NORTHWESTERN UNIVERSITY

Distributed Optimization Methods In Large-Scale Systems With Realistic Constraints

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Electrical Engineering

By

Charikleia Iakovidou

EVANSTON, ILLINOIS

September 2022

ABSTRACT

Distributed Optimization Methods In Large-Scale Systems With Realistic Constraints

Charikleia Iakovidou

Originally motivated by the emergence of networked systems lacking central coordination such as multiprocessors, wireless sensor networks and smart grids, the study of distributed optimization algorithms has been an active field of research spanning multiple decades. More recently, the rapid growth in the availability of high-dimensional datasets has posed the problem of learning efficiently and securely from data located across thousands of devices. In addition, distributed optimization for networks of mobile agents has been gaining significant traction over the last few years due to advancements in robotics and autonomous vehicles research. However, large-scale learning and mobile agent systems have inherent challenges that are typically overlooked in distributed optimization literature. This thesis aims to address some of these concerns.

In the first part of this thesis, we propose and analyze a first order distributed method (S-NEAR-DGD) which utilizes cost-efficient stochastic gradient approximations and can

tolerate inexact communication to alleviate the problems of excessive gradient computation costs and communication bottlenecks in large-scale Machine Learning. S-NEAR-DGD is based on a class of flexible deterministic algorithms (NEAR-DGD) that permit adjusting the amounts of communication and computation performed to best accommodate the application environment. Under strong convexity and Lipschitz gradient continuity, we show the linear convergence of S-NEAR-DGD to an error neighborhood of the optimal solution. Moreover, we provide numerical results demonstrating that S-NEAR-DGD is robust to types of inexact communication which may cause other state-of-the-art methods to diverge.

In the second part of this thesis, we consider the setting of nonconvex distributed optimization which features prominently in machine learning applications. Obtaining convergence guarantees is particularly challenging for nonconvex problems. Utilizing novel Lyapunov functions and under weaker assumptions compared to existing works on the same topic, we prove convergence of the iterates of the NEAR-DGD method to critical points. Moreover, we employ results stemming from dynamical system theory to demonstrate that NEAR-DGD almost always avoids strict saddle points and thus likely converges to minimizers. Our numerical results are promising and indicate that NEAR-DGD performs competitively against state-of-the-art methods.

In the last part of this thesis, we consider the multi-agent rendezvous problem, i.e. guiding of a group of agents to a common meeting point, with applications in multi-robot and multi-vehicle networks. We treat rendezvous as a distributed consensus optimization problem and develop a fully asynchronous algorithm that can handle any number of agents and spaces of any dimension, and which provably converges to an arbitrarily small

neighborhood of the optimal rendezvous point. Our method is robust to outdated information and to potentially erroneous displacements caused by the continuously moving nature of robotic agents.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my advisor, Professor Ermin Wei. I am thankful for her mentorship, encouragement and all the academic and professional opportunities she has provided me with over the years. Under her guidance I have grown both as a researcher and as a person.

I would like to extend my sincere thanks to my committee members, Professor Randall Berry and Professor Randy Freeman, for their support in both research and academic matters. Their feedback and suggestions have been invaluable in improving this thesis.

I would also like to thank the staff of the Electrical & Computer Engineering Department for always being eager to provide help and answer my questions.

Many thanks to my friends and family for being there for me; without their support this thesis would not have been completed. Finally, I would like to thank the instructors, staff and fellow students at my dojo, Thousand Waves; I am deeply grateful for having had the opportunity to train with them for the past three years.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

Beginning with the seminal works [**162, 163, 17**], the development and analysis of distributed optimization algorithms has been an active research area for over three decades. The need to harness the computing power of multiprocessors to solve increasingly complex problems and the emergence of a multitude of networked systems that lack central coordination such as wireless sensor networks [**5, 123, 51, 94, 122, 131, 173, 45, 76, 75, 59, 142, 44**], multi-robot and multi-vehicle networks [**12, 24, 117, 139, 187, 136, 28, 29, 188**] and power systems [**105, 61, 96, 83, 129, 120, 71, 78, 151, 174, 180, 56**], necessitated the design of optimization algorithms that can be implemented in a distributed manner. More recently, the proliferation of datasets coupled with storage constraints, growing computation costs and privacy concerns, has sparked significant interest in decentralized optimization for machine learning [**67, 137, 41, 21, 81, 178, 101, 147, 20, 91, 90**].

The diversity of the applications of distributed optimization makes a "one size fits all" approach unlikely to achieve optimal performance in every setting. Moreover, the distinct constraints inherent in different types of distributed systems are typically overlooked during algorithm design. The goal of this thesis is to develop and study efficient optimization methods that take into consideration the special requirements and limitations present in distributed systems and their emerging applications. For the rest of this work, we focus on the setting where the nodes of a *connected*, *undirected* network $\mathcal{G}(\mathcal{V}, \mathcal{E})$, with $\mathcal{V} = \{1, 2, ..., n\}$ denoting the set of nodes and $\mathcal{E} = \{(i, j) : i \sim j\}$ the set of edges,

collaborate to solve the following composite optimization problem

$$\min_{x\in\mathbb{R}^p} f(x) = \sum_{i=1}^{n} f_i(x), \tag{1.0.1}$$

where $x \in \mathbb{R}^p$ and $f_i : \mathbb{R}^p \to \mathbb{R}$.

Node $i \in \mathcal{V}$ has unique private access to the function component f_i. In addition, due to the absence of a shared memory, each node maintains a local copy $x_i \in \mathbb{R}^p$ of the global variable x. Problem 1.0.1 can then be reformulated to what is commonly referred to as the *consensus optimization problem* [**17**] in the literature

$$\begin{aligned} &\min_{\mathbf{x}\in\mathbb{R}^{np}} \mathbf{f}(\mathbf{x}) = \sum_{i=1}^{n} f_i(x_i) \\ &\text{s.t. } (\mathbf{W}\otimes I_p)\mathbf{x} = \mathbf{x}, \end{aligned} \tag{1.0.2}$$

where $\mathbf{x} = [x_1', x_2', ..., x_n']' \in \mathbb{R}^{np}$ is the column-wise concatenation of local variables x_i, $\mathbf{W} \in [0,1)^{np\times np}$ is a matrix constructed in such a way that the constraint in Problem 1.0.2 is satisfied iff $x_i = x_j$ for all pairs $(i,j) \in \mathcal{E}$, and I_p is the identity matrix of dimension p. We will be referring to $\mathbf{W}$ as the *consensus matrix* throughout this work.

Problems 1.0.1 and 1.0.2 are equivalent. However, unlike problem 1.0.1, the consensus problem is separable with respect to the variables $x_i \in \mathbb{R}^p$ and thus can be solved in a decentralized fashion. A model of distributed computation where each component of the decision vector is evaluated by a different processor was proposed as far back as in [**162, 163, 17**], while the first comprehensive analysis of a distributed (sub)gradient method for solving problem 1.0.2 in a distributed manner was published in [**111**]; starting from initial point $\mathbf{x}_0 = [(x_{1,0})', ..., (x_{n,0})']' \in \mathbb{R}^{np}$, the system iterates of Distributed (Sub)Gradient

Descent (DGD) [**111**], can be expressed as

$$\mathbf{x}_{k+1} = \mathbf{Z}\mathbf{x}_k - \alpha\nabla\mathbf{f}(\mathbf{x}_k), \tag{1.0.3}$$

where $\mathbf{Z} = \mathbf{W} \otimes I_p$, I_p is the identity matrix of dimension p, $\otimes$ denotes the Kronecker product operation, α is a positive steplength, $\nabla\mathbf{f}(\mathbf{x}_k) = [(\nabla f_1(x_{1,k}))', ..., (\nabla f_n(x_{n,k}))']'$ and $x_{i,k}$ the local decision variable at node i and iteration count k.

The building blocks of DGD and distributed optimization algorithms in general can be observed in (1.0.3); every iterate combines the *local optimization* of functions f_i (a gradient step in the case of DGD) with a *communication* or *consensus* step, where nodes update their local variables by forming weighted averages with those of their neighbors (term $\mathbf{Z}\mathbf{x}_k$ in (1.0.3)). Moreover, distributed optimization algorithms can be classified by the order in which computation and consensus are combined to produce an update; DGD is an example of an algorithm employing the Combine-Then-Adapt (CTA) strategy [**141**], where local iterates are first combined into a weighted average followed by an adaptation (computation) step. Conversely, the distributed gradient method known as diffusion [**35**] employs the Adapt-Then-Combine (ATC) strategy, where gradient steps are locally executed first and their results are combined in a weighted average. The system updates in this case can be written as

$$\mathbf{x}_{k+1} = \mathbf{Z}(\mathbf{x}_k - \alpha\nabla\mathbf{f}(\mathbf{x}_k)). \tag{1.0.4}$$

The NEAR-DGD method proposed in [**13**] generalizes (1.0.4) by combining a gradient step with an arbitrary number of nested consensus rounds in a single iteration of the

algorithm. The iterates of NEAR-DGD are given by

$$\mathbf{x}_k = \mathbf{Z}^{t(k)}\mathbf{y}_k \tag{1.0.5}$$

$$\mathbf{y}_{k+1} = \mathbf{x}_k - \alpha\nabla\mathbf{f}(\mathbf{x}_k), \tag{1.0.6}$$

where $\mathbf{Z}^{t(k)} = \mathbf{W}^{t(k)} \otimes I_p$ and $\{t(k)\}$ is a sequence defining the number of consensus rounds $t(k)$ executed at the k^{th} iteration of the algorithm. The strength of NEAR-DGD in comparison to other distributed optimization methods lies in its flexible structure, as the sequence $\{t(k)\}$ can be tuned on a case-by-case basis to best accommodate the underlying application and system properties.

The rest of the Introduction is organized as follows: in the next section we briefly summarize some of the main classes of distributed optimization algorithms. Section 1.2 concerns the problems of communication bottlenecks and excessive computation costs frequently arising in large-scale machine learning applications. In Section 1.3, we delve into the topic on nonconvex distributed optimization, which is prominent in modern large-scale machine learning systems and where deriving convergence guarantees is a challenging task. Finally, in Section 1.4, we focus on the problem of multi-agent rendezvous in robotics and investigate the particular type of asynchrony inherent in networks of mobile agents.

1.1. Summary of distributed optimization algorithms

Distributed optimization algorithms can be roughly divided into the following categories: (sub)gradient algorithms, primal dual methods, Newton-based methods and the Alternating Direction Method of Multipliers (ADMM). We summarize some of the highlights for each category in the paragraphs below.

- **(Sub)gradient algorithms**

  As with DGD, these methods rely on first order information to optimize local functions f_i. Examples include a projected consensus algorithm for constrained optimization, [**112**], distributed Nesterov gradient methods [**74**], diffusion algorithms where consensus is applied on local gradients instead of local variables [**35, 141**], distributed proximal gradient methods [**34**] and gradient methods with multiple nested consensus steps [**13**]. When functions f_i are convex, distributed first order methods generally converge to a neighborhood of the optimal solution of problem 1.0.2 with constant steplengths and require diminishing steplengths for exact convergence.

  A special mention should be made of a class of distributed first order algorithms sometimes referred to as "gradient tracking" (GT) methods [**113, 43, 146, 184, 130, 177**]. Gradient tracking methods maintain an additional variable that converges over time to the true descent direction of problem 1.0.1, i.e. instead of taking a step towards $\nabla f_i(x_i)$, node i progressively moves in the direction of $\nabla f(x_i) = \sum_{i=1}^{n} \nabla f_i(x_i)$. These methods have been shown to admit a primal-dual interpretation [**113, 176**] and are capable of achieving exact convergence to the solution of problem 1.0.2 with constant steplengths when the functions f_i are convex.

- **Primal dual methods**

  The observation that problem 1.0.2 is a nonlinear optimization problem with linear constraints led to the development of distributed primal-dual algorithms [**188, 79, 99, 18, 98, 73, 82**]. These methods aim to locate the saddle points

of a(n augmented) Langrangian function, such as

$$\mathcal{L}_\rho(\mathbf{x}, \mathbf{y}) = \mathbf{f}(\mathbf{x}) + \langle \mathbf{y}, \mathbf{A}\mathbf{x} \rangle + \frac{\rho}{2}\|\mathbf{x}\|_{\mathbf{A}}^2, \tag{1.1.1}$$

where $\mathbf{y}$ is the dual variable, $\mathbf{A}$ is matrix suitably chosen to enforce consensus (eg. $\mathbf{A} = I - \mathbf{Z}$) and ρ a tunable parameter (eg. $\rho = 0$ for non-augmented Lagrangian methods).

Primal-dual algorithms typically update both the primal variable $\mathbf{x}$ and the dual variable $\mathbf{y}$ at every iteration until convergence is reached. Like gradient tracking methods, distributed primal-dual algorithms provably achieve exact convergence with constant steplengths for convex objective functions.

- **Newton-based methods**

  The Newton method iterates for centralized optimization can be written as [**16**]

$$x_{k+1} = x_k - \alpha(\nabla^2 f(x_k))^{-1}\nabla f(x_k), \tag{1.1.2}$$

where $\nabla^2 f(x_k)$ is the Hessian matrix of f at x_k.

Newton methods typically enjoy superior convergence rates compared to gradient methods at the cost of having to compute the inverse of the second-order term $\nabla^2 f(x_k)$. A number of works employ second-order information in order to achieve fast convergence in the distributed setting, including primal-dual methods [**73, 169**] and distributed Newton methods that bypass the difficulty of calculating the Hessian inverse $(\nabla^2 \mathbf{f}(\mathbf{x}))^{-1}$ by replacing it with a suitable approximation [**103, 97, 52**].

- **The Alternating Direction Method of Multipliers (ADMM)**

  Consider the problem

$$\begin{aligned} &\min_{y,z} f(y) + g(z) \\ &\text{s.t. } Fy + Dz = c, \end{aligned} \tag{1.1.3}$$

  where $y \in \mathbb{R}^n$, $z \in \mathbb{R}^m$, $F \in \mathbb{R}^{p \times n}$, $D \in \mathbb{R}^{p \times m}$ and $c \in \mathbb{R}^p$.

  While the objective function of problem 1.1.3 is separable with respect to y and z, the linear constraint $Fy + Dz = c$ is coupled. The Alternating Direction Method of Multipliers [**23**] is a dual decomposition-based method that solves problems of the form of 1.1.3 using the augmented Lagrangian

$$\mathcal{L}_\rho(y, z, \lambda) = f(y) + g(z) - \langle \lambda, Fy + Dz - c \rangle + \frac{\rho}{2} \|Fy + Dz - c\|_2^2, \tag{1.1.4}$$

  where λ is the dual variable and ρ a tunable parameter.

  Each iteration of ADMM consists of three steps performed in a Gauss-Seidel-like order; a minimization of $\mathcal{L}_\rho$ with respect to of y, followed by a minimization with respect to z and an update of the dual variable λ to satisfy the optimality conditions of problem 1.1.3. Distributed variants of ADMM with convergence guarantees have been proposed in [**72, 107, 167, 168**].

- **Other methods**

  Instances include coordinate descent algorithms [**67, 137, 18, 60, 121**], flocking-based methods [**124**] and distributed adaptations of Nesterov's dual averaging algorithm [**51**].

In the next section, we focus on two major challenges in large-scale Machine Learning systems, namely communication bottlenecks and the increasing cost of gradient evaluations. After a summary of the existing literature, we outline our contributions on extending the NEAR-DGD method (1.0.5) (1.0.6) to accommodate these challenges.

1.2. Challenges in Distributed Large-Scale Machine Learning I: Computational & Communication Constraints

1.2.1. Literature review

1.2.1.1. Distributed optimization algorithms with quantized communication. The amount of communication between nodes has long been identified as a major performance bottleneck in decentralized computing, especially as the volume and dimensionality of available data increase [**135, 82**]. Limiting inter-node communication without overly sacrificing accuracy is an active research topic, and popular solutions include applying quantization techniques [**135, 3, 132**], which compress information so it can be transmitted with fewer bits, or sparsification methods [**166, 4**], which aim to decrease the number of transmitted bits while simultaneously enforcing vector sparsity (other approaches can be found in [**81, 165**]). While sparsification methods yield substantial gains in practice [**4**], they increase the variance of the information exchanged via the communication channel [**166**] and for this reason we do not investigate them in this work. Moreover, in any practical setting where the bandwidth of the communication channel is limited, the information exchanged cannot be represented by real-valued vectors with arbitrary precision. This limitation can prevent algorithms from converging to the true optimal value

of Problem (1.0.2) [**47, 46, 128, 10**]. Both of these concerns motivate us to design distributed optimization methods where nodes receive inexact/quantized information from their neighbors.

Multiple works in the literature focus on studying the effects of quantized communication on the convergence of distributed algorithms and on the design of methods robust to quantization error. An energy-based analysis of distributed incremental algorithms under quantized communication which characterized the dependency of the error induced by quantization on the quantization interval was first established in [**132**]. The convergence properties of DGD [**111**] when both communication and storage are quantized were later studied in [**110**], while [**88**] investigates the behavior of the same algorithm under deterministically quantized communication. The distributed dual averaging method [**51**] under both deterministic and probabilistic quantization of transmitted information was studied in [**182**]. Various works focus on distributed averaging (consensus) algorithms in conjunction with bandwidth-limited communication resulting from the application of deterministic quantization schemes [**109, 77, 102, 53**], probabilistic/randomized techniques [**10, 32**] and dynamic communication protocols [**54**].

A number of different approaches have been proposed to guide distributed algorithms with inexact communication towards optimality, such as using weighted averages of incoming quantized information and local estimates [**135, 46**], designing custom quantizers [**84, 47, 128**], employing encoding/decoding schemes [**3, 135, 179**], and utilizing error correction terms [**189, 84, 47**]. Among these, only [**128, 84**] achieve exact convergence with linear rate by employing dynamic quantizers which require the tuning of additional parameters and global information at initialization. Moreover, neither of these methods

allow for adjusting the amounts of computation and communication executed at every iteration.

1.2.1.2. Stochastic gradient in distributed optimization. Consider a supervised learning setup [**22**], where the goal is to construct a model of the relationship between an input variable X and an output variable Y from measured instances (x_k, y_k), $k = 1, ..., M$. Let f_w be a function parametrized by a vector w which serves as an estimate of the true mapping from X to Y. Moreover, let $l(\tilde{y}, y)$ be a *loss function* representing the cost of the model erroneously predicting $\tilde{y}$ instead of the true value y. The *empirical risk* $E_M(f_w)$ of the chosen model is the average value of the loss function across all M existing samples and is given by the expression

$$E_M(f_w) = \frac{1}{M} \sum_{k=1}^{M} l(f_w(x_k), y_k). \tag{1.2.1}$$

Minimizing empirical risk with respect to the parameter vector w is a basic machine learning tool; however, as is evident from Eq. (1.2.1), the computational cost of doing so scales unfavorably with the number of available samples M. As a consequence, the price of a gradient evaluation can become prohibitive in large-scale systems. This problem was partially mitigated by the introduction of Stochastic Gradient Descent (SGD) [**22, 190**] and mini-batching gradient algorithms [**80, 50, 89, 62, 11**], which rely on calculating an approximation of the true gradient at every iteration by appropriately subsampling the original dataset. Various studies and analyses on stochastic gradient based methods have been conducted in centralized settings [**22, 62**], federated learning (client-server model) [**101, 190**] and distributed settings over general topologies [**90, 106, 126**], which

is the setting this thesis adopts. Existing results indicate that general network topologies have potential advantages over client-server architectures [**90**]. Other works have demonstrated that certain distributed stochastic methods achieve a variance reduction effect similar to mini-batching [**125, 148, 127, 145**] (for more comparisons between distributed and centralized stochastic methods, we refer interested readers to [**106, 126**]). Finally, the authors of [**144**] conduct an extensive cost-benefit analysis of distributed stochastic algorithms but only for a limited number of network topologies.

There exists an extensive body of work on distributed stochastic optimization over general networks. Existing approaches include stochastic variants of DGD [**154, 149, 90**], stochastic diffusion algorithms [**160, 106, 127**], primal-dual methods [**33, 82, 65**], gradient-push algorithms [**114, 148**], the dual-averaging method [**51**], accelerated distributed algorithms [**57**] and stochastic distributed gradient tracking methods [**104, 125, 145, 175, 87**]. While some of these methods reach exact convergence (in expectation) with linear rates (eg. stochastic variants of gradient tracking, exact diffusion), they achieve this by utilizing variance reduction techniques that may have excessive memory requirements. Moreover, all of the aforementioned algorithms have a fixed structure and lack adaptability to diverse environments.

