaux*, Les Équations aux Dérivées Partielles (Paris, 1962), 1963, pp. 87–89.

[114] K. Kurdyka, "On gradients of functions definable in o-minimal structures," *Annales de l'institut Fourier*, vol. 48, no. 3, pp. 769–783, 1998.

[115] H. Attouch, J. Bolte, and B. F. Svaiter, "Convergence of descent methods for semi-algebraic and tame problems: Proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods," *Math. Program.*, vol. 137, no. 1, pp. 91–129, Feb. 2013, ISSN: 1436-4646.

[116] H. Attouch, J. Bolte, P. Redont, and A. Soubeyran, "Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the kurdyka-łojasiewicz inequality," *Math. Oper. Res.*, vol. 35, no. 2, pp. 438–457, 2010.

[117] S. Vlaski and A. H. Sayed, "Distributed learning in non-convex environments–part i: Agreement at a linear rate," *arXiv:1907.01848*, 2019.

[118] S. Vlaski and A. H. Sayed, "Distributed learning in non-convex environments–part ii: Polynomial escape from saddle-points," *arXiv:1907.01849*, 2019.

[119] C. Eckart and G. Young, "The approximation of one matrix by another of lower rank," *Psychometrika*, vol. 1, no. 3, pp. 211–218, Sep. 1936, ISSN: 1860-0980.

[120] S. Li, G. Tang, and M. B. Wakin, "The landscape of non-convex empirical risk with degenerate population risk," in *Adv. Neural Inf. Process. Syst.*, 2019, pp. 3502–3512.

[121] M. Dyrholm, C. Christoforou, and L. C. Parra, "Bilinear discriminant component analysis," *J. Mach. Learn. Res.*, vol. 8, pp. 1097–1111, Dec. 2007, ISSN: 1532-4435.

[122] P. Auer, M. Herbster, and M. K. Warmuth, "Exponentially many local minima for single neurons," in *Adv. Neural Inf. Process. Syst.*, 1996, pp. 316–322.

[123] L. Zhao, M. Mammadov, and J. Yearwood, "From convex to nonconvex: A loss function analysis for binary classification," in *2010 IEEE Int. Conf. on Data Mining Workshops*, IEEE, 2010, pp. 1281–1288.

[124] P. Halmos, *Measure Theory*, ser. Graduate Texts in Mathematics. Springer New York, 1976, ISBN: 9780387900889.

[125] G. Qu and N. Li, "Harnessing smoothness to accelerate distributed optimization," *IEEE Trans. Control Netw. Syst.*, vol. 5, no. 3, pp. 1245–1260, 2017.

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

[127] R. Xin, A. K. Sahu, U. A. Khan, and S. Kar, "Distributed stochastic optimization with gradient tracking over strongly-connected networks," *arXiv:1903.07266*, 2019.

[128] H. Attouch and J. Bolte, "On the convergence of the proximal algorithm for nonsmooth functions involving analytic features," *Math. Program.*, vol. 116, no. 1, pp. 5–16, Jan. 2009, ISSN: 1436-4646.

[129] S. Krantz and H. Parks, *A Primer of Real Analytic Functions*. Birkhäuser Boston, 2002.

[130] P. A. Absil, R. Mahony, and R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*. Princeton, NJ, USA: Princeton University Press, 2007, ISBN: 0691132984, 9780691132983.

[131] M. Shub, *Global stability of dynamical systems*. Springer-Verlag, 1987, ISBN: 9780387962955.

[132] P. A. Absil, R. Mahony, and J. Trumpf, "An extrinsic look at the riemannian hessian," in *Geom. Sci. Inf.*, Springer Berlin Heidelberg, 2013, pp. 361–368.

[133] R. A. Horn and C. R. Johnson, *Matrix Analysis*, 2nd. New York, NY, USA: Cambridge University Press, 2012.

[134] Y. Nesterov and B. Polyak, "Cubic regularization of newton method and its global performance," *Mathematical Programming*, vol. 108, pp. 177–205, 2006.