1.2.2. Contributions

In Chapter 2, we propose and analyze the Stochastic-NEAR-DGD (S-NEAR-DGD) method, a distributed algorithm which utilizes stochastic gradient approximations and can tolerate noisy communication to conserve bandwidth and computational resources. Our main contributions are summarized as follows:

(1) S-NEAR-DGD is based on a class of flexible algorithms (NEAR-DGD) [**13**] which permits adjusting the amounts of computation and communication performed during a run of the method. It is, thus, to the best of our knowledge, the only distributed method using stochastic gradient approximations and quantized communication that can be tailored on a case-by-case basis to balance convergence accuracy and total application cost in a diverse set of environments;

(2) We study various techniques for handling communication errors and investigate their effects on the convergence of distributed algorithms. We empirically demonstrate that Gradient Tracking (GT) methods may diverge in the presence of noise in the communication channel, unless the appropriate error correction is implemented. Conversely, purely primal methods such as S-NEAR-DGD appear to be more robust to noisy communication;

(3) We provide theoretical results which prove that S-NEAR-DGD converges to a neighborhood of the optimal solution with linear rate. Parts of the results have appeared in our previous works [**15**] and [**70**], where we considered deterministic quantization and stochastic gradient errors separately. We note that the communication errors considered in Chapter 2 are stochastic and include the ones in [**15**] as a special case. The stochastic nature calls for an error correction mechanism to prevent the communication errors from accumulating, which requires new analysis.

The next section is dedicated to another challenge arising in decentralized large-scale Machine Learning (ML) systems, which is closely related to the recent popularity of artificial Neural Networks (NN) given their superior performance in ML tasks. Namely,

we focus on the setting where the objective function $f : \mathbb{R}^p \to \mathbb{R}$ or Problem (1.0.1) is nonconvex, and where the obtainment of convergence guarantees can be particularly difficult. After reviewing the existing literature on this topic, we outline our contributions.

1.3. Challenges in Distributed Large-Scale Machine Learning II: Nonconvex Optimization

Problems 1.0.1 and 1.0.2 are common in large-scale decentralized machine learning [**67, 137, 21**], where the data is distributed over multiple networked computing units. Nonconvex objective functions feature prominently in machine learning applications, attracting significant interest in the development and analysis of distributed optimization methods for nonconvex problems [**31**]. The convergence of DGD when the function f of Problem1.0.1 is nonconvex has been studied in [**186**]. NEXT [**95**], SONATA [**143, 40**], xFilter [**152**] and MAGENTA [**68**], are some examples of distributed methods that utilize gradient tracking and can handle nonconvex objectives. Other approaches include primal-dual algorithms [**64, 66**] (we note that primal-dual and gradient tracking algorithms are equivalent in some cases [**113**]), the perturbed push-sum method [**159**], zeroth order methods [**63, 158**], and stochastic gradient algorithms [**19, 90, 157, 153**].

Providing second order guarantees when Hessian information is not available is a challenging task. As a result, the majority of the works listed in the previous paragraph establish convergence to critical points only. A recent line of research leverages existing results from dynamical systems theory and the structural properties of certain problems (which include matrix factorization, phase retrieval and dictionary learning, among others) to demonstrate that several centralized first order algorithms converge to minimizers

almost surely when initialized randomly [**85**]. Specifically, if the objective function satisfies the *strict saddle* property, namely, if all critical points are either strict saddles or minimizers, then many first order methods converge to saddles only if they are initialized in a low-dimensional manifold with measure zero. Using similar arguments, almost sure convergence to second order stationary points of Problem 1.0.2 is proven in [**40**] for DOGT, a gradient tracking algorithm for directed networks, and in [**66**] for the first order primal-dual algorithms GPDA and GADMM. The convergence of DGD with constant steplength to a neighborhood of the minimizers of Problem 1.0.2 is also shown in [**40**]. The conditions under which the Distributed Stochastic Gradient method (D-SGD), and Distributed Gradient Flow (DGF), a continuous-time approximation of DGD, avoid saddle points are studied in [**155**] and [**156**], respectively. Finally, the authors of [**159**] prove almost sure convergence to local minima under the assumption that the objective function has no saddle points.

Given the diversity of distributed systems in terms of computing power, connectivity and energy consumption, among other concerns, the ability to adjust the relative amounts of communication and computation on a case-by-case basis is a desirable attribute for a distributed optimization algorithm. While some existing methods are designed to minimize overall communication load (for instance, the authors of [**152**] employ Chebyshev polynomials to improve communication complexity), all of the methods listed above perform fixed amounts of computation and communication at every iteration and lack adaptability to heterogeneous environments.

1.3.1. Contributions

In Chapter 3, we extend the convergence analysis of the NEAR-DGD method, originally proposed in [**13**], from the strongly convex to the nonconvex setting. NEAR-DGD is a distributed first order method with a flexible framework, which allows for the exchange of computation with communication in order to reach a target accuracy level while simultaneously maintaining low overall application cost. We study two instances of NEAR-DGD: a variant performing a fixed number of consensus rounds at every iteration (NEAR-DGD^t), and a time-varying variant where the number of consensus rounds executed increases by one at every iteration (NEAR-DGD$^+$). We design custom Lyapunov functions which track the progress on Problem 1.0.1 and the distance to consensus for both cases, and demonstrate under weaker assumptions compared to similar works in the literature that NEAR-DGD^t converges to the set of critical points of our defined Lyapunov function and to approximate critical points of the function f of Problem 1.0.1, while NEAR-DGD$^+$ converges to the set of critical points of f. Moreover, we show that the gap between the limit points of NEAR-DGD^t and the critical points of f can become arbitrarily small by appropriate selection of algorithm parameters. Finally, we employ recent results based on dynamical systems theory to prove that both variants almost surely avoid strict saddles. Our analysis is shorter and simpler compared to other works due to the convenient form of our Lyapunov functions.

1.4. Challenges in Distributed Mobile Agent Systems: The Case of the Rendezvous Problem

The "rendezvous" problem, first studied in [**6**], concerns steering a group of robots to a common meeting point using exclusively local information, such as the positions or headings of other robots located within a sensing radius. Rendezvousing is an important component in a multitude of complex cooperative control applications, such as search and rescue [**30**], mine countermeasure [**181**], area exploration and monitoring [**138, 2**], refueling and recharging of unmanned ground vehicles (UGVs), unmanned aerial vehicles (UAVs) and autonomous underwater vehicles (AUVs) [**140, 100, 86**], target tracking [**116**], herding [**119**] and emergency evacuation [**39**], to name a few.

The authors of [**6**] proposed the so-called "circumcenter" algorithm to guide a group of robots capable of mutually observing the positions of other robots in their proximity to a rendezvous location on the 2-dimensional plane. Specifically, robots continuously and asynchronously execute cycles consisting of the following steps: *i*) observation of the positions of neighboring robots, *ii*) calculation of target positions based on most recent observations and *iii*) movement towards the target positions. The robots are restricted in the distance they can cover within one cycle, and are unable to modify their directions until a new cycle begins. The synchronous and asynchronous versions of the circumcenter algorithm, again for the 2-dimensional plane, were later elaborated on in [**92**] and [**93**], respectively. The authors of [**92, 93**] adopt a "stop and go" strategy composed of a "sensing" period where agents observe their surroundings, and a "maneuvering" period where agents move to a target point and then rest. Moreover, restrictions are imposed on the distances agents can cover and the total time they can travel within each phase. The

synchronous version of the circumcenter algorithm was studied under weaker assumptions in [**37**], while the authors of [**38**] proposed a class of synchronous circumcenter algorithms for switching topologies and arbitrary dimensional spaces that are robust to link failures and capable of maintaining edges between neighboring robots over time by restricting their motion range and defining a set of allowable graphs.

A different line of works leverages the connection between rendezvousing and what is known as "consensus" in the distributed computing literature [**162, 17, 163**]; namely the exchange of values between neighboring agents and the formation of weighted averages of local and neighbor values with the aim of reaching agreement between agents. Indeed, the rendezvous problem can be viewed as a consensus problem where agent positions serve as the local decision variables, a fact that has been previously noted in the literature [**118, 27**]. Without explicitly mentioning consensus, the authors of [**36**] propose an asynchronous Center-Of-Gravity (COG) algorithm for robotic systems, which similarly to [**6**], operate in "Look-Compute-Move" cycles corresponding to sensing the environment (positions of neighboring robots), computing the COG of the observed robots and moving towards the computed point in a straight line, respectively. Two distinct motion models are considered: an "undisturbed motion model" where robots always reach their destinations, and a "sudden stop model", where robots are guaranteed to move a minimum distance towards the goal point. The convergence of a class of asynchronous consensus processes which includes the sudden stop model of [**36**] as a special case was subsequently shown in [**58**] using paracontractions theory. In the same year, the authors of [**25**] proposed a consensus protocol utilizing column stochastic matrices that is capable of achieving rendezvous called "Consensus-Based Rendezvous" (CRB); they later provided

probabilistic convergence guarantees for CRB in the presence of stochastic noise with bounded variance in [**26**]. In [**171**], consensus schemes are employed to reach rendezvous under asynchronous and intermittent communication over switching and directed graphs; the robots operate in sensing-computing-moving cycles with deterministically bounded length, and do not rest or change direction within a cycle. The work [**108**] considers event-triggered rendezvous for two-wheeled robots under time-varying communication delays; however, the proposed controller relies on one-dimensional consensus protocols and hence each motion coordinate is updated separately at a different time.

Finally, a special mention should be made of the challenge of maintaining network connectivity while rendezvousing, due to the fact that the existence of edges usually depends on distance-based criteria. Graph maintenance is out of the scope of this thesis, but examples of approaches for tackling this issue can be found in [**27, 38, 150, 185, 48**].

1.4.1. Contributions

In Chapter 4, we propose a fully asynchronous algorithm for distributed multi-agent rendezvous and derive its convergence properties. Our work adopts a fundamentally different approach to the rendezvous problem with respect to existing works in the following ways:

(1) We wish to handle the simultaneous optimization of local position-dependent cost functions related to the background application alongside rendezvousing; hence, we model the rendezvous problem as a consensus optimization problem rather than as a consensus problem. Under the assumption of strong convexity, we define the optimal rendezvous point $x^\star$ in an arbitrary p-dimensional space to be

the solution of Problem 1.0.1, i.e.

$$x^\star = \arg\min_x f(x) = \sum_{i=1}^{n} f_i(x),$$

where $f_i : \mathbb{R}^p \to \mathbb{R}$ is the local cost function assigned to agent $i \in \mathcal{V}$.

Note that by setting $f_i = 0$, the consensus optimization problem , i.e. Problem 1.0.2 which is equivalent to Problem 1.0.1, reduces to the standard consensus problem. While there exists an extensive literature on asynchronous distributed optimization algorithms (eg. [**97, 133, 134, 64, 161, 72, 121**]), these works have not been developed for the mobile agent setting and do not provide convergence guarantees for the type of asynchrony we study.

(2) We model the asynchronous activations of agents as the arrivals of local independent Poisson processes, i.e. agent $i \in \mathcal{V}$ is associated to a local Poisson clock with parameter λ_i. The parameters λ_i can be different for each agent, thus allowing for heterogeneity in update frequencies.

(3) With regard to agent mobility, our setting is most similar to the one described in [**171**]. We assume that agents randomly alternate between two non-overlapping states: *i*) an "active" state, where they can perform computations, read and broadcast messages and adjust their velocities; and *ii*) a "inactive" state where they listen for messages from other agents which are stored in local buffers. These states roughly correspond to the "computing" and "moving" phases described in the previous section. In line with existing works on rendezvousing, our agents are unable to change their directions during inactive states. Moreover, they never become stationary, unless they explicitly set their velocities equal to zero.

Unlike [**171**], we do not impose deterministic bounds on the duration of inactive states and unlike [**93, 36**] we do not restrict agent motion range.

In addition to our novel approach, our algorithm enjoys the following properties:

(1) It provably converges to an arbitrarily small small neighborhood of the optimal rendezvous point while achieving approximate rendezvous;

(2) It is fully asynchronous and can handle outdated information.

We provide probabilistic convergence guarantees for our algorithm in Section 4.2 of Chapter 4. We conduct our analysis in discrete time; moreover, our analysis is simple and intuitive and does not rely on complex mathematical tools.

CHAPTER 2

S-NEAR-DGD: A Flexible Distributed Stochastic Gradient Method for Inexact Communication

2.1. Algorithm Development

2.1.1. Notation & Assumptions

In this Chapter, all vectors are column vectors. We use uppercase boldface letters for matrices. We will use the notation 1_n for the vector of ones of dimension n. The element in the i-th row and j-th column of a matrix $\mathbf{H}$ will be denoted by h_{ij}, and the p-th element of a vector v by $[v]_p$. The transpose of a vector v will be denoted by v^T. We will use $\|\cdot\|$ to denote the l_2-norm, i.e. for $v \in \mathbb{R}^p$ $\|v\| = \sqrt{\sum_{i=1}^p [v]_i^2}$, and $\langle v, u\rangle$ to denote the inner product of two vectors v, u. Finally, we define $\mathcal{N}_i$ to be the set of neighbors of node i, i.e., $\mathcal{N}_i = \{j \in \mathcal{V} : (i,j) \in \mathcal{E}\}$.

We adopt the following standard assumptions on the local functions f_i and the consensus matrix $\mathbf{W}$ of Problem 1.0.2.

Assumption 2.1.1. ***(Local Lipschitz gradients)*** *Each local objective function f_i has L_i-Lipschitz continuous gradients, i.e. $\|\nabla f_i(x) - \nabla f_i(y)\| \leq L_i\|x-y\|, \quad \forall x, y \in \mathbb{R}^p$.*

Assumption 2.1.2. ***(Local strong convexity)*** *Each local objective function f_i is μ_i-strongly convex, i.e. $f_i(y) \geq f_i(x) + \langle \nabla f_i(x), y - x\rangle + \frac{\mu_i}{2}\|x-y\|_2^2, \quad \forall x, y \in \mathbb{R}^p$.*

Assumption 2.1.3. ***(Consensus matrix)*** *The matrix* $\mathbf{W} \in \mathbb{R}^{n \times n}$ *has the following properties: i) symmetry, ii) double stochasticity, and iii)* $w_{i,j} > 0$ *iff* $(i,j) \in \mathcal{E}$ *or* $i = j$ *and* $w_{i,j} = 0$ *otherwise.*

Since $\mathbf{W}$ is symmetric it has n real eigenvalues, which we order by $\lambda_n \leq \lambda_{n-1} \leq ... \leq \lambda_2 \leq \lambda_1$ in ascending order. Assumption 2.1.3 implies that $\lambda_1 = 1$ and $\lambda_2 < \lambda_1$ for any connected network. The remaining eigenvalues have absolute values strictly less than 1, i.e., $-1 < \lambda_n$. Moreover, the equality $(\mathbf{W} \otimes I_p)\,\mathbf{x} = \mathbf{x}$ holds if and only if $x_i = x_j$ for all $(i,j) \in \mathcal{E}$ [**111**]. For the rest of this chapter, we will refer to the absolute value of the eigenvalue with the second largest absolute value of $\mathbf{W}$ as β, i.e. $\beta = \max\{|\lambda_2|, |\lambda_n|\}$.

2.1.2. The S-NEAR-DGD method

To accommodate bandwidth-limited communication channel, we assume that whenever node $i \in \{1, ..., n\}$ needs to communicate a vector $v \in \mathbb{R}^p$ to its neighbors, it sends an approximate vector $\mathcal{T}_c[v]$ instead, i.e., $\mathcal{T}_c[\cdot]$ is a randomized operator which modifies the input vector to reduce the bandwidth. Similarly, to model the availability of only inexact gradient information, we assume that instead of the true local gradient $\nabla f_i(x_i)$, node i calculates an approximation $\mathcal{T}_g[\nabla f_i(x_i)]$, where $\mathcal{T}_g[\cdot]$ is a randomized operator denoting the inexact computation. We refer to this method with inexact communication and gradient computation as the Stochastic-NEAR-DGD (S-NEAR-DGD) method.

Each node $i \in \{1, ..., n\}$ initializes and preserves the local variables $x_{i,k}^j$ and $y_{i,k}$. At iteration k of S-NEAR-DGD, node i calculates the stochastic gradient approximation $g_{i,k-1} = \mathcal{T}_g\left[\nabla f_i\left(x_{i,k-1}^{t(k-1)}\right)\right]$ and uses it to take a local gradient step and update its internal variable $y_{i,k}$. Next, it sets $x_{i,k}^0 = y_{i,k}$ and performs $t(k)$ nested consensus rounds, where

Algorithm 1: S-NEAR-DGD at node i

Initialization: Pick $x_{i,0}^{t(0)} = y_{i,0}$
for $k = 1, 2, ...$ **do**
  Compute $g_{i,k-1} = \mathcal{T}_g\left[\nabla f_i\left(x_{i,k-1}^{t(k-1)}\right)\right]$
  Update $y_{i,k} \leftarrow x_{i,k-1}^{t(k-1)} - \alpha g_{i,k-1}$
  Set $x_{i,k}^0 = y_{i,k}$
  for $j = 1, ..., t(k)$ **do**
    Send $q_{i,k}^j = \mathcal{T}_c\left[x_{i,k}^{j-1}\right]$ to neighbors $l \in \mathcal{N}_i$ and receive $q_{l,k}^j$
    Update $x_{i,k}^j \leftarrow \sum_{l=1}^n \left(w_{il} q_{l,k}^j\right) + \left(x_{i,k}^{j-1} - q_{i,k}^j\right)$
  end
end

during each consensus round $j \in \{1, ..., t(k)\}$ it constructs the bandwidth-efficient vector $q_{i,k}^j = \mathcal{T}_c\left[x_{i,k}^{j-1}\right]$, forwards it to its neighboring nodes $l \in \mathcal{N}_i$ and receives the vectors $q_{l,k}^j$ from neighbors. Finally, during each consensus round, node i updates its local variable $x_{i,k}^j$ by forming a weighted average of the vectors $q_{l,k}^j$, $l = 1, ..., n$ and adding the residual error correction term $\left(x_{i,k}^{j-1} - q_{i,k}^j\right)$. The entire procedure is presented in Algorithm 1.

Let $\mathbf{x}_0^{t(0)} = \mathbf{y}_0 = [y_{1,0}; ...; y_{n,0}]$ be the concatenation of local initial points $y_{i,0}$ at nodes $i = 1, ..., n$ as defined in Algorithm 1. The system-wide iterates of S-NEAR-DGD at iteration count k and j-th consensus round can be written compactly as,

$$\mathbf{y}_k = \mathbf{x}_{k-1}^{t(k-1)} - \alpha \mathbf{g}_{k-1}, \tag{2.1.1a}$$

$$\mathbf{x}_k^j = \mathbf{x}_k^{j-1} + (\mathbf{Z} - I_{np})\, \mathbf{q}_k^j, \quad j = 1, ..., t(k), \tag{2.1.1b}$$

where $\mathbf{x}_k^0 = \mathbf{y}_k$, $\mathbf{Z} = (\mathbf{W} \otimes I_p) \in \mathbb{R}^{np \times np}$, $g_{i,k-1} = \mathcal{T}_g\left[\nabla f_i\left(x_{i,k-1}^{t(k-1)}\right)\right]$, $q_{i,k}^j = \mathcal{T}_c\left[x_{i,k}^{j-1}\right]$ for $j = 1, ..., t(k)$ and $\mathbf{g}_{k-1}$ and $\mathbf{q}_k^j$ are the long vectors formed by concatenating $g_{i,k-1}$ and $q_{i,k}^j$ over i respectively.

Moreover, due to the double stochasticity of $\mathbf{W}$, the following relations hold for the average iterates $\bar{y}_k = \frac{1}{n}\sum_{i=1}^n y_{i,k}$ and $\bar{x}_k^j = \frac{1}{n}\sum_{i=1}^n x_{i,k}^j$ for all k and j,

$$\bar{y}_k = \bar{x}_{k-1}^{t(k-1)} - \alpha \bar{g}_{k-1}, \tag{2.1.2a}$$

$$\bar{x}_k^j = \bar{x}_k^{j-1}, \quad j = 1, ..., t(k), \tag{2.1.2b}$$

where $\bar{g}_{k-1} = \frac{1}{n}\sum_{i=1}^n g_{i,k-1}$.

The operators $\mathcal{T}_c[\cdot]$ and $\mathcal{T}_g[\cdot]$ can be interpreted as $\mathcal{T}_c\left[x_{i,k}^{j-1}\right] = x_{i,k}^{j-1} + \epsilon_{i,k}^j$, and $\mathcal{T}_g\left[\nabla f_i\left(x_{i,k-1}^{t(k-1)}\right)\right] = \nabla f_i\left(x_{i,k-1}^{t(k-1)}\right) + \zeta_{i,k}$, where $\epsilon_{i,k}^j$ and $\zeta_{i,k}$ are random error vectors. We list our assumptions on these vectors and the operators $\mathcal{T}_c[\cdot]$ and $\mathcal{T}_g[\cdot]$ below.

Assumption 2.1.4. ***(Properties of*** $\mathcal{T}_c[\cdot]$***)*** *The operator* $\mathcal{T}_c[\cdot]$ *is iid for all* $i = 1, ..., n$, $j = 1, ..., t(k)$ *and* $k \geq 1$. *Moreover, the errors* $\epsilon_{i,k}^j = \mathcal{T}_c\left[x_{i,k}^{j-1}\right] - x_{i,k}^{j-1}$ *have zero mean and bounded variance for all* $i = 1, ..., n$, $j = 1, ..., t(k)$ *and* $k \geq 1$, *i.e.,*

$$\mathbb{E}_{\mathcal{T}_c}\left[\epsilon_{i,k}^j \middle| x_{i,k}^{j-1}\right] = 0, \quad \mathbb{E}_{\mathcal{T}_c}\left[\|\epsilon_{i,k}^j\|^2 \middle| x_{i,k}^{j-1}\right] \leq \sigma_c^2,$$

where σ_c *is a positive constant and the expectation is taken over the randomness of* $\mathcal{T}_c$.

Example 1. ***(Probabilistic quantizer)***

An example of an operator satisfying Assumption 2.1.4 is the probabilistic quantizer in [**182**]*, defined as follows: for a scalar* $x \in \mathbb{R}$, *its quantized value* $\mathcal{Q}[x]$ *is given by*

$$\mathcal{Q}[x] = \begin{cases} \lfloor x \rfloor & \text{with probability } (\lceil x \rceil - x)\,\Delta \\ \lceil x \rceil & \text{with probability } (x - \lfloor x \rfloor)\,\Delta, \end{cases}$$

where $\lfloor x \rfloor$ *and* $\lceil x \rceil$ *denote the operations of rounding down and up to the nearest integer multiple of* $1/\Delta$*, respectively, and* Δ *is a positive integer.*

It is shown in [**182**] that $\mathbb{E}\Big[x - \mathcal{Q}\left[x\right]\Big] = 0$ and $\mathbb{E}\left[\left|x - \mathcal{Q}\left[x\right]\right|^2\right] \leq \frac{1}{4\Delta^2}$. For any vector $v = [v_i]_{i=\{1,...,p\}}$ in $\mathbb{R}^p$, we can then apply the operator $\mathcal{Q}$ element-wise to obtain $\mathcal{T}_c\left[v\right] = \Big[\mathcal{Q}\left[v_i\right]\Big]_{i=\{1,...,p\}}$ in $\mathbb{R}^p$ with $\mathbb{E}_{\mathcal{T}_c}\left[v - \mathcal{T}_c\left[v\right] \big| v\right] = \mathbf{0}$ and $\mathbb{E}_{\mathcal{T}_c}\left[\left\|v - \mathcal{T}_c\left[v\right]\right\|^2 \big| v\right] \leq \frac{p}{4\Delta^2} = \sigma_c^2$.

Assumption 2.1.5. ***(Properties of*** $\mathcal{T}_g\left[\cdot\right]$***)*** *The operator* $\mathcal{T}_g\left[\cdot\right]$ *is iid for all* $i = 1, ..., n$ *and* $k \geq 1$*. Moreover, the errors* $\zeta_{i,k} = \mathcal{T}_g\left[\nabla f_i\left(x_{i,k-1}^{t(k-1)}\right)\right] - \nabla f_i\left(x_{i,k-1}^{t(k-1)}\right)$ *have zero mean and bounded variance for all* $i = 1, ..., n$ *and* $k \geq 1$*,*

$$\mathbb{E}_{\mathcal{T}_g}\left[\zeta_{i,k} \Big| x_{i,k-1}^{t(k-1)}\right] = 0, \quad \mathbb{E}_{\mathcal{T}_g}\left[\left\|\zeta_{i,k}\right\|^2 \Big| x_{i,k-1}^{t(k-1)}\right] \leq \sigma_g^2,$$

where σ_g *is a positive constant and the expectation is taken over the randomness of* $\mathcal{T}_g$*.*

Assumption 2.1.5 is standard in the analysis of distributed stochastic gradient methods [**154, 114, 125, 90**].

We make one final assumption on the independence of the operators $\mathcal{T}_c\left[\cdot\right]$ and $\mathcal{T}_g\left[\cdot\right]$, namely that the process of generating stochastic gradient approximations does not affect the process of random quantization and vice versa.

Assumption 2.1.6. ***(Independence)*** *The operators* $\mathcal{T}_g\left[\cdot\right]$ *and* $\mathcal{T}_c\left[\cdot\right]$ *are independent for all* $i = 1, ..., n$*,* $j = 1, ..., t(k)$ *and* $k \geq 1$*.*

Before we conclude this section, we note that there are many possible choices for the operators $\mathcal{T}_c\left[\cdot\right]$ and $\mathcal{T}_g\left[\cdot\right]$ and each would yield a different algorithm instance in the family of NEAR-DGD-based methods. For example, both $\mathcal{T}_c\left[\cdot\right]$ and $\mathcal{T}_g\left[\cdot\right]$ can be identity

Method	Communication	Computation
NEAR-DGD [**13**], NEAR-DGD^{t_c,t_g} [**14**]	$\mathcal{T}_c\left[x_{i,k}^j\right] = x_{i,k}^j$	$\mathcal{T}_g\left[\nabla f_i\left(x_{i,k}^{t(k)}\right)\right] = \nabla f_i\left(x_{i,k}^{t(k)}\right)$
NEAR-DGD+Q [**15**]	$\mathcal{T}_c\left[x_{i,k}^j\right] = x_{i,k}^j + \epsilon_{i,k}^{j+1}$ (D)	$\mathcal{T}_g\left[\nabla f_i\left(x_{i,k}^{t(k)}\right)\right] = \nabla f_i\left(x_{i,k}^{t(k)}\right)$
SG-NEAR-DGD [**70**]	$\mathcal{T}_c\left[x_{i,k}^j\right] = x_{i,k}^j$	$\mathcal{T}_g\left[\nabla f_i\left(x_{i,k}^{t(k)}\right)\right] = \nabla f_i\left(x_{i,k}^{t(k)}\right) + \zeta_{i,k+1}$ (R)
S-NEAR-DGD	$\mathcal{T}_c\left[x_{i,k}^j\right] = x_{i,k}^j + \epsilon_{i,k}^{j+1}$ (R)	$\mathcal{T}_g\left[\nabla f_i\left(x_{i,k}^{t(k)}\right)\right] = \nabla f_i\left(x_{i,k}^{t(k)}\right) + \zeta_{i,k+1}$ (R)

Table 2.1. Summary of NEAR-DGD-based methods. (D) and (R) denote deterministic and random error vectors respectively.

operators as in [**13**]. We considered quantized communication using deterministic (D) algorithms (e.g. rounding to the nearest integer with no uncertainty) in [**15**], while a variant of NEAR-DGD that utilizes stochastic gradient approximations only was presented in [**70**]. This chapter unifies and generalizes these methods. We summarize the related works in Table 2.1, denoting deterministic and random error vectors with (D) and (R), respectively.

2.2. Convergence Analysis

In this section, we present our theoretical results on the convergence of S-NEAR-DGD. We assume that Assumptions 2.1.1-2.1.6 hold for the rest of this chapter. We first focus on the instance of our algorithm where the number of consensus rounds is constant at every iteration, i.e., $t(k) = t$ in (2.1.1b) for some positive integer $t > 0$. We refer to this method as S-NEAR-DGD^t. Next, we will analyze a second variant of S-NEAR-DGD, where the number of consensus steps increases by one at every iteration, namely $t(k) = k$, for $k \geq 1$. We will refer to this new version as S-NEAR-DGD$^+$.

Before our main analysis, we introduce some additional notation and a number of preliminary results.

2.2.1. Preliminaries

We will use the notation $\mathcal{F}_k^j$ to denote the σ-algebra containing all the information generated by S-NEAR-DGD up to and including the k-th inexact gradient step (calculated using using $\mathbf{g}_{k-1}$) and j subsequent nested consensus rounds. This includes the initial point $\mathbf{x}_0 = \mathbf{y}_0$, the vectors $\{\mathbf{x}_\tau^l : 1 \leq l \leq t(\tau)$ if $1 \leq \tau < k$ and $1 \leq l \leq j$ if $\tau = k\}$, the vectors $\mathbf{y}_\tau$ for $1 \leq \tau \leq k$, the vectors $\{\mathbf{q}_\tau^l : 1 \leq l \leq t(\tau)$ if $1 \leq \tau < k$ and $1 \leq l \leq j$ if $\tau = k\}$ and the vectors $\mathbf{g}_\tau$ for $0 \leq \tau \leq k-1$. For example, $\mathcal{F}_k^0$ would denote the σ-algebra containing all the information up to and including the vector $\mathbf{y}_k$ generated at the k-th gradient step (notice that $\mathcal{F}_k^0$ contains the inexact gradient $\mathbf{g}_{k-1}$, but not $\mathbf{g}_k$), while $\mathcal{F}_k^l$ would store all the information produced by S-NEAR-DGD up to and including $\mathbf{x}_k^l$, generated at the l^{th} consensus round after the k-th gradient step using $\mathbf{g}_{k-1}$.

We also introduce 4 lemmas here; Lemmas 2.2.1 and 2.2.2 will be used to show that the iterates generated by S-NEAR-DGD^t are bounded and to characterize their distance to the solution of Problem (1.0.1). Next, in Lemmas 2.2.3 and 2.2.4 we prove that the total communication and computation errors in a single iteration of the S-NEAR-DGD^t method have zero mean and bounded variance. These two error terms play a key role in our main analysis of convergence properties.

The following lemma is adapted from [**115**, Theorem 2.1.15, Chapter 2].

Lemma 2.2.1. ***(Gradient descent)*** *Let $h : \mathbb{R}^d \to \mathbb{R}$ be a μ-strongly convex function with L-Lipschitz gradients and define $x^\star := \arg\min_x h(x)$. Then the gradient method*

$x_{k+1} = x_k - \alpha \nabla f(x_k)$ with steplength $\alpha < \frac{2}{\mu+L}$, generates a sequence $\{x_k\}$ such that

$$\|x_{k+1} - x^\star\|^2 \leq \left(1 - \frac{2\alpha\mu L}{\mu + L}\right) \|x_k - x^\star\|^2 .$$

Lemma 2.2.2. ***(Convexity and smoothness)*** *The global function $\mathbf{f} : \mathbb{R}^{np} \to \mathbb{R}$, $\mathbf{f}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n f_i(x_i)$ is μ-strongly convex and L-smooth, where $\mu = \min_i \mu_i$ and $L = \max_i L_i$. In addition, the average function $\bar{f} : \mathbb{R}^p \to \mathbb{R}$, $\bar{f}(x) = \frac{1}{n}\sum_{i=1}^n f_i(x)$ is $\mu_{\bar{f}}$-strongly convex and $L_{\bar{f}}$-smooth, where $\mu_{\bar{f}} = \frac{1}{n}\sum_{i=1}^n \mu_i$ and $L_{\bar{f}} = \frac{1}{n}\sum_{i=1}^n L_i$.*

Proof. This is a direct consequence of Assumptions 2.1.1 and 2.1.2. □

Lemma 2.2.3. ***(Bounded communication error)*** *Let $\mathcal{E}_{t,k}^c := \mathbf{x}_k^t - \mathbf{Z}^t \mathbf{y}_k$ be the total communication error at the k-th iteration of S-NEAR-DGD^t, i.e. $t(k) = t$ in* (2.1.1a) *and* (2.1.1b)*. Then the following relations hold for $k \geq 1$,*

$$\mathbb{E}_{\mathcal{T}_c}\left[\mathcal{E}_{t,k}^c \middle| \mathcal{F}_k^0\right] = \mathbf{0}, \quad \mathbb{E}_{\mathcal{T}_c}\left[\left\|\mathcal{E}_{t,k}^c\right\|^2 \middle| \mathcal{F}_k^0\right] \leq \frac{4n\sigma_c^2}{1-\beta^2}.$$

Proof. Let $\tilde{\mathbf{Z}} := \mathbf{Z} - I_{np}$. Setting $\mathbf{x}_k^0 = \mathbf{y}_k$ and applying (2.1.1b), the error $\mathcal{E}_{t,k}^c$ can be expressed as $\mathcal{E}_{t,k}^c = \mathbf{x}_k^{t-1} + \tilde{\mathbf{Z}}\mathbf{q}_k^t - \mathbf{Z}^t \mathbf{x}_k^0$. Adding and subtracting the quantity $\sum_{j=1}^{t-1}\left(\mathbf{Z}^{t-j}\mathbf{x}_k^j\right) = \sum_{j=1}^{t-1}\left(\mathbf{Z}^{t-j}\left(\mathbf{x}_k^{j-1} + \tilde{\mathbf{Z}}\mathbf{q}_k^j\right)\right)$ (by (2.1.1b)) yields,

$$\begin{aligned}\mathcal{E}_{t,k}^c = \tilde{\mathbf{Z}}\left(\mathbf{q}_k^t - \mathbf{x}^{t-1}\right) - \sum_{j=1}^{t-2}\left(\mathbf{Z}^{t-j}\mathbf{x}_k^j\right) + \sum_{j=2}^{t-1}\left(\mathbf{Z}^{t-j}\mathbf{x}_k^{j-1}\right) \\ + \sum_{j=2}^{t-1}\left(\mathbf{Z}^{t-j}\tilde{\mathbf{Z}}\mathbf{q}_k^j\right) + \mathbf{Z}^{t-1}\tilde{\mathbf{Z}}\left(\mathbf{q}_k^1 - \mathbf{x}_k^0\right),\end{aligned}$$

where have taken the $(t-1)^{th}$ term out of $-\sum_{j=1}^{t-1}\left(\mathbf{Z}^{t-j}\mathbf{x}_k^j\right)$ and the 1^{st} term out of $\sum_{j=1}^{t-1}\left(\mathbf{Z}^{t-j}\left(\mathbf{x}_k^{j-1}+\tilde{\mathbf{Z}}\mathbf{q}_k^j\right)\right)$.

We observe that $\sum_{j=1}^{t-2}\left(\mathbf{Z}^{t-j}\mathbf{x}_k^j\right)=\sum_{j=2}^{t-1}\left(\mathbf{Z}^{t-j+1}\mathbf{x}_k^{j-1}\right)$, and after rearranging and combining the terms of the previous relation we obtain,

$$\mathcal{E}_{t,k}^c=\sum_{j=1}^{t}\left(\mathbf{Z}^{t-j}\tilde{\mathbf{Z}}\left(\mathbf{q}_k^j-\mathbf{x}_k^{j-1}\right)\right). \tag{2.2.1}$$

Let $d_k^j=\mathbf{q}_k^j-\mathbf{x}_k^{j-1}$. Noticing that $d_k^j=\left[\epsilon_{i,k}^j;...;\epsilon_{n,k}^j\right]$ as defined in Assumption 2.1.4, it follows that $\mathbb{E}_{\mathcal{T}_c}\left[d_k^j\middle|\mathcal{F}_k^{j-1}\right]=\mathbf{0}$ for $1\leq j\leq t$. Due to the fact that $\mathcal{F}_k^0\subseteq\mathcal{F}_k^1\subseteq...\subseteq\mathcal{F}_k^{j-1}$, applying the tower property of conditional expectation yields,

$$\mathbb{E}_{\mathcal{T}_c}\left[d_k^j\middle|\mathcal{F}_k^0\right]=\mathbb{E}_{\mathcal{T}_c}\left[\mathbb{E}_{\mathcal{T}_c}\left[d_k^j\middle|\mathcal{F}_k^{j-1}\right]\middle|\mathcal{F}_k^0\right]=\mathbf{0}. \tag{2.2.2}$$

Combining the preceding relation with (2.2.1) and due to the linearity of expectation, we obtain $\mathbb{E}_{\mathcal{T}_c}\left[\mathcal{E}_{t,k}^c\middle|\mathcal{F}_k^0\right]=\mathbf{0}$. This completes the first part of the proof.

Let $D_k^j=\mathbf{Z}^{t-j}\tilde{\mathbf{Z}}d_k^j$. By the spectral properties of $\mathbf{Z}$, we have

$$\left\|\mathbf{Z}^{t-j}\tilde{\mathbf{Z}}\right\|=\max_{i>1}|\lambda_i^{t-j}||\lambda_i-1|\leq 2\beta^{t-j}.$$

We thus obtain for $1 \leq j \leq t$,

$$\begin{aligned}
\mathbb{E}_{\mathcal{T}_c}\left[\left\|D_k^j\right\|^2 \middle| \mathcal{F}_k^0\right] &\leq 4\beta^{2(t-j)}\mathbb{E}_{\mathcal{T}_c}\left[\left\|d_k^j\right\|^2 \middle| \mathcal{F}_k^0\right] \\
&= 4\beta^{2(t-j)}\mathbb{E}_{\mathcal{T}_c}\left[\mathbb{E}_{\mathcal{T}_c}\left[\left\|d_k^j\right\|^2 \middle| \mathcal{F}_k^{j-1}\right] \middle| \mathcal{F}_k^0\right] \\
&= 4\beta^{2(t-j)}\mathbb{E}_{\mathcal{T}_c}\left[\sum_{i=1}^{n}\mathbb{E}_{\mathcal{T}_c}\left[\left\|\epsilon_{i,k}^j\right\|^2 \middle| \mathcal{F}_k^{j-1}\right] \middle| \mathcal{F}_k^0\right] \\
&\leq 4\beta^{2(t-j)} n\sigma_c^2,
\end{aligned} \tag{2.2.3}$$

where we derived the second inequality using the tower property of conditional expectation and applied Assumption 2.1.4 to get the last inequality.

Assumption 2.1.4 implies that for $i_1 \neq i_2$ and $j_1 \neq j_2$, $\epsilon_{i_1,k}^{j_1}$ and $\epsilon_{i_2,k}^{j_2}$ and by extension $d_k^{j_1}$ and $d_k^{j_2}$ are independent. Eq. (2.2.2) then yields $\mathbb{E}_{\mathcal{T}_c}\left[\left\langle D_k^{j_1}, D_k^{j_2}\right\rangle \middle| \mathcal{F}_k^0\right] = \mathbf{0}$. Combining this fact and linearity of expectation yields $\mathbb{E}_{\mathcal{T}_c}\left[\left\|\mathcal{E}_{t,k}^c\right\|^2 \middle| \mathcal{F}_k^0\right] = \mathbb{E}_{\mathcal{T}_c}\left[\left\|\sum_{j=1}^t D_k^j\right\|^2 \middle| \mathcal{F}_k^0\right] = \sum_{j=1}^t \mathbb{E}_{\mathcal{T}_c}\left[\left\|D_k^j\right\|^2 \middle| \mathcal{F}_k^0\right]$. Applying (2.2.3) to this last relation yields,

$$\mathbb{E}_{\mathcal{T}_c}\left[\left\|\mathcal{E}_{t,k}^c\right\|^2 \middle| \mathcal{F}_k^0\right] \leq 4n\sigma_c^2 \sum_{j=1}^{t}\beta^{2(t-j)} \leq \frac{4n\sigma_c^2}{1-\beta^2},$$

where we used $\sum_{j=1}^t \beta^{2(t-j)} = \sum_{j=0}^{t-1}\beta^{2j} = \frac{1-\beta^{2t}}{1-\beta^2} \leq \frac{1}{1-\beta^2}$ to get the last inequality.

□

Lemma 2.2.4. ***(Bounded computation error)*** *Let $\mathcal{E}_k^g := \mathbf{g}_{k-1} - \nabla\mathbf{f}\left(\mathbf{x}_{k-1}^t\right)$ be the computation error at the k-th iteration of S-NEAR-DGD^t. Then the following statements*

hold for all $k \geq 1$,

$$\mathbb{E}_{\mathcal{T}_g}\left[\mathcal{E}_k^g \middle| \mathcal{F}_{k-1}^t\right] = \mathbf{0}, \quad \mathbb{E}_{\mathcal{T}_g}\left[\left\|\mathcal{E}_k^g\right\|^2 \middle| \mathcal{F}_{k-1}^t\right] \leq n\sigma_g^2.$$

Proof. We observe that $\mathcal{E}_k^g = [\zeta_{1,k}; ...; \zeta_{n,k}]$ as defined in Assumption 2.1.5. Due to the unbiasedness of $\mathcal{T}_g[\cdot]$, we obtain

$$\mathbb{E}_{\mathcal{T}_g}\left[\mathcal{E}_k^g \middle| \mathcal{F}_{k-1}^t\right] = \mathbb{E}_{\mathcal{T}_g}\left[\mathbf{g}_{k-1} - \nabla \mathbf{f}\left(\mathbf{x}_{k-1}^t\right) \middle| \mathcal{F}_{k-1}^t\right] = \mathbf{0},$$

For the magnitude square of $\mathcal{E}_k^g$ we have,

$$\left\|\mathcal{E}_k^g\right\|^2 = \left\|\mathbf{g}_{k-1} - \nabla \mathbf{f}\left(\mathbf{x}_{k-1}^t\right)\right\|^2 = \sum_{i=1}^{n} \left\|\zeta_{i,k}\right\|^2.$$

Taking the expectation conditional to $\mathcal{F}_{k-1}^t$ on both sides of the equation above and using Assumption 2.1.5 establishes the desired results. □

We are now ready to proceed with our main analysis of the convergence properties of S-NEAR-DGD.

2.2.2. Main Analysis

For simplicity, from this point on we will use the notation $\mathbb{E}[\cdot]$ to denote the expected value taken over the randomness of both $\mathcal{T}_c$ and $\mathcal{T}_g$. We begin our convergence analysis by proving that the iterates generated by S-NEAR-DGD^t are bounded in expectation in Lemma 2.2.5. Next, we demonstrate that the distance between the local iterates produced by our method and their average is bounded in Lemma 2.2.6. In Lemma 2.2.7, we prove an intermediate result stating that the distance between the average iterates of

S-NEAR-DGD^t and the optimal solution is bounded. We then use this result to show the linear of convergence of S-NEAR-DGD^t to a neighborhood of the optimal solution in Theorem 2.2.8, and we characterize the size of this error neighborhood in terms of network and problem related quantities and the precision of the stochastic gradients and the noisy communication channel. We prove convergence to a neighborhood of the optimal solution for the local iterates of S-NEAR-DGD^t in Corollary 2.2.9. We conclude our analysis by proving that the average iterates of S-NEAR-DGD$^+$ converge with geometric rate to an improved error neighborhood compared to S-NEAR-DGD^t in Theorem 2.2.10.