[135] A. Agafonov, P. Dvurechensky, G. Scutari, A. Gasnikov, D. Kamzolov, A. Lukashevich, and A. Daneshmand, "An accelerated second-order method for distributed stochastic optimization," *arXiv:2103.14392*, 2021.

[136] R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction," *Advances in neural information processing systems*, vol. 26, pp. 315–323, 2013.

[137] H. Hendrikx, F. Bach, and L. Massoulie, "An optimal algorithm for decentralized finite sum optimization," *arXiv:2005.10675*, 2020.

[138] R. Bhatia, *Matrix analysis*. Springer Science & Business Media, 2013, vol. 169.

[139] 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, Dec. 2015, pp. 1756–1764.

[140] 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.

[141] Y. Zhang and L. Xiao, "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.

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

[143] H. Lu, R. M. Freund, and Y. Nesterov, "Relatively smooth convex optimization by first-order methods, and applications," *SIAM J. on Optimization*, vol. 28, no. 1, pp. 333–354, 2020.

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

[145] G. Qu and N. Li, "Harnessing Smoothness to Accelerate Distributed Optimization," *IEEE Control Netw. Syst.*, vol. 5, no. 3, pp. 1245–1260, 2018.

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

[147] Q. Ling, W. Shi, G. Wu, and A. Ribeiro, "DLM: Decentralized linearized alternating direction method of multipliers," *IEEE Trans. Signal Process.*, vol. 63, no. 15, pp. 4051–4064, 2015.

[148] A. Mokhtari, W. Shi, Q. Ling, and A. Ribeiro, "Dqm: Decentralized quadratically approximated alternating direction method of multipliers," *IEEE Transactions on Signal Processing*, vol. 64, no. 19, pp. 5158–5173, 2016.

[149] 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.*, vol. 62, no. 7, pp. 1750–1761, 2014.

[150] D. Jakovetic, J. M. F. Moura, and J. Xavier, "Linear convergence rate of a class of distributed augmented lagrangian algorithms," *IEEE Trans. Autom. Control*, vol. 60, no. 4, pp. 922–936, Apr. 2015.

[151] Z. Li, W. Shi, and M. Yan, "A decentralized proximal-gradient method with network independent step-sizes and separated convergence rates," *IEEE Transactions on Signal Processing*, vol. 67, no. 17, pp. 4494–4506, 2019.

[152] D. Jakovetic, "A Unification and Generalization of Exact Distributed First-Order Methods," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 5, no. 1, pp. 31–46, 2019.

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

[154] C. Xi, V. S. Mai, R. Xin, E. H. Abed, and U. A. Khan, "Linear convergence in optimization over directed graphs with row-stochastic matrices," *IEEE Trans. Autom. Control*, vol. 63, no. 10, pp. 3558–3565, Oct. 2018.

[155] C. Xi and U. A. Khan, "A linear algorithm for optimization over directed graphs with geometric convergence," *IEEE Contr. Syst. Lett.*, vol. 2, no. 3, pp. 315–320, 2018.

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

[157] J. Zeng and W. Yin, "ExtraPush for convex smooth decentralized optimization over directed networks," *J. Comput. Math.*, vol. 35, no. 4, pp. 383–396, 2017.

[158] M. Maros and J. Jalden, "On the Q-linear convergence of Distributed Generalized ADMM under non-strongly convex function components," *IEEE Trans. Signal Inf. Process. Netw.*, vol. PP, no. 99, pp. 1–1, Jan. 2019.

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

[160] K. Yuan, B. Ying, X. Zhao, and A. H. Sayed, "Exact diffusion for distributed optimization and learning—part ii: Convergence analysis," *IEEE Transactions on Signal Processing*, vol. 67, no. 3, pp. 724–739, 2018.

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

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

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

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

[165] C. G. Lopes and A. H. Sayed, "Diffusion Least-Mean Squares Over Adaptive Networks: Formulation and Performance Analysis," *IEEE Trans. Signal Process.*, vol. 56, no. 7, pp. 3122–3136, 2008.