Lemma 2.2.5. ***(Bounded iterates)*** *Let $\mathbf{x}_k$ and $\mathbf{y}_k$ be the iterates generated by S-NEAR-DGD^t ($t(k) = t$ in Eq. (2.1.1b) and (2.1.1a)) starting from initial point $\mathbf{y}_0 = \mathbf{x}_0 \in \mathbb{R}^{np}$ and let the steplength α satisfy*

$$\alpha < \frac{2}{\mu + L},$$

where $\mu = \min_i \mu_i$ and $L = \max_i L_i$.

Then $\mathbf{x}_k$ and $\mathbf{y}_k$ are bounded in expectation for $k \geq 1$, i.e.,

$$\mathbb{E}\left[\|\mathbf{y}_k\|^2\right] \leq D + \frac{(1+\kappa)^2 n\sigma_g^2}{2L^2} + \frac{2(1+\kappa)^2 n\sigma_c^2}{\alpha\,(1-\beta^2)\,L^2},$$

$$\mathbb{E}\left[\left\|\mathbf{x}_k^t\right\|^2\right] \leq D + \frac{(1+\kappa)^2 n\sigma_g^2}{2L^2} + \frac{2(1+\kappa)^2 n\sigma_c^2}{\alpha\,(1-\beta^2)\,L^2} + \frac{4n\sigma_c^2}{1-\beta^2},$$

where $D = 2\mathbb{E}\left[\|\mathbf{y}_0 - \mathbf{u}^\star\|^2\right] + 2\,(1+4\nu^{-3})\,\|\mathbf{u}^\star\|^2$, $\mathbf{u}^\star = [u_1^\star; u_2^\star; ...; u_n^\star] \in \mathbb{R}^{np}$, $u_i^\star = \arg\min_x f_i(x)$, $\nu = \frac{2\alpha\mu L}{\mu+L}$ and $\kappa = L/\mu$ is the condition number of Problem 1.0.2.

Proof. Consider,

$$\begin{aligned}\|\mathbf{y}_{k+1}-\mathbf{u}^\star\|^2 &= \left\|\mathbf{x}_k^t-\alpha\mathbf{g}_k-\mathbf{u}^\star\right\|^2\\ &= \left\|\mathbf{x}_k^t-\alpha\nabla\mathbf{f}(\mathbf{x}_k^t)-\mathbf{u}^\star-\alpha\mathcal{E}_{k+1}^g\right\|^2\\ &= \left\|\mathbf{x}_k^t-\alpha\nabla\mathbf{f}(\mathbf{x}_k^t)-\mathbf{u}^\star\right\|^2+\alpha^2\left\|\mathcal{E}_{k+1}^g\right\|^2-2\alpha\left\langle\mathbf{x}_k^t-\alpha\nabla\mathbf{f}(\mathbf{x}_k^t)-\mathbf{u}^\star,\mathcal{E}_{k+1}^g\right\rangle,\end{aligned}$$

where we used (2.1.1a) to get the first equality and added and subtracted $\alpha\nabla\mathbf{f}\left(\mathbf{x}_k^t\right)$ and applied the computation error definition $\mathcal{E}_{k+1}^g := \mathbf{g}_k-\nabla\mathbf{f}\left(\mathbf{x}_k^t\right)$ to obtain the second equality.

Taking the expectation conditional to $\mathcal{F}_k^t$ on both sides of the inequality above and applying Lemma 2.2.4 yields,

$$\mathbb{E}\left[\|\mathbf{y}_{k+1}-\mathbf{u}^\star\|^2\,\middle|\,\mathcal{F}_k^t\right] \leq \left\|\mathbf{x}_k^t-\alpha\nabla\mathbf{f}(\mathbf{x}_k^t)-\mathbf{u}^\star\right\|^2+\alpha^2 n\sigma_g^2. \tag{2.2.4}$$

For the first term on the right-hand side of (2.2.4), after combining Lemma 2.2.1 with Lemma 2.2.2 and due to $\alpha < \frac{2}{\mu+L}$ we acquire,

$$\left\|\mathbf{x}_k^t-\alpha\nabla\mathbf{f}(\mathbf{x}_k^t)-\mathbf{u}^\star\right\|^2 \leq (1-\nu)\left\|\mathbf{x}_k^t-\mathbf{u}^\star\right\|^2,$$

where $\nu = \frac{2\alpha\mu L}{\mu+L} = \frac{2\alpha L}{1+\kappa} < 1$.

Expanding the term on the right hand side of the above relation yields,

$$\begin{aligned}
\left\|\mathbf{x}_k^t - \mathbf{u}^\star\right\|^2 &= \left\|\mathcal{E}_{t,k}^c + \mathbf{Z}^t\mathbf{y}_k - \mathbf{Z}^t\mathbf{u}^\star + \mathbf{Z}^t\mathbf{u}^\star - \mathbf{u}^\star\right\|^2 \\
&= \left\|\mathcal{E}_{t,k}^c\right\|^2 + \left\|\mathbf{Z}^t\left(\mathbf{y}_k - \mathbf{u}^\star\right) - \left(I - \mathbf{Z}^t\right)\mathbf{u}^\star\right\|^2 + 2\left\langle\mathcal{E}_{t,k}^c, \mathbf{Z}^t\mathbf{y}_k - \mathbf{u}^\star\right\rangle \\
&\leq \left\|\mathcal{E}_{t,k}^c\right\|^2 + (1+\nu)\left\|\mathbf{Z}^t\left(\mathbf{y}_k - \mathbf{u}^\star\right)\right\|^2 \\
&\quad + \left(1+\nu^{-1}\right)\left\|\left(I - \mathbf{Z}^t\right)\mathbf{u}^\star\right\|^2 + 2\left\langle\mathcal{E}_{t,k}^c, \mathbf{Z}^t\mathbf{y}_k - \mathbf{u}^\star\right\rangle \\
&\leq \left\|\mathcal{E}_{t,k}^c\right\|^2 + (1+\nu)\left\|\mathbf{y}_k - \mathbf{u}^\star\right\|^2 + 4\left(1+\nu^{-1}\right)\left\|\mathbf{u}^\star\right\|^2 + 2\left\langle\mathcal{E}_{t,k}^c, \mathbf{Z}^t\mathbf{y}_k - \mathbf{u}^\star\right\rangle,
\end{aligned}$$

where we added and subtracted the quantities $\mathbf{Z}^t\mathbf{y}_k$ and $\mathbf{Z}^t\mathbf{u}^\star$ and applied the communication error definition $\mathcal{E}_{t,k}^c := \mathbf{x}_k^t - \mathbf{Z}^t\mathbf{y}_k$ to get the first equality. We used the standard inequality $\pm 2\langle a, b\rangle \leq c\|a\|^2 + c^{-1}\|b\|^2$ that holds for any two vectors a, b and positive constant $c > 0$ to obtain the first inequality. Finally, we derived the last inequality using the relations $\|\mathbf{Z}^t\| = 1$ and $\|I - \mathbf{Z}^t\| < 2$ that hold due to Assumption 2.1.3.

Due to the fact that $\mathcal{F}_k^0 \subseteq \mathcal{F}_k^t$, combining the preceding three relations and taking the expectation conditional on $\mathcal{F}_k^0$ on both sides of (2.2.4) yields,

$$\begin{aligned}
\mathbb{E}\left[\left\|\mathbf{y}_{k+1} - \mathbf{u}^\star\right\|^2\middle|\mathcal{F}_k^0\right] &\leq \left(1-\nu^2\right)\left\|\mathbf{y}_k - \mathbf{u}^\star\right\|^2 + \alpha^2 n\sigma_g^2 + (1-\nu)\mathbb{E}\left[\left\|\mathcal{E}_{t,k}^c\right\|^2\middle|\mathcal{F}_k^0\right] \\
&\quad + 4\nu^{-1}\left(1-\nu^2\right)\left\|\mathbf{u}^\star\right\|^2 + 2(1-\nu)\mathbb{E}\left[\left\langle\mathcal{E}_{t,k}^c, \mathbf{Z}^t\mathbf{y}_k - \mathbf{u}^\star\right\rangle\middle|\mathcal{F}_k^0\right] \\
&\leq \left(1-\nu^2\right)\left\|\mathbf{y}_k - \mathbf{u}^\star\right\|^2 + \alpha^2 n\sigma_g^2 + (1-\nu)\frac{4n\sigma_c^2}{1-\beta^2} + 4\nu^{-1}\left(1-\nu^2\right)\left\|\mathbf{u}^\star\right\|^2,
\end{aligned}$$

where we applied Lemma 2.2.3 to get the last inequality. Taking the total expectation on both sides of the relation above and applying recursively over iterations $0, 1, \ldots, k$ yields,

$$\begin{aligned}\mathbb{E}\Big[\big\|\mathbf{y}_k - \mathbf{u}^\star\big\|^2\Big] &\leq \left(1-\nu^2\right)^k \mathbb{E}\left[\|\mathbf{y}_0 - \mathbf{u}^\star\|^2\right] + \alpha^2\nu^{-2}n\sigma_g^2 \\ &\quad + (\nu^{-2}-\nu^{-1})\frac{4n\sigma_c^2}{1-\beta^2} + 4\left(\nu^{-3}-\nu^{-1}\right)\|\mathbf{u}^\star\|^2 \\ &\leq \mathbb{E}\left[\|\mathbf{y}_0 - \mathbf{u}^\star\|^2\right] + \alpha^2\nu^{-2}n\sigma_g^2 + \frac{4\nu^{-2}n\sigma_c^2}{1-\beta^2} + 4\nu^{-3}\|\mathbf{u}^\star\|^2,\end{aligned}$$

where we used $\sum_{h=0}^{k-1}(1-\nu^2)^h \leq \nu^{-2}$ to get the first inequality and $\nu > 0$ to get the second inequality.

Moreover, the statement $\|\mathbf{y}_k\|^2 = \|\mathbf{y}_k - \mathbf{u}^\star + \mathbf{u}^\star\|^2 \leq 2\|\mathbf{y}_k - \mathbf{u}^\star\|^2 + 2\|\mathbf{u}^\star\|^2$ trivially holds. Taking the total expectation on both sides of this relation, yields,

$$\begin{aligned}\mathbb{E}\Big[\big\|\mathbf{y}_k\big\|^2\Big] &\leq 2\mathbb{E}\left[\|\mathbf{y}_k - \mathbf{u}^\star\|^2\right] + 2\|\mathbf{u}^\star\|^2 \\ &\leq 2\mathbb{E}\left[\|\mathbf{y}_0 - \mathbf{u}^\star\|^2\right] + 2\alpha^2\nu^{-2}n\sigma_g^2 + \frac{8\nu^{-2}n\sigma_c^2}{1-\beta^2} + 2\left(1+4\nu^{-3}\right)\|\mathbf{u}^\star\|^2.\end{aligned} \tag{2.2.5}$$

Applying the definitions of D and κ to (2.2.5) yields the first result of this lemma.

Finally, the following statement also holds

$$\begin{aligned}\big\|\mathbf{x}_k^t\big\|^2 &= \big\|\mathcal{E}_{t,k}^c + \mathbf{Z}^t\mathbf{y}_k\big\|^2 \\ &= \big\|\mathcal{E}_{t,k}^c\big\|^2 + \big\|\mathbf{Z}^t\mathbf{y}_k\big\|^2 + 2\left\langle\mathcal{E}_{t,k}^c, \mathbf{Z}^t\mathbf{y}_k\right\rangle \\ &\leq \big\|\mathcal{E}_{t,k}^c\big\|^2 + \|\mathbf{y}_k\|^2 + 2\left\langle\mathcal{E}_{t,k}^c, \mathbf{Z}^t\mathbf{y}_k\right\rangle,\end{aligned}$$

where we used the non-expansiveness of $\mathbf{Z}$ for the last inequality.

Taking the expectation conditional on $\mathcal{F}_k^0$ on both sides of the preceding relation and applying Lemma 2.2.3 yields,

$$\mathbb{E}\left[\left\|\mathbf{x}_k^t\right\|^2 \middle| \mathcal{F}_k^0\right] \leq \frac{4n\sigma_c^2}{1-\beta^2} + \left\|\mathbf{y}_k\right\|^2.$$

Taking the total expectation on both sides of the relation above, applying (2.2.5) and the definitions of D, ν and κ concludes this proof.

□

We will now use the preceding lemma to prove that the distance between the local and the average iterates generated by S-NEAR-DGD^t is bounded. This distance can be interpreted as a measure of consensus violation, with small values indicating small disagreement between nodes.

Lemma 2.2.6. ***(Bounded distance to average)*** *Let $x_{i,k}^t$ and $y_{i,k}$ be the local iterates produced by S-NEAR-DGD^t at node i and iteration k and let $\bar{x}_k^t := \sum_{i=1}^n x_{i,k}^t$ and $\bar{y}_k := \sum_{i=1}^n y_{i,k}$ denote the average iterates across all nodes. Then the distance between the local and average iterates is bounded in expectation for all $i = 1, ..., n$ and $k = 1, 2, ...$, namely,*

$$\mathbb{E}\left[\left\|x_{i,k}^t - \bar{x}_k^t\right\|^2\right] \leq \mathbb{E}\left[\left\|\mathbf{x}_k^t - \mathbf{M}\mathbf{x}_k^t\right\|^2\right] \leq \beta^{2t} D + \frac{\beta^{2t}(1+\kappa)^2 n\sigma_g^2}{2L^2} + \frac{2\beta^{2t}(1+\kappa)^2 n\sigma_c^2}{\alpha^2\left(1-\beta^2\right)L^2} + \frac{4n\sigma_c^2}{1-\beta^2},$$

and

$$\mathbb{E}\left[\left\|y_{i,k} - \bar{y}_k\right\|^2\right] \leq \mathbb{E}\left[\left\|\mathbf{y}_k - \mathbf{M}\mathbf{y}_k\right\|^2\right] \leq D + \frac{(1+\kappa)^2 n\sigma_g^2}{2L^2} + \frac{2(1+\kappa)^2 n\sigma_c^2}{\alpha^2\left(1-\beta^2\right)L^2},$$

where $\mathbf{M} = \left(\frac{1_n 1_n^T}{n} \otimes I_p\right) \in \mathbb{R}^{np}$ *is the averaging matrix, constant* D *is defined in Lemma 2.2.5,* $\kappa = L/\mu$ *is the condition number of Problem 1.0.2,* $L = \max_i L_i$ *and* $\mu = \min_i L_i$.

Proof. Observing that $\sum_{i=1}^n \left\|x_{i,k}^t - \bar{x}_k^t\right\|^2 = \|\mathbf{x}_k^t - \mathbf{M}\mathbf{x}_k^t\|^2$, we obtain,

$$\left\|x_{i,k}^t - \bar{x}_k^t\right\|^2 \leq \left\|\mathbf{x}_k^t - \mathbf{M}\mathbf{x}_k^t\right\|^2, \quad i = 1, ..., n. \tag{2.2.6}$$

We can bound the right-hand side of (2.2.6) as

(2.2.7)

$$\begin{aligned}
\left\|\mathbf{x}_k^t - \mathbf{M}\mathbf{x}_k\right\|^2 &= \left\|\mathcal{E}_{t,k}^c + \mathbf{Z}^t\mathbf{y}_k - \mathbf{M}\mathbf{x}_k - \mathbf{M}\mathbf{y}_k + \mathbf{M}\mathbf{y}_k\right\|^2 \\
&= \left\|\mathcal{E}_{t,k}^c + \left(\mathbf{Z}^t - \mathbf{M}\right)\mathbf{y}_k - \mathbf{M}\mathbf{x}_k + \mathbf{M}\mathbf{Z}^t\mathbf{y}_k\right\|^2 \\
&= \left\|\left(I - \mathbf{M}\right)\mathcal{E}_{t,k}^c\right\|^2 + \left\|\left(\mathbf{Z}^t - \mathbf{M}\right)\mathbf{y}_k\right\|^2 + 2\left\langle\left(I - \mathbf{M}\right)\mathcal{E}_{t,k}^c, \left(\mathbf{Z}^t - \mathbf{M}\right)\mathbf{y}_k\right\rangle \\
&\leq \left\|\mathcal{E}_{t,k}^c\right\|^2 + \beta^{2t}\left\|\mathbf{y}_k\right\|^2 + 2\left\langle\left(I - \mathbf{M}\right)\mathcal{E}_{t,k}^c, \left(\mathbf{Z}^t - \mathbf{M}\right)\mathbf{y}_k\right\rangle,
\end{aligned}$$

where we applied the definition of the communication error $\mathcal{E}_{t,k}^c$ of Lemma 2.2.3 and added and subtracted $\mathbf{M}\mathbf{y}_k$ to obtain the first equality. We used the fact that $\mathbf{M}\mathbf{Z}^t = \mathbf{M}$ to get the second equality. We derive the last inequality from Cauchy-Schwarz and the spectral properties of $\mathbf{Z}^t = \mathbf{W}^t \otimes I_p$ and $\mathbf{M} = \left(\frac{1_n 1_n^T}{n}\right) \otimes I_p$; both $\mathbf{W}^t$ and $\frac{1_n 1_n^T}{n}$ have a maximum eigenvalue at 1 associated with the eigenvector 1_n, implying that the null space of $\mathbf{W}^t - \frac{1_n 1_n^T}{n}$ is parallel to 1_n and $\|\mathbf{Z}^t - \mathbf{M}\| = \left\|\mathbf{W}^t - \frac{1_n 1_n^T}{n}\right\| = \beta^t$.

Taking the expectation conditional to $\mathcal{F}_k^0$ on both sides of (2.2.7) and applying Lemma 2.2.3 yields,

$$\mathbb{E}\left[\left\|\mathbf{x}_k^t - \mathbf{M}\mathbf{x}_k\right\|^2 \middle| \mathcal{F}_k^0\right] \leq \frac{4n\sigma_c^2}{1 - \beta^2} + \beta^{2t}\left\|\mathbf{y}_k\right\|^2.$$

Taking the total expectation on both sides and applying Lemma 2.2.5 yields the first result of this lemma.

Similarly, the following inequality holds for the $\mathbf{y}_k$ iterates,

$$\|y_{i,k} - \bar{y}_k\|^2 \leq \|\mathbf{y}_k - \mathbf{M}\mathbf{y}_k\|^2, \quad i = 1, ..., n. \tag{2.2.8}$$

For the right-hand side of (2.2.8), we have,

$$\|\mathbf{y}_k - \mathbf{M}\mathbf{y}_k\|^2 = \|(I - \mathbf{M})\,\mathbf{y}_k\|^2 \leq \|\mathbf{y}_k\|^2,$$

where we have used the fact that $\|I - \mathbf{M}\| = 1$.

Taking the total expectation on both sides and applying Lemma 2.2.5 concludes this proof.

□

The bounds established in Lemma 2.2.6 indicate that there are at least three factors preventing the local iterates produced by S-NEAR-DGD^t from reaching consensus: errors related to network connectivity, represented by β, and errors caused by the inexact computation process and the noisy communication channel associated with the constants σ_g and σ_c respectively.

Before presenting our main theorem, we state one more intermediate result on the distance of the average $\bar{y}_k$ iterates to the solution of Problem (1.0.2).

Lemma 2.2.7. ***(Bounded distance to minimum)*** *Let* $\bar{y}_k := \frac{1}{n}\sum_{i=1}^{n} y_{i,k}$ *denote the average of the local* $y_{i,k}$ *iterates generated by S-NEAR-DGD*^{t} *under steplength* α *satisfying*

$$\alpha < \frac{2}{\mu_{\bar{f}} + L_{\bar{f}}},$$

where $\mu_{\bar{f}} = \frac{1}{n}\sum_{i=1}^{n} \mu_i$ *and* $L_{\bar{f}} = \frac{1}{n}\sum_{i=1}^{n} L_i$.