[166] K. Yuan, Q. Ling, and W. Yin, "On the Convergence of Decentralized Gradient Descent," *SIAM J. Optim.*, vol. 26, no. 3, pp. 1835–1854, Jan. 2016.

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

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

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

[170] P. Di Lorenzo and G. Scutari, "NEXT: In-network nonconvex optimization," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 2, no. 2, pp. 120–136, Jun. 2016.

[171] G. Qu and N. Li, "Accelerated distributed nesterov gradient descent for smooth and strongly convex functions," in *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Sep. 2016, pp. 209–216.

[172] C. Xi and U. A. Khan, "Add-opt: Accelerated distributed directed optimization," *IEEE Trans. Autom. Control*, vol. 63, no. 5, pp. 1329–1339, May 2018.

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

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

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

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

[177] J. Xu, S. Zhu, Y. C. Soh, and L. Xie, "Convergence of Asynchronous Distributed Gradient Methods Over Stochastic Networks," *IEEE Transactions on Automatic Control*, vol. 63, no. 2, pp. 434–448, 2018.

[178] W. Shi, Q. Ling, G. Wu, and W. Yin, "A proximal gradient algorithm for decentralized composite optimization," *IEEE Trans. Signal Process.*, vol. 63, no. 22, pp. 6013–6023, 2015.

[179] 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.*, vol. 59, no. 8, pp. 3889–3902, 2011.

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

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

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

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

[184] 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, Jun. 2009.

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

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

[187] G. Scutari, F. Facchinei, and L. Lampariello, "Parallel and distributed methods for constrained nonconvex optimization–Part I: Theory," *IEEE Trans. Signal Process.*, vol. 65, no. 8, pp. 1929–1944, Apr. 2017.

[188] 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, Apr. 2005, pp. 63–70.

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

[190] B. Gharesifard and J. Cortés, "When does a digraph admit a doubly stochastic adjacency matrix?" In *Proc. of the 2010 American Control Conference*, Jun. 2010, pp. 2440–2445. DOI: 10.1109/ACC.2010.5530578.

[191] A. Nedić and A. Ozdaglar, "Convergence rate for consensus with delays," *Journal of Global Optimization*, vol. 47, no. 3, pp. 437–456, 2010.

[192] Y. Tian, Y. Sun, and G. Scutari, "Achieving linear convergence in distributed asynchronous multi-agent optimization," *IEEE Trans. on Automatic Control*, to appear 2020.

[193] R. Bekkerman, M. Bilenko, and J. Langford, *Scaling up Machine Learning: Parallel and Distributed Approaches*. Cambridge University Press, 2011.

[194] X. Lian, C. Zhang, H. Zhang, C.-J. Hsieh, W. Zhang, and J. Liu., "Can decentralized algorithms outperform centralized algorithms? A case study for decentralized parallel stochastic gradient descent," in *Advances in Neural Information Processing Systems*, 2017, pp. 5330–5340.

[195] A. Mokhtari, H. Daneshmand, A. Lucchi, T. Hofmann, and A. Ribeiro, "Adaptive newton method for empirical risk minimization to statistical accuracy," in *Proceedings of the Advances in Neural Information Processing Systems*, 2016, pp. 4062–4070.

[196] L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," *SIAM Review*, vol. 60, no. 2, pp. 223–311, 2018.

[197] C. Ma and M. Takac, "Distributed inexact damped newton method: Data partitioning and work-balancing," in *Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[198] M. Jahani, X. He, C. Ma, A. Mokhtari, D. Mudigere, A. Ribeiro, and M. Takac, "Efficient distributed hessian free algorithm for large-scale empirical risk minimization via accumulating sample strategy," in *Twenty Third International Conference on Artificial Intelligence and Statistics*, 2020, pp. 2634–2644.