Then the following inequality holds for $k = 1, 2, ...$

$$\mathbb{E}\left[\left\|\bar{y}_{k+1} - x^\star\right\|^2 \middle| \mathcal{F}_k^t\right] \leq \rho \left\|\bar{x}_k^t - x^\star\right\|^2 + \frac{\alpha^2\sigma_g^2}{n} + \frac{\alpha\rho L^2 \Delta_{\mathbf{x}}}{n\gamma_{\bar{f}}},$$

where $x^\star = \arg\min_x f(x)$, $\rho = 1 - \alpha\gamma_{\bar{f}}$, $\gamma_{\bar{f}} = \frac{\mu_{\bar{f}} L_{\bar{f}}}{\mu_{\bar{f}} + L_{\bar{f}}}$, $L = \max_i L_i$, $\Delta_{\mathbf{x}} = \left\|\mathbf{x}_k^t - \mathbf{M}\mathbf{x}_k^t\right\|^2$ *and* $\mathbf{M} = \left(\frac{1_n 1_n^T}{n} \otimes I_p\right) \in \mathbb{R}^{np}$ *is the averaging matrix.*

Proof. Applying (2.1.2a) to $(k+1)^{th}$ iteration we obtain,

$$\bar{y}_{k+1} = \bar{x}_k^t - \alpha\bar{g}_k,$$

where $\bar{g}_k = \frac{1}{n}\sum_{i=1}^{n} g_{i,k} = \frac{1}{n}\sum_{i=1}^{n}\left(\zeta_{i,k+1} + \nabla f_i\left(x_{i,k}^t\right)\right)$. Let $h_k = \frac{1}{n}\sum_{i=1}^{n}\nabla f_i\left(x_{i,k}^t\right)$. Adding and subtracting αh_k to the right-hand side of the preceding relation and taking the square norm on both sides yields,

$$\begin{aligned}\left\|\bar{y}_{k+1} - x^\star\right\|^2 &= \left\|\bar{x}_k^t - \alpha h_k - x^\star\right\|^2 + \alpha^2\left\|h_k - \bar{g}_k\right\|^2 + 2\alpha\left\langle \bar{x}_k^t - \alpha h_k - x^\star, h_k - \bar{g}_k\right\rangle \\ &= \left\|\bar{x}_k^t - \alpha h_k - x^\star\right\|^2 + \frac{\alpha^2}{n^2}\left\|\sum_{i=1}^{n}\zeta_{i,k+1}\right\|^2 - \frac{2\alpha}{n}\sum_{i=1}^{n}\left\langle \bar{x}_k^t - \alpha h_k - x^\star, \zeta_{i,k+1}\right\rangle.\end{aligned}$$

Moreover, let $\tilde{\rho} = \frac{\alpha\gamma_{\bar{f}}}{1-2\alpha\gamma_{\bar{f}}} > 0$. We can re-write the first term on the right-hand side of the inequality above as,

$$\begin{aligned}\left\|\bar{x}_k^t - \alpha h_k - x^\star\right\|^2 &\leq (1+\tilde{\rho})\left\|\bar{x}_k^t - \alpha\nabla\bar{f}\left(\bar{x}_k^t\right) - x^\star\right\|^2 + \alpha^2\left(1+\tilde{\rho}^{-1}\right)\left\|h_k - \nabla\bar{f}\left(\bar{x}_k^t\right)\right\|^2 \\ &\leq \left(1-\alpha\gamma_{\bar{f}}\right)\left\|\bar{x}_k^t - x^\star\right\|^2 + \alpha^2\left(1+\tilde{\rho}^{-1}\right)\left\|h_k - \nabla\bar{f}\left(\bar{x}_k^t\right)\right\|^2,\end{aligned}$$

where we added and subtracted the quantity $\alpha\nabla\bar{f}\left(\bar{x}_k^t\right)$ and used the relation $\pm 2\langle a, b\rangle \leq c\|a\|^2 + c^{-1}\|b\|^2$ that holds for any two vectors a, b and positive constant c to obtain the first inequality. We derive the second inequality after combining Lemmas 2.2.2 and 2.2.1 that hold due to $\alpha < \frac{2}{\mu_{\bar{f}}+L_{\bar{f}}}$ and $x^\star = \arg\min_x \bar{f}(x)$.

We notice that $\mathbb{E}\left[\zeta_{i,k+1}\middle|\mathcal{F}_k^t\right] = \mathbf{0}$ and that

$$\mathbb{E}\left[\left\|\sum_{i=1}^n \zeta_{i,k+1}\right\|^2\middle|\mathcal{F}_k^t\right] = \mathbb{E}\left[\sum_{i=1}^n \left\|\zeta_{i,k+1}\right\|^2\middle|\mathcal{F}_k^t\right] + \mathbb{E}\left[\sum_{i_1\neq i_2}\left\langle\zeta_{i_1,k+1},\zeta_{i_2,k+1}\right\rangle\middle|\mathcal{F}_k^t\right] \leq n\sigma_g^2,$$

due to Assumption 2.1.5 and the linearity of expectation. Combining all of the preceding relations and taking the expectation conditional on $\mathcal{F}_k^t$, yields,

(2.2.9)

$$\mathbb{E}\left[\left\|\bar{y}_{k+1} - x^\star\right\|^2\middle|\mathcal{F}_k^t\right] = \left(1-\alpha\gamma_{\bar{f}}\right)\left\|\bar{x}_k^t - x^\star\right\| + \alpha^2\left(1+\tilde{\rho}^{-1}\right)\left\|h_k - \nabla\bar{f}\left(\bar{x}_k^t\right)\right\|^2 + \frac{\alpha^2\sigma_g^2}{n}$$

Finally, for any set of vectors $v_i \in \mathbb{R}^p$, $i = 1, ..., n$ we have

$$\left\|\sum_{i=1}^n v_i\right\|^2 = \sum_{h=1}^p\left(\sum_{i=1}^n [v_i]_h\right)^2 \leq n\sum_{h=1}^p\sum_{i=1}^n [v_i]_h^2 = n\sum_{i=1}^n \|v_i\|^2,$$

where we used the fact that $\pm 2ab \leq a^2 + b^2$ for any pair of scalars a, b to get the first inequality and reversed the order of summation to get the last equality. We can use this

result to obtain,

$$\begin{aligned}\left\|h_k - \nabla \bar{f}\left(\bar{x}_k^t\right)\right\|^2 &= \frac{1}{n^2}\left\|\sum_{i=1}^{n}\left(\nabla f_i\left(x_{i,k}^t\right) - \nabla f_i\left(\bar{x}_k^t\right)\right)\right\|^2 \\ &\leq \frac{n}{n^2}\sum_{i=1}^{n}\left\|\nabla f_i\left(x_{i,k}^t\right) - \nabla f_i\left(\bar{x}_k^t\right)\right\|^2 \\ &\leq \frac{L^2}{n}\sum_{i=1}^{n}\left\|x_{i,k}^t - \bar{x}_k^t\right\|^2 \\ &= \frac{L^2}{n}\left\|\mathbf{x}_k^t - \mathbf{M}\mathbf{x}_k^t\right\|^2,\end{aligned}$$

where we used Assumption 2.1.1 to get the second inequality.

Substituting the immediately previous relation in (2.2.9), observing $1+\tilde{\rho}^{-1} = \left(1-\alpha\gamma_{\bar{f}}\right)/\alpha\gamma_{\bar{f}}$ and applying the definition of ρ yields the final result.

□

We have now obtained all necessary results to prove the convergence of S-NEAR-DGD^t to a neighborhood of the optimal solution in the next theorem.

Theorem 2.2.8. ***(Convergence of S-NEAR-DGD^t)*** *Let $\bar{x}_k^t := \frac{1}{n}\sum_{i=1}^{n} x_{i,k}^t$ denote the average of the local $x_{i,k}^t$ iterates generated by S-NEAR-DGD^t from initial point $\mathbf{y}_0$ and let the steplength α satisfy,*

$$\alpha < \min\left\{\frac{2}{\mu+L}, \frac{2}{\mu_{\bar{f}}+L_{\bar{f}}}\right\},$$

where $\mu = \min_i L_i$, $L = \max_i L_i$, $\mu_{\bar{f}} = \frac{1}{n}\sum_{i=1}^{n}\mu_i$ and $L_{\bar{f}} = \frac{1}{n}\sum_{i=1}^{n} L_i$.

Then the distance of $\bar{x}_k^t$ to the optimal solution $x^\star$ of Problem (1.0.2) is bounded in expectation for $k = 1, 2, ...,$

$$\begin{aligned}\mathbb{E}\left[\left\|\bar{x}_{k+1}^t - x^\star\right\|^2\right] &\leq \rho\mathbb{E}\left[\left\|\bar{x}_k^t - x^\star\right\|^2\right] + \frac{\alpha\beta^{2t}\rho L^2 D}{n\gamma_{\bar{f}}} \\ &+ \frac{\alpha^2\sigma_g^2}{n} + \frac{\alpha\beta^{2t}\left(1+\kappa\right)^2\rho\sigma_g^2}{2\gamma_{\bar{f}}} + \frac{4\alpha\rho L^2\sigma_c^2}{\left(1-\beta^2\right)\gamma_{\bar{f}}} + \frac{2\beta^{2t}\left(1+\kappa\right)^2\rho\sigma_c^2}{\alpha\left(1-\beta^2\right)\gamma_{\bar{f}}},\end{aligned} \tag{2.2.10}$$

and

$$\begin{aligned}\mathbb{E}\left[\left\|\bar{x}_k^t - x^\star\right\|^2\right] &\leq \rho^k\mathbb{E}\left[\left\|\bar{x}_0 - x^\star\right\|^2\right] + \frac{\beta^{2t}\rho L^2 D}{n\gamma_{\bar{f}}^2} \\ &+ \frac{\alpha\sigma_g^2}{n\gamma_{\bar{f}}} + \frac{\beta^{2t}\left(1+\kappa\right)^2\rho\sigma_g^2}{2\gamma_{\bar{f}}^2} + \frac{4\rho L^2\sigma_c^2}{\left(1-\beta^2\right)\gamma_{\bar{f}}^2} + \frac{2\beta^{2t}\left(1+\kappa\right)^2\rho\sigma_c^2}{\alpha^2\left(1-\beta^2\right)\gamma_{\bar{f}}^2},\end{aligned} \tag{2.2.11}$$

where $\bar{x}_0 = \frac{1}{n}\sum_{i=1}^n y_{i,0}$, $\rho = 1 - \alpha\gamma_{\bar{f}}$, $\gamma_{\bar{f}} = \frac{\mu_{\bar{f}} L_{\bar{f}}}{\mu_{\bar{f}} + L_{\bar{f}}}$, $\kappa = L/\mu$ is the condition number of Problem 1.0.2 and the constant D is defined in Lemma 2.2.5.

Proof. Applying (2.1.2b) to the $(k+1)^{th}$ iteration yields,

$$\bar{x}_{k+1}^j = \bar{x}_{k+1}^{j-1},\ j = 1, ..., t,$$

which in turn implies that $\bar{x}_{k+1}^j = \bar{y}_{k+1}$ for $j = 1, ..., t$.

Hence, the relation $\left\|\bar{x}_{k+1}^t - x^\star\right\|^2 = \left\|\bar{y}_{k+1} - x^\star\right\|^2$ holds. Taking the expectation conditional to $\mathcal{F}_k^t$ on both sides of this equality and applying Lemma 2.2.7 yields,

$$\mathbb{E}\left[\left\|\bar{x}_{k+1}^t - x^\star\right\|^2 \middle| \mathcal{F}_k^t\right] \leq \rho\left\|\bar{x}_k^t - x^\star\right\|^2 + \frac{\alpha^2\sigma_g^2}{n} + \frac{\alpha\rho L^2\Delta_{\mathbf{x}}}{n\gamma_{\bar{f}}},$$

where $\Delta_{\mathbf{x}} = \left\|\mathbf{x}_k^t - \mathbf{M}\mathbf{x}_k^t\right\|^2$.

Taking the total expectation on both sides of the relation above and applying Lemma 2.2.6 yields,

$$\begin{aligned}\mathbb{E}\left[\left\|\bar{x}_{k+1}^t - x^\star\right\|^2\right] \leq \rho\mathbb{E}\left[\left\|\bar{x}_k^t - x^\star\right\|^2\right] + \frac{\alpha\beta^{2t}\rho L^2 D}{n\gamma_{\bar{f}}} + \frac{\alpha^2\sigma_g^2}{n} \\ + \frac{\alpha\beta^{2t}(1+\kappa)^2\rho\sigma_g^2}{2\gamma_{\bar{f}}} + \frac{4\alpha\rho L^2\sigma_c^2}{\gamma_{\bar{f}}\left(1-\beta^2\right)} + \frac{2\beta^{2t}(1+\kappa)^2\rho\sigma_c^2}{\alpha\left(1-\beta^2\right)\gamma_{\bar{f}}}.\end{aligned} \tag{2.2.12}$$

We notice that $\rho < 1$ and after applying (2.2.12) recursively and then using the bound $\sum_{h=0}^{k-1}\rho^h \leq (1-\rho)^{-1}$ we obtain,

$$\begin{aligned}\mathbb{E}\left[\left\|\bar{x}_k^t - x^\star\right\|^2\right] \leq \rho^k\mathbb{E}\left[\|\bar{x}_0 - x^\star\|^2\right] + \frac{\alpha\beta^{2t}\rho L^2 D}{n\gamma_{\bar{f}}(1-\rho)} + \frac{\alpha^2\sigma_g^2}{n(1-\rho)} \\ + \frac{\alpha\beta^{2t}(1+\kappa)^2\rho\sigma_g^2}{2\gamma_{\bar{f}}(1-\rho)} + \frac{4\alpha\rho L^2\sigma_c^2}{\gamma_{\bar{f}}\left(1-\beta^2\right)(1-\rho)} + \frac{2\beta^{2t}(1+\kappa)^2\rho\sigma_c^2}{\alpha\left(1-\beta^2\right)\gamma_{\bar{f}}(1-\rho)}.\end{aligned}$$

Applying the definition of ρ completes the proof.

□

Theorem 2.2.8 indicates that the average iterates of S-NEAR-DGD^t converge in expectation to a neighborhood of the optimal solution $x^\star$ of Problem (1.0.1). We have quantified the dependence of this neighborhood on the connectivity of the network and the errors due to imperfect computation and communication through the terms containing the quantities β, σ_g and σ_c, respectively. We observe that the 2^{nd} and 3^{rd} error terms in (2.2.11) scale favorably with the number of nodes n, yielding a variance reduction effect proportional to network size. Our bounds indicate that higher values of the steplength α yield faster convergence rates ρ. On the other hand, α has a mixed effect on the size of the error neighborhood; the 2^{nd} term in (2.2.11) is associated with inexact computation and increases with α, while the last term in (2.2.11) is associated with noisy communication

and decreases with α. The size of the error neighborhood increases with the condition number κ as expected, while the dependence on the algorithm parameter t indicates that performing additional consensus steps mitigates the error due to network connectivity and the errors induced by the operators $\mathcal{T}_g[\cdot]$ and $\mathcal{T}_c[\cdot]$.

In the next corollary, we will use Theorem 2.2.8 and Lemmas 2.2.7 and 2.2.5 to show that the the local iterates produced by S-NEAR-DGD^t also have bounded distance to the solution of Problem (1.0.1).

Corollary 2.2.9. ***(Convergence of local iterates (S-NEAR-DGD^t)*** *Let $x_{i,k}^t$ and $y_{i,k}$ be the local iterates generated by S-NEAR-DGD^t at node i and iteration k from initial point $\mathbf{x}_0 = \mathbf{y}_0 = [y_{1,0}; ...; y_{n,0}] \in \mathbb{R}^{np}$ and let the steplength α satisfy*

$$\alpha < \min\left\{\frac{2}{\mu + L}, \frac{2}{\mu_{\bar{f}} + L_{\bar{f}}}\right\},$$

where $\mu = \min_i \mu_i$, $L = \max_i L_i$, $\mu_{\bar{f}} = \frac{1}{n}\sum_{i=1}^n \mu_i$ and $L_{\bar{f}} = \frac{1}{n}\sum_{i=1}^n L_i$.

Then for $i = 1, ..., n$ and $k \geq 1$ the distance of the local iterates to the solution of Problem (1.0.2) is bounded, i.e.

$$\mathbb{E}\left[\left\|x_{i,k}^t - x^\star\right\|^2\right] \leq 2\rho^k \mathbb{E}\left[\|\bar{x}_0 - x^\star\|^2\right] + 2\beta^{2t}\left(1 + \frac{C}{n}\right) D + \frac{2\alpha\sigma_g^2}{n\gamma_{\bar{f}}} + \frac{\beta^{2t}(1+\kappa)^2(n+C)\sigma_g^2}{L^2} + \frac{8(n+C)\sigma_c^2}{1-\beta^2} + \frac{4\beta^{2t}(1+\kappa)^2(n+C)\sigma_c^2}{\alpha^2(1-\beta^2)L^2},$$

and,

$$\mathbb{E}\left[\left\|y_{i,k} - x^\star\right\|^2\right] \leq 2\rho^k \mathbb{E}\left[\|\bar{x}_0 - x^\star\|^2\right] + 2\left(1 + \frac{\beta^{2t}C}{n}\right) D + \frac{2\alpha\sigma_g^2}{n\gamma_{\bar{f}}} + \frac{(1+\kappa)^2(n+\beta^{2t}C)\sigma_g^2}{L^2} + \frac{8C\sigma_c^2}{1-\beta^2} + \frac{4(1+\kappa)^2(n+\beta^{2t}C)\sigma_c^2}{\alpha^2(1-\beta^2)L^2},$$

where $C = \frac{\rho L^2}{\gamma_{\bar{f}}^2}$, $\rho = 1 - \alpha\gamma_{\bar{f}}$, $\gamma_{\bar{f}} = \frac{\mu_{\bar{f}} L_{\bar{f}}}{\mu_{\bar{f}} + L_{\bar{f}}}$, $\bar{x}_0 = \sum_{i=1}^{n} y_{i,0}$ *and the constant* D *is defined in Lemma 2.2.5.*

Proof. For all $i \in \{1, ...n\}$ and $k \geq 1$, the following relation holds for the $x_{i,k}^t$ iterates,

$$\begin{aligned} \left\|x_{i,k}^t - x^\star\right\|^2 &= \left\|x_{i,k}^t - \bar{x}_k^t + \bar{x}_k^t - x^\star\right\|^2 \\ &\leq 2\left\|x_{i,k}^t - \bar{x}_k^t\right\|^2 + 2\left\|\bar{x}_k^t - x^\star\right\|^2, \end{aligned} \tag{2.2.13}$$

where we added and subtracted $\bar{x}_k^t$ to get the first equality.

Taking the total expectation on both sides of (2.2.13) and applying Lemma 2.2.6, Theorem 2.2.8 and the definition of C yields the first result of this corollary.

Similarly, for the $y_{i,k}$ local iterates we have,

$$\begin{aligned} \|y_{i,k} - x^\star\|^2 &= \|y_{i,k} - \bar{y}_k + \bar{y}_k - x^\star\|^2 \\ &\leq 2\|y_{i,k} - \bar{y}_k\|^2 + 2\|\bar{y}_k - x^\star\|^2 \\ &= 2\|y_{i,k} - \bar{y}_k\|^2 + 2\left\|\bar{x}_k^t - x^\star\right\|^2, \end{aligned} \tag{2.2.14}$$

where we derive the first equality by adding and subtracting $\bar{y}_k$ and used (2.1.2b) to obtain the last equality.

Taking the total expectation on both sides of (2.2.14) and applying Theorem 2.2.8, Lemma 2.2.6 and the definition of C completes the proof. □

Corollary 2.2.9 concludes our analysis of the S-NEAR-DGD^t method. For the remainder of this section, we derive the convergence properties of S-NEAR-DGD$^+$, i.e. $t(k) = k$ for $k \geq 1$ in (2.1.1a) and (2.1.1b).

Theorem 2.2.10. ***(Convergence of S-NEAR-DGD$^+$)*** *Consider the S-NEAR-DGD$^+$ method, i.e. $t(k) = k$ for $k \geq 1$. Let $\bar{x}_k^k = \frac{1}{n}\sum_{i=1}^n x_{i,k}^k$ be the average iterates produced by S-NEAR-DGD$^+$ and let the steplength α satisfy*

$$\alpha < \min\left\{\frac{2}{\mu + L}, \frac{2}{\mu_{\bar{f}} + L_{\bar{f}}}\right\}.$$

Then the distance of $\bar{x}_k^k$ to $x^\star$ is bounded for $k = 1, 2, ...,$ namely

$$\mathbb{E}\left[\left\|\bar{x}_k^k - x^\star\right\|^2\right] \leq \rho^k \mathbb{E}\left[\|\bar{x}_0 - x^\star\|^2\right] + \frac{\eta\theta^k \alpha\rho L^2 D}{n\gamma_{\bar{f}}} + \frac{\alpha\sigma_g^2}{n\gamma_{\bar{f}}}$$
$$+ \frac{\eta\theta^k \alpha\left(1+\kappa\right)^2 \rho\sigma_g^2}{2\gamma_{\bar{f}}} + \frac{4\rho L^2\sigma_c^2}{\left(1-\beta^2\right)\gamma_{\bar{f}}^2} + \frac{2\eta\theta^k\left(1+\kappa\right)^2\rho\sigma_c^2}{\alpha\left(1-\beta^2\right)\gamma_{\bar{f}}},$$

where $\eta = \left|\beta^2 - \rho\right|^{-1}$ and $\theta = \max\{\rho, \beta^2\}$.