[199] S. Soori, K. Mishchenko, A. Mokhtari, M. M. Dehnavi, and M. Gurbuzbalaban, "Daveqn: A distributed averaged quasi-newton method with local superlinear convergence rate," in *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, vol. 108, 2020, pp. 1965–1976.

[200] C. A. Uribe and A. Jadbabaie, "A distributed cubic-regularized newton method for smooth convex optimization over networks," *arXiv:2007.03562*, 2020.

[201] J. Zhang, K. You, and T. Basar, "Distributed adaptive newton methods with globally superlinear convergence," *arXiv:2002.07378*, 2020.

[202] H. Hendrikx, L. Xiao, S. Bubeck, F. Bach, and L. Massoulie, "Statistically preconditioned accelerated gradient method for distributed optimization," in *Proceedings of the 37th International Conference on Machine Learning*, vol. 119, 13–18 Jul 2020, pp. 4203–4227.

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

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

[205] 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, 2017, pp. 3027–3036.

[206] G. Lan, S. Lee, and Y. Zhou, "Communication-efficient algorithms for decentralized and stochastic optimization," *Mathematical Programming*, pp. 1–48, 2017.

[207] C. A. Uribe, S. Lee, A. Gasnikov, and A. Nedić, "A dual approach for optimal algorithms in distributed optimization over networks," *Optimization Methods and Software*, pp. 1–40, 2020.

[208] A. Rogozin, V. Lukoshkin, A. Gasnikov, D. Kovalev, and E. Shulgin, "Towards accelerated rates for distributed optimization over time-varying networks," *arXiv:2009.11069*, 2020.

[209] E. Gorbunov, A. Rogozin, A. Beznosikov, D. Dvinskikh, and A. Gasnikov, "Recent theoretical advances in decentralized distributed convex optimization," *arXiv:2011.13259*, 2020.

[210] A. Jadbabaie, A. Ozdaglar, and M. Zargham, "A distributed newton method for network optimization," in *Proceedings of the 48h IEEE Conference on Decision and Control (CDC) held jointly with 2009 28th Chinese Control Conference*, 2009, pp. 2736–2741.

[211] E. Wei, A. Ozdaglar, and A. Jadbabaie, "A distributed newton method for network utility maximization—part ii: Convergence," *IEEE Transactions on Automatic Control*, vol. 58, no. 9, pp. 2176–2188, 2013.

[212] R. Tutunov, H. Bou-Ammar, and A. Jadbabaie, "Distributed newton method for large-scale consensus optimization," *IEEE Transactions on Automatic Control*, vol. 64, no. 10, pp. 3983–3994, 2019.

[213] A. Mokhtari, W. Shi, Q. Ling, and A. Ribeiro, "A decentralized second-order method with exact linear convergence rate for consensus optimization," *IEEE Transactions on Signal Processing*, vol. 2, no. 4, pp. 507–522, 2016.

[214] A. Mokhtari, Q. Ling, and A. Ribeiro, "Network newton distributed optimization methods," *IEEE Transactions on Signal Processing*, vol. 65, pp. 146–161, 1 2017.

[215] A. Mokhtari, W. Shi, Q. Ling, and A. Ribeiro, "DQM: Decentralized quadratically approximated alternating direction method of multipliers," *IEEE Transactions on Signal Processing*, vol. 64, no. 19, pp. 5158–5173, 2016.

[216] M. Eisen, A. Mokhtari, and A. Ribeiro, "A primal-dual quasi-newton method for exact consensus optimization," *IEEE Transactions on Signal Processing*, vol. 67, no. 23, pp. 5983–5997, 2019.

[217] Z. Jiaojiao, Q. Ling, and A. So, "A newton tracking algorithm with exact linear convergence rate for decentralized consensus optimization," in *59th IEEE Conference on Decision and Control (CDC)*, 2020.

[218] S. J. Reddi, J. Konečny, P. Richtárik, B. Póczós, and A. Smola, "Aide: Fast and communication efficient distributed optimization," *arXiv:1608.06879*, 2016.