Proof. Replacing t with k in (2.2.10) in Theorem 2.2.8 yields,

$$\mathbb{E}\left[\left\|\bar{x}_{k+1}^{k+1} - x^\star\right\|^2\right] \leq \rho\mathbb{E}\left[\left\|\bar{x}_k^k - x^\star\right\|^2\right] + \frac{\alpha\beta^{2k}\rho L^2 D}{n\gamma_{\bar{f}}} + \frac{\alpha^2\sigma_g^2}{n}$$
$$+ \frac{\alpha\beta^{2k}\left(1+\kappa\right)^2\rho\sigma_g^2}{2\gamma_{\bar{f}}} + \frac{4\alpha\rho L^2\sigma_c^2}{\left(1-\beta^2\right)\gamma_{\bar{f}}} + \frac{2\beta^{2k}\left(1+\kappa\right)^2\rho\sigma_c^2}{\alpha\left(1-\beta^2\right)\gamma_{\bar{f}}}.$$

Applying recursively for iterations $1, 2, \ldots, k$, we obtain,

(2.2.15)

$$\mathbb{E}\left[\left\|\bar{x}_k^k - x^\star\right\|^2\right] \leq \rho^k\mathbb{E}\left[\|\bar{x}_0 - x^\star\|^2\right]$$
$$+ S_1\left(\frac{\alpha\rho L^2 D}{n\gamma_{\bar{f}}} + \frac{\alpha\left(1+\kappa\right)^2\rho\sigma_g^2}{2\gamma_{\bar{f}}} + \frac{2\left(1+\kappa\right)^2\rho\sigma_c^2}{\alpha\left(1-\beta^2\right)\gamma_{\bar{f}}}\right) + S_2\left(\frac{\alpha^2\sigma_g^2}{n} + \frac{4\alpha\rho L^2\sigma_c^2}{\left(1-\beta^2\right)\gamma_{\bar{f}}}\right)$$

where $S_1 = \sum_{j=0}^{k-1}\rho^j\beta^{2(k-1-j)}$ and $S_2 = \sum_{j=0}^{k-1}\rho^j$.

Let $\psi = \frac{\rho}{\beta^2}$. Then $S_1 = \beta^{2(k-1)} \sum_{j=0}^{k-1} \psi^j = \beta^{2(k-1)} \frac{1-\psi^k}{1-\psi} = \frac{\beta^{2k}-\rho^k}{\beta^2-\rho} \leq \eta\theta^k$. Applying this result and the bound $S_2 \leq \frac{1}{1-\rho} = \frac{1}{\alpha\gamma_{\bar{f}}}$ to (2.2.15) yields the final result. □

Theorem 2.2.10 indicates that S-NEAR-DGD$^+$ converges with geometric rate $\theta = \max\{\rho, \beta^2\}$ to a neighborhood of the optimal solution $x^\star$ of Problem 1.0.1 with size

$$\limsup_{k\to\infty} \mathbb{E}\left[\left\|\bar{x}_k^k - x^\star\right\|^2\right] = \frac{\alpha\sigma_g^2}{n\gamma_{\bar{f}}} + \frac{4\rho L^2\sigma_c^2}{(1-\beta^2)\,\gamma_{\bar{f}}^2}. \tag{2.2.16}$$

The first error term on right-hand side of Eq. (2.2.16) depends on the variance of the gradient error σ_g and is inversely proportional to the network size n. This scaling with n, which has a similar effect to centralized mini-batching, is a trait that our method shares with a number of distributed stochastic gradient algorithms. The last error term depends on the variance of the communication error σ_c and increases with β, implying that badly connected networks accumulate more communication error over time.

Conversely, Eq. (2.2.11) of Theorem 2.2.8 yields,

$$\begin{aligned}\limsup_{k\to\infty} \mathbb{E}\left[\left\|\bar{x}_k^t - x^\star\right\|^2\right] &= \frac{\alpha\sigma_g^2}{n\gamma_{\bar{f}}} + \frac{4\rho L^2\sigma_c^2}{(1-\beta^2)\,\gamma_{\bar{f}}^2} \\ &\quad + \frac{\beta^{2t}\rho}{\gamma_{\bar{f}}^2}\left(\frac{L^2 D}{n} + \frac{(1+\kappa)^2\sigma_g^2}{2} + \frac{2(1+\kappa)^2\sigma_c^2}{\alpha(1-\beta^2)}\right).\end{aligned} \tag{2.2.17}$$

Comparing (2.2.16) and (2.2.17), we observe that (2.2.17) contains three additional error terms, all of which depend directly on the algorithm parameter t. Our results imply that S-NEAR-DGD^t generally converges to a worse error neighborhood than S-NEAR-DGD$^+$, and approaches the error neighborhood of S-NEAR-DGD$^+$ as $t \to \infty$.

2.3. Numerical results

2.3.1. Comparison to existing algorithms

To quantify the empirical performance of S-NEAR-DGD, we consider the following regularized logistic regression problem for the classification of the mushrooms dataset [**49**],

$$\min_{x\in\mathbb{R}^p} f(x) = \frac{1}{M}\sum_{s=1}^{M} \log(1 + e^{-b_s\langle A_s, x\rangle}) + \frac{1}{M}\|x\|_2^2,$$

where $M = 8124$ is the total number of samples, $A \in \mathbb{R}^{M\times p}$ is a feature matrix, $p = 118$ is the problem dimension and $b \in \{-1, 1\}^M$ is a vector of labels.

We evenly distributed the samples among $n = 14$ nodes and assigned to node i the function $f_i(x) = |S_i|^{-1}\sum_{s\in S_i} \log(1 + e^{-b_s\langle A_s,x\rangle}) + M^{-1}\|x\|_2^2$, where S_i is the set of sample indices accessible to node i. We modeled the network as a connected, random graph generated using the Erdős-Rényi model [**55**] with edge probability 0.5. To construct the stochastic gradient approximations each node randomly samples with replacement $B = 16$ indices from its local distribution and computes a mini-batch gradient (we note that due to the finite number of samples, these gradients satisfy Assumption 2.1.5). To simulate the inexact communication operator $\mathcal{T}_c[\cdot]$ we implemented the probabilistic quantizer in [**182**], described in Example 1.

Moreover, we tested a number of different approaches to handle the communication noise, summarized in Table 2.2. Specifically, variant Q.1 is the scheme S-NEAR-DGD uses in step (2.1.1b), and includes a consensus step using the quantized variables $q_{i,k}^j$ and the addition of the error correction term $\left(x_{i,k}^{j-1} - q_{i,k}^j\right)$. Variant Q.2 considers a more naïve approach, where a simple weighted average of the noisy variables $q_{i,k}$ is calculated without

the addition of error correction. Finally, in variant Q.3 we assume that node i either does not have access to its local quantized variable $q_{i,k}$ or prefers to use the original quantity $x_{i,k}^{j-1}$ whenever possible and thus computes the weighted average using its original local variable and noisy versions from its neighbors. For algorithms that perform consensus step on gradients instead of the decision variables, similar schemes were implemented. We compared S-NEAR-DGD^t with $t = 2$ and $t = 5$, to versions of DGD [**111, 154**],

Table 2.2. Quantized consensus step variations

Variant name	Consensus update
Q.1	$x_{i,k}^j \leftarrow \sum_{l=1}^n \left(w_{il} q_{l,k}^j\right) + \left(x_{i,k}^{j-1} - q_{i,k}^j\right)$
Q.2	$x_{i,k}^j \leftarrow \sum_{l=1}^n \left(w_{il} q_{i,k}^j\right)$
Q.3	$x_{i,k}^j \leftarrow w_{ii} x_{i,k}^{j-1} + \sum_{l \in \mathcal{N}_i} \left(w_{il} q_{l,k}^j\right)$

EXTRA [**146**] and DIGing [**113**] with noisy gradients and communication. All methods use the same mini-batch gradients at every iteration and exchange variables that are quantized with the same probabilistic protocol. We implemented the consensus step variations Q.1, Q.2 and Q.3 for all methods[1]. All algorithms shared the same steplength $\alpha = 1$. Our results averaged over 20 trials of the experiment are shown in Fig. 2.1 (we note that smaller steplengths yield similar data trends). We plot the squared error $\|\bar{x}_k - x^\star\|^2$ against the number of iterations for quantization parameter values $\Delta = 10$ (left) and $\Delta = 10^5$ (right). The most important observation in Fig. 2.1 is that the GT methods (EXTRA, DIGing), diverge without error correction in Q.2 and Q.3 regardless of the quantization resolution. This aligns with recent findings indicating that GT methods are more sensitive to the network topology, which directly affects the quantization error,

[1]Combining DIGing with Q.1 and using the true local gradients instead of stochastic approximations recovers the iterates of Q-NEXT [**84**]. However, the authors of [**84**] accompany their method with a dynamic quantizer that we did not implement for our numerical experiments.

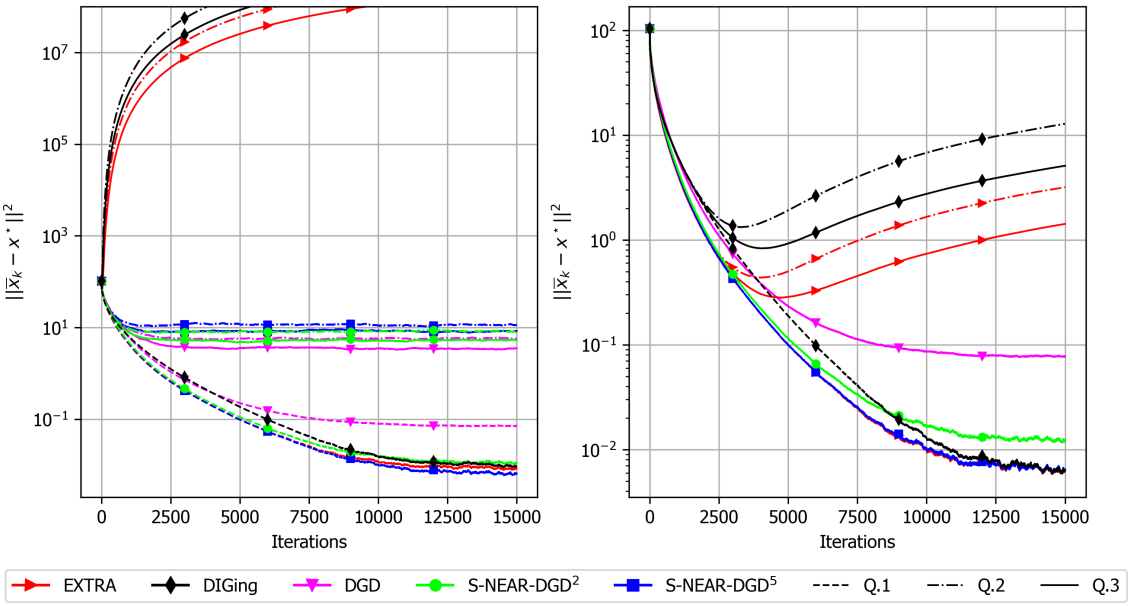


Figure 2.1. Error plots, $\Delta = 10$ (left) and $\Delta = 10^5$ (right)

compared to purely primal methods [**183**]. For coarse quantization ($\Delta = 10$, Fig. 2.1, left), S-NEAR-DGD^5 combined with Q.1 slightly outperforms the remaining methods in terms of convergence accuracy. When the quantization is fine ($\Delta = 10^5$, Fig. 2.1, right), we observe that the choice of consensus variant has no effect on the performance of the primal methods (DGD, S-NEAR-DGD), making S-NEAR-DGD the most attractive option when quantized versions of the local variables, necessary for the implementation of Q.1, are not available. When $\Delta = 10^{-5}$, S-NEAR-DGD^5, EXTRA and DIGing overlap as the number of iterations increases under variant Q.1.

2.3.2. Scalability

To evaluate the scalability of our method and the effect of network type on convergence accuracy and speed, we tested 5 network sizes ($n = 5, 10, 15, 20, 25$) and 5 different network

types: *i*) complete, *ii*) random (connected Erdős-Rényi, edge probability 0.4), *iii*) 4-cyclic (i.e. starting from a ring graph, every node is connected to 4 immediate neighbors), *iv*) ring and *v*) path. We compared 2 instances of the NEAR-DGD^t method, *i*) $t = 1$ and *ii*) $t = 7$. We opted to exclude NEAR-DGD$^+$ from this set of experiments to facilitate result interpretability in terms of communication load. We set $\alpha = 1$ and $\Delta = 10^2$ for all instances of the experiment, while the batch size for the calculation of the local stochastic gradients was set to $B = 16$ in all cases. Different methods applied to networks of identical size selected (randomly, with replacement) the same samples at each iteration.

Our results are summarized in Figure 2.2. In Fig. 2.2, top left, we terminated all experiments after $T = 2 \cdot 10^4$ iterations and plotted the normalized function value error $\left(f\left(\bar{x}_k\right) - f\left(x^\star\right)\right)/f(x^\star)$, averaged over the last $\tau = 10^3$ iterations. We observe that networks with better connectivity converge closer to the true optimal value, implying that the terms inversely proportional to $(1-\beta^2)$ dominate the error neighborhood in Eq. (2.2.17). Adding more nodes improves convergence accuracy for well-connected graphs (complete, random), possibly due to the "variance reduction" effect on the stochastic gradients discussed in the previous section. For the remaining graphs, however, this beneficial effect is outweighed by the decrease in connectivity that results from the introduction of additional nodes and consequently, large values of n yield worse convergence neighborhoods. Increasing the number of consensus steps per iteration has a favorable effect on accuracy, an observation consistent with our theoretical findings in the previous section.

For the next set of plots, we analyze the run time and cost of the algorithm until termination. The presence of stochastic noise makes the establishment of a termination criterion that performs well for all parameter combinations a challenging task. Inspired

by [**170**], we tracked an approximate time average $\bar{f}$ of the function values $f(\bar{x}_k)$ using Welford's method for online sample mean computation, i.e. $\bar{f}_k = \bar{f}_{k-1} + \frac{f(\bar{x}_k) - \bar{f}_{k-1}}{k}$, for $k = 1, 2, ...$, with $\bar{f}_0 = f(\bar{x}_0)$. We terminate the algorithm at iteration count k if $\left| \frac{\bar{f}_k - \bar{f}_{k-1}}{\bar{f}_{k-1}} \right| < \epsilon$, where ϵ is a tolerance parameter.

In Figure 2.2, top right, we graph the number of steps (or gradient evaluations) until the termination criterion described in the previous paragraph is satisfied for $\epsilon = 10^{-5}$. We observe a similar trend to Figure 2.2, top left, indicating that poor connectivity has an adverse effect on both accuracy and the rate of convergence, although the latter is not predicted by the theory. Increasing the number of consensus steps per iteration reduces the total number of steps needed to satisfy the stopping criterion. Finally, in the bottom row of Fig. 2.2 we plot the total application cost per node until termination, which we calculated using the cost framework first introduced in [**13**],

$$\text{Cost} = c_c \times \#\text{Communications} + c_g \times \#\text{Computations},$$

where c_c and c_g are constants representing the application-specific costs of communication and computation respectively.

In Fig. 2.2, bottom right, the communication is a 100 times cheaper than computation, i.e. $c_c = 0.01 \cdot c_g$. Increasing the number of consensus steps per iteration almost always yields faster convergence in terms of total cost, excluding some cases where the network is already reasonably well-connected (eg. complete graph). In Fig. 2.2, bottom left, the costs of computation and communication are equal, i.e. $c_c = c_g$, and increasing the number of consensus steps per iteration results in higher total cost in all cases.

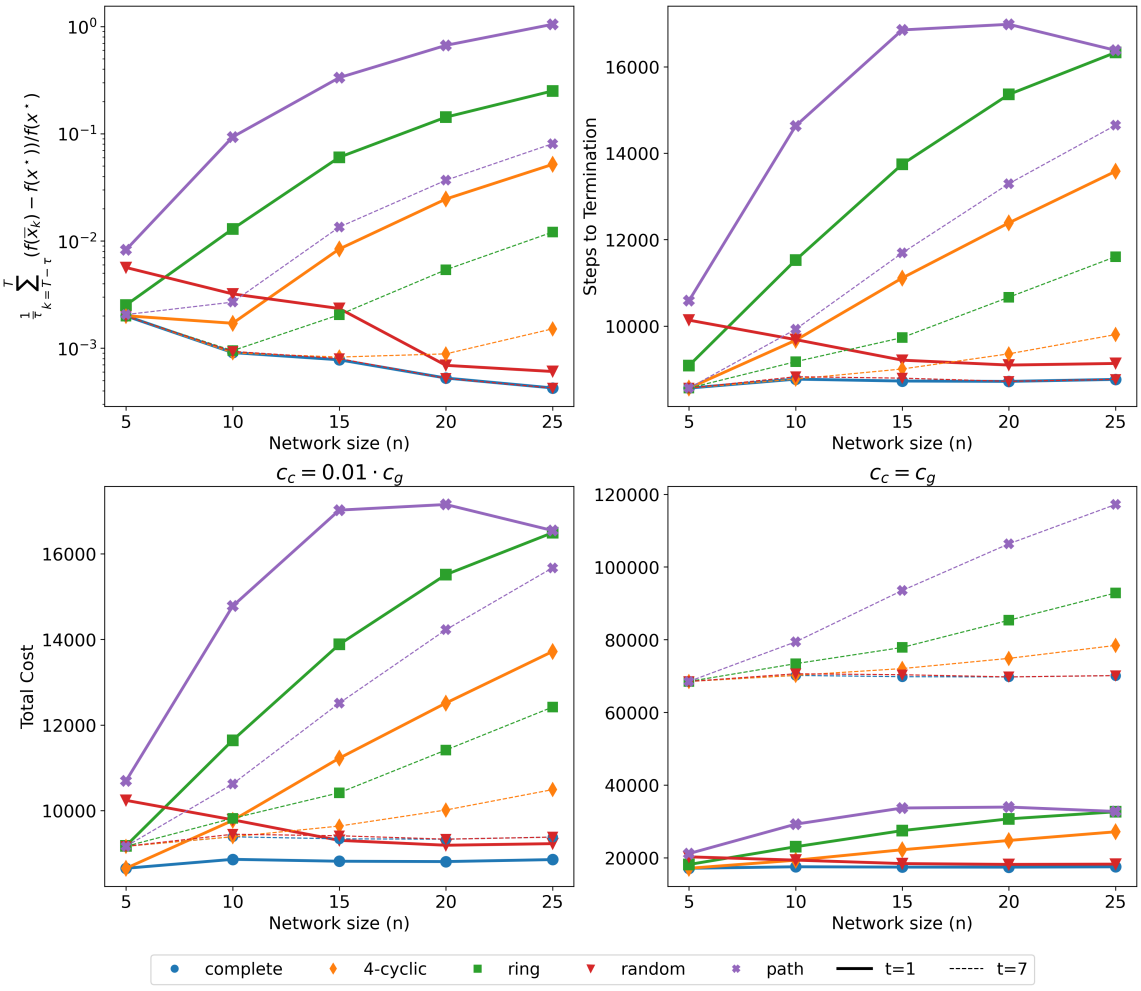


Figure 2.2. Dependence on network type and size. Function value error averaged over the last τ iterations out of T total iterations (top left), steps/gradient evaluations until termination (top right), total cost until termination when communication is cheaper than computation (bottom left), and when communication and computation have the same cost (bottom right).

In Figure 2.2, bottom, the costs of computation and communication are equal, i.e. $c_c = c_g$, and increasing the number of consensus steps per iteration results in higher total cost for all cases.

2.4. Summary

We proposed a first order method (S-NEAR-DGD) for distributed optimization over fixed, undirected networks, that can tolerate stochastic gradient approximations and noisy information exchange to alleviate the loads of computation and communication. The strength of our method lies in its flexible framework, which alternates between gradient steps and a number of nested consensus rounds that can be adjusted to best match the application requirements. Our analysis indicates that S-NEAR-DGD converges in expectation to a neighborhood of the optimal solution when the local objective functions are strongly convex and have Lipschitz continuous gradients. We have quantified the dependence of this neighborhood on algorithm parameters, problem-related quantities and the topology and size of the network. Empirical results demonstrate that our algorithm performs comparably or better than state-of-the-art methods, depending on the implementation of the quantized consensus.

CHAPTER 3

Nested Distributed Gradient Methods for Non-Convex Optimization With Second Order Guarantees

3.1. Convergence Analysis

In this section, we generalize the convergence properties of the NEAR-DGD method (1.0.5), (1.0.6) from the strongly convex [**13**] to the nonconvex setting. Before stating our results, we introduce some additional notation and list our assumptions for the remainder of this chapter.

3.1.1. Notation

In this Chapter, all vectors are column vectors. We will use the notation v' to refer to the transpose of a vector v. The average of the vectors $v_i \in \mathbb{R}^p$ contained in $\mathbf{v} = [v_1', ..., v_n']' \in \mathbb{R}^{np}$ will be denoted by $\bar{v}$, i.e. $\bar{v} = \frac{1}{n}\sum_{i=1}^n v_i$. We use uppercase boldface letters for matrices and will denote the element in the i^{th} row and j^{th} column of matrix $\mathbf{H}$ with h_{ij}. We will refer to the i^{th} (real) eigenvalue in ascending order (i.e. 1st is the smallest) of a matrix $\mathbf{H}$ as $\lambda_i(\mathbf{H})$. We use the notation 1_n for the vector of ones of dimension n. We will use $\|\cdot\|$ to denote the l_2-norm, i.e. for $v \in \mathbb{R}^p$ we have $\|v\| = \sqrt{\sum_{i=1}^p [v]_i^2}$ where $[v]_i$ is the i-th element of v. The inner product of vectors v, u will be denoted by $\langle v, u\rangle$. The symbol $\otimes$ will denote the Kronecker product operation. Finally, we define the averaging matrix $\mathbf{M} := \left(\frac{1_n 1_n'}{n} \otimes I_p\right)$.

Assumption 3.1.1. ***(Consensus matrix)*** *The matrix* $\mathbf{W} \in \mathbb{R}^{n \times n}$ *of Problem 1.0.2 has the following properties: i) symmetry, ii) double stochasticity, iii)* $w_{ij} > 0$ *if and only if* $(i,j) \in \mathcal{E}$ *or* $i = j$ *and* $w_{ij} = 0$ *otherwise and iv) positive-definiteness.*

We can construct a matrix $\tilde{\mathbf{W}}$ satisfying properties $(i) - (iii)$ of Assumption 3.1.1 by defining its elements to be max degree or Metropolis-Hastings weights [**172**], for instance. Such matrices are not necessarily positive-definite, so we can further enforce property (iv) using simple linear transformations (for example, we could define $\mathbf{W} = (1-\delta)^{-1}(\tilde{\mathbf{W}} - \delta I_n)$, where $\delta < \lambda_1(\tilde{\mathbf{W}})$ is a constant). For the rest of this work, we will be referring to the 2^{nd} largest eigenvalue of $\mathbf{W}$ as β, i.e. $\beta = \lambda_{n-1}(\mathbf{W})$.

We also adopt the following standard assumptions for the global function $\mathbf{f} : \mathbb{R}^{np} \to \mathbb{R}$ of Problem 1.0.2. Note that unlike other works on this topic (eg. [**186, 40**]), we do not make any assumptions on the local functions $f_i : \mathbb{R}^p \to \mathbb{R}$.

Assumption 3.1.2. ***(Lipschitz gradients)*** *The global objective function* $\mathbf{f} : \mathbb{R}^{np} \to \mathbb{R}$ *has L-Lipschitz continuous gradients, i.e.* $\|\nabla \mathbf{f}(\mathbf{x}) - \nabla \mathbf{f}(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^{np}$.

Assumption 3.1.3. ***(Coercivity)*** *The global objective function* $\mathbf{f} : \mathbb{R}^{np} \to \mathbb{R}$ *is coercive, i.e.* $\lim_{k \to \infty} \mathbf{f}(\mathbf{x}_k) = \infty$ *for every sequence* $\{\mathbf{x}_k\}$ *such that* $\|\mathbf{x}_k\| \to \infty$.

We need one more assumption to guarantee the convergence of NEAR-DGD. Namely, we require that the Lyapunov functions we will employ in our analysis are "sharp" around their critical points, up to a reparametrization. This property is formally summarized below.

Definition 1. ***(Kurdyka-Łojasiewicz (KL) property)*** [9] *A function $h : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ has the KL property at $x^\star \in dom(\partial h)$ if there exists $\eta \in (0, +\infty]$, a neighborhood U of $x^\star$, and a continuous concave function $\phi : [0, \eta) \to \mathbb{R}^+$ such that:*

(1) *$\phi(0) = 0$;*

(2) *ϕ is $\mathcal{C}^1$ (continuously differentiable) on $(0, \eta)$;*

(3) *for all $s \in (0, \eta)$, $\phi'(s) > 0$; and*

(4) *for all $x \in U \cap \{x : h(x^\star) < h(x) < h(x^\star) + \eta\}$, the KL inequality holds:*

$$\phi'(h(x) - h(x^\star)) \cdot dist(0, \partial h(x)) \geq 1.$$

Proper lower semicontinuous functions which satisfy the KL inequality at each point of $dom(\partial h)$ are called KL functions.

In the next subsection, we provide first order guarantees (i.e. show convergence to critical points) for two instances of the NEAR-DGD method. First, we study a variant of NEAR-DGD where a fixed number of consensus rounds are executed at every iteration, i.e. $t(k) = t$ for some $t \in \mathbb{N}^+$ in Eq. (1.0.5), and to which we will refer as NEAR-DGD^t. Next, we examine a variant of NEAR-DGD where the number of consensus rounds increases by 1 at every iteration, i.e. $t(k) = k$ in Eq. (1.0.5). We will refer to this latter variant as NEAR-DGD$^+$.

3.1.2. First order guarantees

We begin this subsection with a general result which will be necessary for proving the convergence of the iterates of NEAR-DGD. Namely, we outline a number of sufficient

conditions related to the KL property under which a sequence achieves local convergence (namely, it converges if initialized from a suitable neighborhood).

Lemma 3.1.4. (Local convergence) *Suppose the function $h : \mathbb{R}^p \to \mathbb{R}$ is a KL function and let ϕ, U and η be the objects in Def. 1. Moreover, suppose that there exists a sequence $\{x_k\}$, a sequence $\{\nu_k\}$ and a point $x^\star \in \mathbb{R}^p$ such that i) $h(x^\star) \leq h(x_k) < h(x^\star)+\eta$ for all $k \geq 0$; and ii) if for some index $\tau \geq 0$ the point $x_\tau \in \{x_k\}$ satisfies the KL inequality with respect to $x^\star$, then the following inequality holds*

$$\|x_{\tau+1} - x_\tau\| \leq c(\phi(l_\tau) - \phi(l_{\tau+1})) + \nu_\tau, \tag{3.1.1}$$

where $l_\tau = h(x_\tau) - h(x^\star)$ and c is a positive constant.

Finally, suppose that $\sum_{k=0}^{\infty} \nu_k \leq \bar{\nu}$ where $\bar{\nu}$ is a non-negative constant, and that the initial point $x_0 \in \mathbb{R}^p$ and the constants c and $\bar{\nu}$ satisfy

$$c\phi(l_0) + \|x_0 - x^\star\| + \bar{\nu} < r,$$

where $\mathcal{B}(x^\star, r) \subset U$.

Then the sequence $\{x_k\}$ is finite, i.e. $\sum_{k=0}^{\infty} \|x_{k+1} - x_k\| < \infty$, and thus, convergent.

Proof. We prove the result by induction. Since $\|x_0 - x^\star\| \leq r$ and $h(x^\star) \leq h(x_0) < h(x^\star) + \eta$ the KL inequality holds at x_0 and (3.1.1) holds for $\tau = 0$. Let us assume that $x_k \in \mathcal{B}(x^\star, r)$ up to and including some index $\tau > 0$, which implies that the KL inequality and by extension (3.1.1) hold for all $k \leq \tau$. We apply the triangle inequality twice to

obtain

$$\begin{aligned}\|x_{\tau+1}-x^\star\| &\leq \|x_{\tau+1}-x_0\|+\|x_0-x^\star\| \\ &= \left\|\sum_{j=0}^{\tau}(x_{j+1}-x_j)\right\|+\|x_0-x^\star\| \\ &\leq \sum_{j=0}^{\tau}\|x_{j+1}-x_j\|+\|x_0-x^\star\|.\end{aligned}$$

Applying Eq. (3.1.1) to the preceding relation and evaluating the resulting telescoping sum yields

$$\begin{aligned}\|x_{\tau+1}-x^\star\| &\leq c\left(\phi\left(l_0\right)-\phi\left(l_{\tau+1}\right)\right)+\sum_{j=0}^{\tau}\nu_j+\|x_0-x^\star\| \\ &\leq c\phi\left(l_0\right)+\|x_0-x^\star\|+\bar{\nu}<r.\end{aligned}$$

Hence, $x_{\tau+1}\in U$, which in turn implies that $x_k\in U$ for all $k\geq 0$. Combined with $h(x^\star)\leq h(x_k)<h(x^\star)+\eta$ for all $k\geq 0$, we conclude that the KL inequality holds for all $k\geq 0$ and thus we can sum Eq. (3.1.1) from $k=0$ to infinity to obtain

$$\sum_{k=0}^{\infty}\|x_{k+1}-x_k\|\leq c\phi(l_0)+\bar{\nu}<\infty.$$

Hence, the sequence $\{x_k\}$ is finite and Cauchy (convergent). □

3.1.2.1. First order guarantees for NEAR-DGD^t. We now present our theoretical results on the convergence of NEAR-DGD^t, i.e. $t(k)=t$ in (1.0.5) for some $t\in\mathbb{N}^+$. We introduce the following Lyapunov function, which will play a key role in our analysis,

$$\mathcal{L}_t\left(\mathbf{y}\right)=\mathbf{f}\left(\mathbf{Z}^t\mathbf{y}\right)+\frac{1}{2\alpha}\|\mathbf{y}\|_{\mathbf{Z}^t}^2-\frac{1}{2\alpha}\|\mathbf{y}\|_{\mathbf{Z}^{2t}}^2. \tag{3.1.2}$$

Assumption 3.1.5. ***(KL Lyapunov function)*** *The Lyapunov function* $\mathcal{L}_t : \mathbb{R}^{np} \to \mathbb{R}$ *is a KL function.*

Assumption 3.1.5 covers a broad range of functions, including real analytic, semialgebraic and globally subanalytic functions (see [**8**] for more details). For instance, if the function $\mathbf{f}$ is real analytic, then $\mathcal{L}_t$ is the sum of real analytic functions and by extension KL.

We note that using (3.1.2), we can express the $\mathbf{x}_k$ iterates of NEAR-DGD^t as,

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha \nabla \mathcal{L}_t \left(\mathbf{y}_k \right). \tag{3.1.3}$$

We will demonstrate that the sequence of $\{\mathbf{y}_k\}$ iterates of NEAR-DGD^t generated by Eq. (1.0.6) converges to a critical point of the Lyapunov function $\mathcal{L}_t$ (3.1.2). We begin our analysis by showing that the sequence $\{\mathcal{L}_t(\mathbf{y}_k)\}$ is non-increasing in the following Lemma.

Lemma 3.1.6. ***(Sufficient Descent)*** *Let* $\{\mathbf{y}_k\}$ *be the sequence of NEAR-DGD*^{t} *iterates generated by* (1.0.6) *and suppose that the steplength* α *satisfies* $\alpha < 2/L$*, where* L *is defined in Assumption 3.1.2. Then the following inequality holds for the sequence* $\{\mathcal{L}_t(\mathbf{y}_k)\}$*,*

$$\mathcal{L}_t \left(\mathbf{y}_{k+1} \right) \leq \mathcal{L}_t \left(\mathbf{y}_k \right) - \rho \left\| \mathbf{y}_{k+1} - \mathbf{y}_k \right\|^2,$$

where $\rho = (2\alpha)^{-1} \min_i \left(\lambda_i^t(\mathbf{Z}) \left(1 + (1 - \alpha L) \, \lambda_i^t(\mathbf{Z}) \right) \right) > 0$.

Proof. Combining (1.0.5) and (3.1.2), we obtain for $k \geq 0$,

$$\mathcal{L}_t(\mathbf{y}_k) = \mathbf{f}(\mathbf{x}_k) + \frac{1}{2\alpha}\langle \mathbf{y}_k, \mathbf{x}_k \rangle - \frac{1}{2\alpha}\|\mathbf{x}_k\|^2. \tag{3.1.4}$$

Let $\mathbf{d}_x := \mathbf{x}_{k+1} - \mathbf{x}_k$. Assumption 3.1.2 then yields $\mathbf{f}(\mathbf{x}_{k+1}) \leq \mathbf{f}(\mathbf{x}_k) + \langle \nabla \mathbf{f}(\mathbf{x}_k), \mathbf{d}_x \rangle + \frac{L}{2}\|\mathbf{d}_x\|^2 = \mathbf{f}(\mathbf{x}_k) - \frac{1}{\alpha}\langle \mathbf{y}_{k+1} - \mathbf{x}_k, \mathbf{d}_x \rangle + \frac{L}{2}\|\mathbf{d}_x\|^2$, where we acquire the last equality from (1.0.6). Substituting this relation in (3.1.4) applied at the $(k+1)^{th}$ iteration, we obtain,

$$\begin{aligned}
\mathcal{L}_t(\mathbf{y}_{k+1}) &\leq \mathbf{f}(\mathbf{x}_k) - \frac{1}{2\alpha}\langle \mathbf{y}_{k+1} - \mathbf{x}_k, \mathbf{d}_x \rangle \\
&\quad + \frac{L}{2}\|\mathbf{d}_x\|^2 + \frac{1}{2\alpha}\langle \mathbf{y}_{k+1}, \mathbf{x}_{k+1} \rangle - \frac{1}{2\alpha}\|\mathbf{x}_{k+1}\|^2 \\
&= \mathcal{L}_t(\mathbf{y}_k) - \frac{1}{2\alpha}\langle \mathbf{y}_k, \mathbf{x}_k \rangle + \frac{1}{2\alpha}\|\mathbf{x}_k\|^2 \\
&\quad - \frac{1}{\alpha}\langle \mathbf{y}_{k+1} - \mathbf{x}_k, \mathbf{d}_x \rangle + \frac{L}{2}\|\mathbf{d}_x\|^2 + \frac{1}{2\alpha}\langle \mathbf{y}_{k+1}, \mathbf{x}_{k+1} \rangle - \frac{1}{2\alpha}\|\mathbf{x}_{k+1}\|^2,
\end{aligned}$$

where we obtain the equality after further application of (3.1.4). After setting $\mathbf{d}_y := \mathbf{y}_{k+1} - \mathbf{y}_k$ and re-arranging the terms, we obtain,

$$\begin{aligned}
\mathcal{L}_t(\mathbf{y}_{k+1}) &\leq \mathcal{L}_t(\mathbf{y}_k) - \frac{1}{2\alpha}\langle \mathbf{y}_k, \mathbf{x}_k \rangle \\
&\quad + \frac{1}{\alpha}\langle \mathbf{y}_{k+1}, \mathbf{x}_k \rangle - \frac{1}{2\alpha}\langle \mathbf{y}_{k+1}, \mathbf{x}_{k+1} \rangle - \left(\frac{1}{2\alpha} - \frac{L}{2}\right)\|\mathbf{d}_x\|^2 \\
&= \mathcal{L}_t(\mathbf{y}_k) - \frac{1}{2\alpha}\|\mathbf{y}_k\|^2_{\mathbf{Z}^t} \\
&\quad + \frac{1}{\alpha}\langle \mathbf{y}_{k+1}, \mathbf{y}_k \rangle_{\mathbf{Z}^t} - \frac{1}{2\alpha}\|\mathbf{y}_{k+1}\|^2_{\mathbf{Z}^t} - \left(\frac{1}{2\alpha} - \frac{L}{2}\right)\|\mathbf{d}_x\|^2 \\
&= \mathcal{L}_t(\mathbf{y}_k) - \frac{1}{2\alpha}\|\mathbf{d}_y\|^2_{\mathbf{Z}^t} - \left(\frac{1}{2\alpha} - \frac{L}{2}\right)\|\mathbf{d}_x\|^2.
\end{aligned}$$

Let $\mathbf{H} := (2\alpha)^{-1}\mathbf{Z}^t\left(I + (1 - \alpha L)\mathbf{Z}^t\right)$, which is a positive definitionnite matrix due to Assumption 3.1.1 and the fact that $\alpha < 2/L$. Moreover, $\|\mathbf{d}_x\|^2 = \|\mathbf{d}_y\|^2_{\mathbf{Z}^{2t}}$ by Eq. (1.0.5).

We can then re-write the immediately previous relation as $\mathcal{L}_t(\mathbf{y}_{k+1}) \leq \mathcal{L}_t(\mathbf{y}_k) - \|\mathbf{y}_{k+1} - \mathbf{y}_k\|^2_{\mathbf{H}}$. Applying the definition of $\rho = \lambda_1(\mathbf{H})$ concludes the proof. $\square$

An important consequence of Lemma 3.1.6 is that NEAR-DGD^t can tolerate a bigger range of steplengths than previously indicated ($\alpha < 2/L$ vs. $\alpha < 1/L$ in [**13**]). Moreover, Lemma 3.1.6 implies that the sequence $\{\mathcal{L}_t(\mathbf{y}_k)\}$ is upper bounded by $\mathcal{L}_t(\mathbf{y}_0)$. We use this fact to prove that the iterates of NEAR-DGD^t are also bounded in the next Lemma.

Lemma 3.1.7. ***(Boundedness)*** *Let $\{\mathbf{x}_k\}$ and $\{\mathbf{y}_k\}$ be the sequences of NEAR-DGD^t ($t(k) = t$) iterates generated by* (1.0.5) *and* (1.0.6)*, respectively, from initial point $\mathbf{y}_0$ and under steplength $\alpha < 2/L$. Then the following hold: i) the sequence $\{\mathcal{L}_t(\mathbf{y}_k)\}$ is lower bounded, and ii) there exist universal positive constants B_x and B_y such that $\|\mathbf{x}_k\| \leq B_x$ and $\|\mathbf{y}_{k+1}\| \leq B_y$ for all $k \geq 0$ and $t \in \mathbb{N}^+$.*

Proof. By Assumption 3.1.3, the function $\mathbf{f}$ is lower bounded and therefore $\mathcal{L}_t$ is also lower bounded (sum of lower bounded functions). This proves the first claim of this Lemma.

To prove the second claim, we first notice that Lemma 3.1.6 implies that the sequence $\{\mathcal{L}_t(\mathbf{y}_k)\}$ is upper bounded by $\mathcal{L}_t(\mathbf{y}_0)$. Let us define the set $\mathcal{X}_0 := \{\mathbf{Z}^t\mathbf{y}_0, t \in \mathbb{N}^+\}$. The set $\mathcal{X}_0$ is compact, since $\|\mathbf{Z}^t\mathbf{y}_0\| \leq \|\mathbf{y}_0\|$ for all $t \in \mathbb{N}^+$ due to the non-expansiveness of $\mathbf{Z}$. Hence, by the continuity of $\mathbf{f}$ and the Weierstrass Extreme Value Theorem, there exists $\hat{\mathbf{x}}_0 \in \mathcal{X}_0$ such that $\mathbf{f}(\mathbf{x}_0) \leq \mathbf{f}(\hat{\mathbf{x}}_0)$ for all $\mathbf{x}_0 \in \mathcal{X}_0$. Moreover, Assumption 3.1.1 yields $\|\mathbf{y}_0\|^2_{\mathbf{Z}^t(I-\mathbf{Z}^t)} \leq \|\mathbf{y}_0\|^2$ for all positive integers t, and therefore $\mathcal{L}_t(\mathbf{y}_0) \leq \hat{\mathcal{L}}$ for all $t \in \mathbb{N}^+$, where $\hat{\mathcal{L}} = \mathbf{f}(\hat{\mathbf{x}}_0) + (2\alpha)^{-1}\|\mathbf{y}_0\|^2$.

Since $\hat{\mathcal{L}} \geq \mathcal{L}_t(\mathbf{y}_0) \geq \mathcal{L}_t(\mathbf{y}_k) \geq \mathbf{f}(\mathbf{Z}^t\mathbf{y}_k) = \mathbf{f}(\mathbf{x}_k)$ for all $k \geq 0$ and $t > 0$, the sequence $\{\mathbf{f}(\mathbf{x}_k)\}$ is upper bounded. Hence, by Assumption 3.1.3, there exists positive constant B_x such that $\|\mathbf{x}_k\| \leq B_x$ for $k \geq 0$ and $t > 0$. Moreover, Assumption 3.1.2 yields $\mathbf{f}(\mathbf{y}_{k+1}) \leq \mathbf{f}(\mathbf{x}_k) + \langle \nabla \mathbf{f}(\mathbf{x}_k), \mathbf{y}_{k+1} - \mathbf{x}_k \rangle + \frac{L}{2}\|\mathbf{y}_{k+1} - \mathbf{x}_k\|^2 = \mathbf{f}(\mathbf{x}_k) - \alpha\|\nabla \mathbf{f}(\mathbf{x}_k)\|^2 + \frac{\alpha^2 L}{2}\|\nabla \mathbf{f}(\mathbf{x}_k)\|^2 = \mathbf{f}(\mathbf{x}_k) - \alpha\left(1 - \frac{\alpha L}{2}\right)\|\nabla \mathbf{f}(\mathbf{x}_k)\|^2 \leq \mathbf{f}(\mathbf{x}_k)$, where we obtain the first equality from (1.0.6) and last inequality from the fact that $\alpha < 2/L$. This relation combined with Assumption 3.1.3 implies that there exists constant $B_y > 0$ such that $\|\mathbf{y}_{k+1}\| \leq B_y$ for $k > 0$ and $t > 0$, which concludes the proof. □

Next, we use Lemma 3.1.7 to show that the distance between the local iterates generated by NEAR-DGD^t and their average is bounded.

Lemma 3.1.8. ***(Bounded distance to mean)*** *Let $x_{i,k}$ and $y_{i,k}$ be the local NEAR-DGD^t iterates produced under steplength $\alpha < 2/L$ by* (1.0.5) *and* (1.0.6)*, respectively, and define the average iterates $\bar{x}_k := \frac{1}{n}\sum_{i=1}^n x_{i,k}$ and $\bar{y}_k := \frac{1}{n}\sum_{i=1}^n y_{i,k}$. Then the distance between the local and the average iterates is bounded for $i = 1, ..., n$ and $k = 1, 2, ...,$ i.e.*

$$\|x_{i,k} - \bar{x}_k\| \leq \beta^t B_y, \text{ and } \|y_{i,k} - \bar{y}_k\| \leq B_y,$$

where B_y is a positive constant defined in Lemma 3.1.7.