[219] X.-T. Yuan and P. Li, "On convergence of distributed approximate newton methods: Globalization, sharper bounds and beyond," *arXiv:1908.02246*, 2019.

[220] S. Wang, F. Roosta-Khorasani, P. Xu, and M. W. Mahoney, "Giant: Globally improved approximate newton method for distributed optimization," in *Proceedings of the 32nd 32nd International Conference on Neural Information Processing Systems*, vol. 37, 2018, pp. 2338–2348.

[221] P. Dvurechensky, D. Kamzolov, A. Lukashevich, S. Lee, E. Ordentlich, C. A. Uribe, and A. Gasnikov, "Hyperfast second-order local solvers for efficient statistically preconditioned distributed optimization," *arXiv:2102.08246*, 2021.

[222] Y. Sun, A. Daneshmand, and G. Scutari, "Distributed optimization based on gradient-tracking revisited: Enhancing convergence rate via surrogation," *arXiv:1905.02637*, 2019.

[223] V. Vapnik, *The nature of statistical learning theory*. Springer science & business media, 2013.

[224] O. Bousquet, *Concentration inequalities and empirical processes theory applied to the analysis of learning algorithms*. PhD thesis, Ecole Polytechnique: Department of Applied Mathematics Paris, France, 2002.

[225] P. L. Bartlett, M. I. Jordan, and J. D. McAulffe, "Convexity, classification, and risk bounds," *Journal of the American Statistical Association*, vol. 101, no. 473, pp. 138–156, 2006.

[226] R. Frostig, R. Ge, S. M. Kakade, and A. Sidford, "Competing with the empirical risk minimizer in a single pass," in *Conference on learning theory (COLT)*, 2015, pp. 728–763.

[227] S.-S. Shai and S. Ben-David, *Understanding Machine Learning: From Theory to Algorihtms*. Cambridge University Press, 2014.

[228] K. Sridharan, S. Shalev-Shwartz, and N. Srebro, "Fast rates for regularized objectives," *Advances in neural information processing systems*, vol. 21, pp. 1545–1552, 2008.

[229] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan, "Learnability, stability and uniform convergence," *Journal of Machine Learning Research*, vol. 11, pp. 2635–2670, 2010.

[230] P. Di Lorenzo and G. Scutari, "NEXT: In-network nonconvex optimization," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 2, no. 2, pp. 120–136, Jun. 2016.

[231] A. Nedić, A. Olshevsky, and M. G. Rabbat, "Network topology and communication – computation tradeoffs in decentralized optimization," *Proceedings of the IEEE*, vol. 106, no. 5, pp. 953–976, 2018.

[232] R. Berthier, F. Bach, and P. Gaillard, "Accelerated gossip in networks of given dimension using jacobi polynomial iterations," *SIAM J. on Mathematics of Data Science*, vol. 1, pp. 24–47, 2 2020.

[233] F. Bach, "Self-concordant analysis for logistic regression," *Electronic Journal of Statistics*, vol. 4, pp. 384–414. 2010.

[234] Y. Nesterov, *Lectures on convex optimization*. Springer, 2018, vol. 137.

[235] T. Sun and Q. Tran-Dinh, "Generalized self-concordant functions: A recipe for newton-type methods," *Mathematical Programming*, vol. 178, pp. 145–213, 2019.

[236] D. Kovalev, A. Salim, and P. Richtárik, "Optimal and practical algorithms for smooth and strongly convex decentralized optimization," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[237] A. Mokhtari, Q. Ling, and A. Ribeiro, "Network newton distributed optimization methods," *IEEE Transactions on Signal Processing*, vol. 65, no. 1, pp. 146–161, 2016.

[238] G. W. Flake and S. Lawrence, "Efficient SVM regression training with SMO," *Machine Learning*, vol. 46, no. 1, pp. 271–290, 2002.

[239] L. Xiao, S. Boyd, and S.-J. Kim, "Distributed average consensus with least-mean-square deviation," *Journal of parallel and distributed computing*, vol. 67, no. 1, pp. 33–46, 2007.