Proof. Multiplying both sides of (1.0.5) with $\mathbf{M} = \left(\frac{1_n 1_n'}{n} \otimes I_p\right)$ yields $\bar{x}_k = \bar{y}_k$. Moreover, we observe that $\|\mathbf{v}_k - \mathbf{M}\mathbf{v}_k\|^2 = \sum_{i=1}^n \|v_{i,k} - \bar{v}_k\|^2$ for any vector $\mathbf{v} \in \mathbb{R}^{np}$. Hence,

$$\begin{aligned}\|x_{i,k} - \bar{x}_k\| = \|x_{i,k} - \bar{y}_k\| &\leq \|\mathbf{x}_k - \mathbf{M}\mathbf{y}_k\| \\ &\leq \left\|\mathbf{Z}^t\mathbf{y}_k - \mathbf{M}\mathbf{y}_k\right\| \leq \beta^t\|\mathbf{y}_k\|,\end{aligned}$$

where we derive the last inequality from the spectral properties of $\mathbf{Z}$ and $\mathbf{M}$ (we note that the matrix $1_n 1_n'/n$ has a single non-zero eigenvalue at 1 associated with the eigenvector 1_{np}).

Similarly, for the local iterates $y_{i,k}$ we obtain,

$$\|y_{i,k} - \bar{y}_k\| \leq \|\mathbf{y}_k - \mathbf{M}\mathbf{y}_k\| = \|(I - \mathbf{M})\,\mathbf{y}_k\| \leq \|\mathbf{y}_k\|.$$

Applying Lemma 3.1.7 to the two preceding inequalities completes the proof. □

We are now ready to state the first Theorem of this section, namely that there exists a subsequence of $\{\mathbf{y}_k\}$ that converges to a critical point of $\mathcal{L}_t$.

Theorem 3.1.9. ***(Subsequence convergence)*** *Let $\{\mathbf{y}_k\}$ be the sequence of NEAR-DGD^t iterates generated by (1.0.6) with steplength $\alpha < 2/L$. Then $\{\mathbf{y}_k\}$ has a convergent subsequence whose limit point is a critical point of (3.1.2).*

Proof. By Lemma 3.1.7, the sequence $\{\mathbf{y}_k\}$ is bounded and therefore there exists a convergent subsequence $\{\mathbf{y}_{k_s}\}_{s\in\mathbb{N}} \to \mathbf{y}^\infty$ as $s \to \infty$. In addition, recursive application of Lemma 3.1.6 over iterations $0, 1, ..., k$ yields,

$$\mathcal{L}_t(\mathbf{y}_k) \leq \mathcal{L}_t(\mathbf{y}_0) - \rho \sum_{j=0}^{k-1} \|\mathbf{y}_{j+1} - \mathbf{y}_j\|^2,$$

where the sequence $\{\mathcal{L}_t(\mathbf{y}_k)\}$ is non-increasing and bounded from below by Lemmas 3.1.6 and 3.1.7.

Hence, $\{\mathcal{L}_k(\mathbf{y}_k)\}$ converges and the above relation implies that $\sum_{k=1}^{\infty} \|\mathbf{y}_{k+1} - \mathbf{y}_k\|^2 < +\infty$ and $\|\mathbf{y}_{k+1} - \mathbf{y}_k\| \to 0$. Moreover, $\|\mathbf{x}_{k+1} - \mathbf{x}_k\| = \|\mathbf{y}_{k+1} - \mathbf{y}_k\|_{\mathbf{Z}^{2t}} \leq \|\mathbf{y}_{k+1} - \mathbf{y}_k\|$

by the non-expansiveness of $\mathbf{Z}$ and thus $\|\mathbf{x}_{k+1} - \mathbf{x}_k\| \to 0$. Finally, Eq. 3.1.3 yields $\|\nabla \mathcal{L}_t(\mathbf{y}_k)\| = \alpha^{-1}\|\mathbf{x}_{k+1} - \mathbf{x}_k\| \to 0$. We conclude that $\|\nabla \mathcal{L}_t(\mathbf{y}_{k_s})\| \to 0$ as $s \to \infty$ and therefore $\nabla \mathcal{L}_t(\mathbf{y}^\infty) = \mathbf{0}$. □

We note that Assumption 3.1.5 is not necessary for Theorem 3.1.9 to hold. However, Theorem 3.1.9 does not guarantee the convergence of NEAR-DGD^t; we will need Assumption 3.1.5 to prove that NEAR-DGD^t converges in Theorem 3.1.11. Before that, we introduce the following two preliminary Lemmas that hold only under Assumption 3.1.5.

Lemma 3.1.10. ***(Bounded difference under the KL property)*** *Let $\mathbf{x}_k$ and $\mathbf{y}_k$ be the NEAR-DGD^t iterates generated by* (1.0.5) *and* (1.0.6)*, respectively, under steplength $\alpha < 2/L$. Moreover, suppose that the KL inequality with respect to some point $\mathbf{y}^\star \in \mathbb{R}^{np}$ holds at $\mathbf{y}_k$, i.e.,*

$$\phi'(\mathcal{L}_t(\mathbf{y}_k) - \mathcal{L}_t(\mathbf{y}^\star))\|\nabla \mathcal{L}_t(\mathbf{y}_k)\| \geq 1. \tag{3.1.5}$$

Then the following relation holds,

$$\|\mathbf{v}_{k+1} - \mathbf{v}_k\| \leq \frac{1}{\alpha\rho}\left(\phi(l_k) - \phi(l_{k+1})\right),$$

where $\|\mathbf{v}_{k+1} - \mathbf{v}_k\|$ can be $\|\mathbf{x}_{k+1} - \mathbf{x}_k\|$ or $\|\mathbf{y}_{k+1} - \mathbf{y}_k\|$ and $l_k := \mathcal{L}_t(\mathbf{y}_k) - \mathcal{L}_t(\mathbf{y}^\star)$.

Proof. Lemma 3.1.6 yields $\rho\|\mathbf{y}_{k+1} - \mathbf{y}_k\|^2 \leq \mathcal{L}_t(\mathbf{y}_k) - \mathcal{L}_t(\mathbf{y}_{k+1}) = l_k - l_{k+1}$ for $k \geq 0$. We can multiply both sides of this relation with $\phi'(l_k) > 0$ to obtain $\rho\phi'(l_k)\|\mathbf{y}_{k+1} - \mathbf{y}_k\|^2 \leq -\phi'(l_k)(l_{k+1} - l_k) \leq \phi(l_k) - \phi(l_{k+1})$, where we derive the last inequality from the concavity of ϕ. In addition, using Eq. 3.1.3, we can re-write (3.1.5) as $\alpha^{-1}\phi'(l_k)\|\mathbf{x}_{k+1} - \mathbf{x}_k\| \geq 1$.

Combining these relations, we acquire,

$$\frac{\alpha\rho \left\|\mathbf{y}_{k+1}-\mathbf{y}_k\right\|^2}{\left\|\mathbf{x}_{k+1}-\mathbf{x}_k\right\|} \leq \phi\left(l_k\right)-\phi\left(l_{k+1}\right).$$

Observing that $\left\|\mathbf{x}_{k+1}-\mathbf{x}_k\right\| \leq \left\|\mathbf{y}_{k+1}-\mathbf{y}_k\right\|$ due to the non-expansiveness of $\mathbf{Z}$ and re-arranging the terms of the relation above yields the final result. □

Next, we combine our previous results to prove the global convergence of the $\mathbf{y}_k$ iterates of NEAR-DGD^t in Theorem 3.1.11.

Theorem 3.1.11. ***(Global Convergence)*** *Let $\{\mathbf{y}_k\}$ be the sequence of NEAR-DGD^t iterates produced by (1.0.6) under steplength $\alpha < 2/L$ and let $\mathbf{y}^\infty$ be a limit point of a convergent subsequence of $\{\mathbf{y}_k\}$ as defined in Theorem 3.1.9.*

Then under Assumption 3.1.5 the following statements hold: i) there exists an index $k_0 \in \mathbb{N}^+$ such that the KL inequality with respect to $\mathbf{y}^\infty$ holds for all $k \geq k_0$, and ii) the sequence $\{\mathbf{y}_k\}$ converges to $\mathbf{y}^\infty$.

Proof. We first observe that by Lemma 3.1.6 the sequence $\{\mathcal{L}_t(\mathbf{y}_k)\}$ is non-increasing, and therefore $\mathcal{L}_t(\mathbf{y}^\infty) \leq \mathcal{L}_t(\mathbf{y}_k)$ for all $k \geq 0$. Moreover, by Lemma 3.1.10 the inequality $\left\|\mathbf{y}_{k+1}-\mathbf{y}_k\right\| \leq (\alpha\rho)^{-1}\left(\phi\left(l_k\right)-\phi\left(l_{k+1}\right)\right)$ holds for all $\mathbf{y}_k$ satisfying the KL inequality with respect to $\mathbf{y}^\infty$, where $l_k = \mathcal{L}_t(\mathbf{y}_k) - \mathcal{L}_t(\mathbf{y}^\infty)$. If Assumption 3.1.5 holds, then the objects U and η in Def. 1 exist and by the continuity of ϕ, it is possible to find an index k_0 satisfying

the following relations,

$$(\alpha\rho)^{-1}\phi\left(\mathcal{L}_t(\mathbf{y}_{k_0}) - \mathcal{L}_t(\mathbf{y}^\infty)\right) + \|\mathbf{y}_{k_0} - \mathbf{y}^\infty\| < r,$$

$$\mathcal{L}_t(\mathbf{y}_k) \in [\mathcal{L}_t(\mathbf{y}^\infty), \mathcal{L}_t(\mathbf{y}^\infty) + \eta), \quad \forall k \geq k_0,$$

where $\mathcal{B}(\mathbf{y}^\infty, r) \subset U$.

The global convergence of NEAR-DGD^t follows from applying Lemma 3.1.4 on the sequence $\{\mathbf{y}_k\}_{k\geq k_0}$ with $h = \mathcal{L}_t$, $c = (\alpha\rho)^{-1}$, $\{\nu_k\} = \{\bar{\nu}\} = \{0\}$ and $\mathbf{y}^\star = \mathbf{y}^\infty$. Since $\mathbf{y}^\infty$ is the limit point of a subsequence of $\{\mathbf{y}_k\}$ and $\{\mathbf{y}_k\}$ is convergent, we conclude that $\{\mathbf{y}_k\} \to \mathbf{y}^\infty$. □

Since $\mathbf{Z}$ is a non-singular matrix, Theorem 3.1.11 implies that the sequence $\{\mathbf{x}_k\}$ also converges. Moreover, using arguments similar to [**7**], we can prove the following result on the convergence rate of $\{\mathbf{x}_k\}$.

Lemma 3.1.12. ***(Rates)*** *Let $\{\mathbf{x}_k\}$ be the sequence of iterates produced by (1.0.5), $\mathbf{x}^\infty = \mathbf{Z}^t\mathbf{y}^\infty$ where $\mathbf{y}^\infty$ is the limit point of the sequence $\{\mathbf{y}_k\}$ and suppose $\phi(s) = cs^{1-\theta}$ in Assumption 3.1.5 for some constant $c > 0$ and $\theta \in [0, 1)$ (for a discussion on ϕ, we direct readers to [**8**]). Then the following hold:*

(1) *If $\theta = 0$, $\{\mathbf{x}_k\}$ converges in a finite number of iterations.*

(2) *If $\theta \in (0, 1/2]$, then constants $c > 0$ and $Q \in [0, 1)$ exist such that $\|\mathbf{x}_k - \mathbf{x}^\infty\| \leq cQ^k$.*

(3) *If $\theta \in (1/2, 1)$, then there exists a constant $c > 0$ such that $\|\mathbf{x}_k - \mathbf{x}^\infty\| \leq ck^{-\frac{1-\theta}{2\theta-1}}$.*

Proof. *i)* $\theta = 0$: From the definition of ϕ and $\theta = 0$ we have $\phi'(l_k) = c(1-\theta)l_k^{-\theta} = c$. Let $I := \{k \in \mathbb{N} : \mathbf{x}_{k+1} \neq \mathbf{x}_k\}$ (by the non-singularity of $\mathbf{Z}$, it also follows that $\mathbf{y}_{k+1} \neq \mathbf{y}_k$

for $k \in I$). Then for large k the KL inequality holds at $\mathbf{y}_k$ and we obtain $\|\nabla \mathcal{L}_t(\mathbf{y}_k)\| \geq c^{-1}$, or equivalently by (3.1.3), $\|\mathbf{x}_{k+1} - \mathbf{x}_k\| \geq \alpha c^{-1}$. Application of Lemma 3.1.6 combined with the fact that $\|\mathbf{x}_{k+1} - \mathbf{x}_k\| \leq \|\mathbf{y}_{k+1} - \mathbf{y}_k\|$ yields $\mathcal{L}_t(\mathbf{y}_{k+1}) \leq \mathcal{L}_t(\mathbf{y}_k) - \rho\alpha^2 c^{-2}$. Given the convergence of the sequence $\{\mathcal{L}_t\}$, we conclude that the set I is finite and the method converges in a finite number of steps.

ii) $\theta \in (0, 1)$: Let $S_k := \sum_{j=k}^{\infty} \|\mathbf{x}_{j+1} - \mathbf{x}_j\|$ where $\mathbf{x}^\infty = \mathbf{Z}^t \mathbf{y}^\infty$. Since $\|\mathbf{x}_k - \mathbf{x}^\infty\| \leq S_k$, it suffices to bound S_k. Using Lemma 3.1.10 with $\mathbf{y}^\star = \mathbf{y}^\infty$ and for $k \geq k_0$, where k_0 is defined in Theorem 3.1.11, we obtain,

$$S_k \leq \frac{1}{\alpha\rho} \sum_{j=k}^{\infty} (\phi(l_j) - \phi(l_{j+1})) = \frac{1}{\alpha\rho} \phi(l_k) = \frac{1}{\nu} l_k^{1-\theta}, \tag{3.1.6}$$

where $\nu = \alpha\rho/c$.

Moreover, Eq. 3.1.3 yields $\|\nabla \mathcal{L}_t(\mathbf{y}_k)\| = \alpha^{-1} \|\mathbf{x}_{k+1} - \mathbf{x}_k\| = \alpha^{-1}(S_k - S_{k+1})$. Using this relation and the definition of ϕ, we can express the KL inequality as,

$$\mu l_k^{-\theta} (S_k - S_{k+1}) \geq 1, \tag{3.1.7}$$

where $\mu = \alpha^{-1} c(1 - \theta)$.

If $\theta \in (0, 1/2]$, raising both sides of the preceding inequality to the power of $\gamma = \frac{1-\theta}{\theta} > 1$ and re-arranging the terms yields $\mu^\gamma (S_k - S_{k+1})^\gamma \geq l_k^{1-\theta}$. Due to the fact that $S_k - S_{k+1} = \alpha \|\nabla \mathcal{L}_t(\mathbf{y}_k)\| \to 0$, there exists some index k such that $S_k - S_{k+1} > (S_k - S_{k+1})^\gamma$ and $\mu^\gamma (S_k - S_{k+1}) \geq l_k^{1-\theta}$. Combining this relation with (3.1.6), we obtain $\nu S_k \leq \mu^\gamma (S_k - S_{k+1}) \Leftrightarrow S_{k+1} \leq \left(1 - \frac{\nu}{\mu^\gamma}\right) S_k$.

If $\theta \in (1/2, 1)$, raising both sides of (3.1.6) to the power of $\theta/(1-\theta) > 1$ yields $S_k^{\theta/(1-\theta)} \leq \nu^{-\theta/(1-\theta)} l_k^{\theta}$. After substituting this relation in (3.1.7) and re-arranging we obtain $1 \leq C\,(S_k - S_{k+1})\,(S_k^{\theta/(1-\theta)})^{-1}$, where $C = \mu\nu^{-\theta/(1-\theta)}$. Define $h : (0, +\infty) \to \mathbb{R}$ to be $h(s) = s^{-\theta/(1-\theta)}$. The preceding relation then yields $1 \leq C(S_k - S_{k+1})h(S_k) \leq C\int_{S_{k+1}}^{S_k} h(s)ds = C\zeta^{-1}\left(S_k^{\zeta} - S_{k+1}^{\zeta}\right)$, where $\zeta = (1-2\theta)/(1-\theta) < 0$. After setting $\tilde{C} = -C^{-1}\zeta > 0$ and re-arranging, we obtain $\tilde{C} \leq S_{k+1}^{\zeta} - S_k^{\zeta}$. Summing this relation over iterations $k = k_0, ..., t-1$ yields $(t-k_0)\tilde{C} \leq S_t^{\zeta} - S_{k_0}^{\zeta} \Leftrightarrow S_t \leq \left((t-k_0)\tilde{C} + S_{k_0}^{\zeta}\right)^{1/\zeta} \leq ct^{1/\zeta}$, for some $c > 0$. □

We conclude this subsection with one more result on the distance to optimality of the local $x_{i,k}$ iterates of NEAR-DGD^t and their average $\bar{x}_k = \frac{1}{n}\sum_{i=1}^n x_{i,k}$ as $k \to \infty$.

Corollary 3.1.13. ***(Distance to optimality)*** *Suppose that* $\{\mathbf{y}_k\} \to \mathbf{y}^{\infty}$ *and let* $\mathbf{x}^{\infty} = \mathbf{Z}^t\mathbf{y}^{\infty}$. *Moreover, let* $\bar{x}^{\infty} = \bar{y}^{\infty} = \frac{1}{n}\sum_{i=1}^n x_i^{\infty}$. *Then* $\bar{x}^{\infty}$ *is an approximate critical point of* f,

$$\|\nabla f(\bar{x}^{\infty})\| \leq \beta^t\sqrt{n}LB_y$$

where B_y *is a positive constant defined in Lemma 3.1.7.*

Proof. We observe that $\mathbf{M}\nabla\mathbf{f}(\mathbf{M}\mathbf{y}^{\infty}) = \frac{1}{n}\cdot 1_n \otimes \nabla f(\bar{y}^{\infty})$ and hence $\|\mathbf{M}\nabla\mathbf{f}(\mathbf{M}\mathbf{y}^{\infty})\| = n^{-1}\|1_n \otimes \nabla f(\bar{y}^{\infty})\| = (\sqrt{n})^{-1}\|\nabla f(\bar{y}^{\infty})\|$, where we obtain the last equality due to the fact that $\|1_n \otimes v\|^2 = n\|v\|^2$ for any vector v.

Moreover, $\mathbf{y}^{\infty}$ is a critical point of (3.1.2) and therefore satisfies $\nabla\mathcal{L}_t(\mathbf{y}^{\infty}) = \mathbf{Z}^t\nabla\mathbf{f}(\mathbf{Z}^t\mathbf{y}^{\infty}) + \frac{1}{\alpha}\mathbf{Z}^t\mathbf{y}^{\infty} - \frac{1}{\alpha}\mathbf{Z}^{2t}\mathbf{y}^{\infty} = \mathbf{0}$. From the double stochasticity of $\mathbf{Z}$, multiplying the above relation with $\mathbf{M}$ yields $\mathbf{M}\nabla\mathcal{L}_t(\mathbf{y}^{\infty}) = \mathbf{M}\nabla\mathbf{f}(\mathbf{Z}^t\mathbf{y}^{\infty}) = \mathbf{0}$. After combining all the preceding results,

we obtain,

$$\|\nabla f(\bar{x}^\infty)\| = \sqrt{n}\|\mathbf{M}\nabla\mathbf{f}(\mathbf{M}\mathbf{y}^\infty) - \mathbf{M}\nabla\mathbf{f}(\mathbf{Z}^t\mathbf{y}^\infty)\|$$

$$\leq \sqrt{n}L\|\mathbf{M}\mathbf{y}^\infty - \mathbf{Z}^t\mathbf{y}^\infty\| \leq \beta^t\sqrt{n}L\|\mathbf{y}^\infty\|,$$

where used the spectral properties of $\mathbf{M}$ and Assumption 3.1.2 to get the first inequality and the spectral properties of $\mathbf{Z}$ to get the second inequality. Applying Lemma 3.1.7 yields the result of this Corollary. □

We have now concluded our work on the first order guarantees for the NEAR-DGD^t variant. Next, we provide similar guarantees for NEAR-DGD$^+$; while our proof has a similar structure to our proof for NEAR-DGD^t, the time-variant nature of NEAR-DGD$^+$ requires the design of a new Lyapunov function. It is also necessary to establish that the iterates of NEAR-DGD$^+$ achieve consensus exponentially fast, which is not the case with NEAR-DGD^t.

3.1.2.2. First order guarantees for NEAR-DGD$^+$. We now analyze the convergence of NEAR-DGD$^+$, i.e. $t(k) = k$ in (1.0.5). Specifically, we will demonstrate that the sequence of $\{\mathbf{x}_k\}$ iterates of NEAR-DGD$^+$ generated by Eq. (1.0.5) converges to a critical point of the function $f : \mathbb{R}^p \to \mathbb{R}$ of Problem 1.0.1 locally at the agent level. Similar results for nonconvex problems have been established for the DGD method (1.0.3) under diminishing steplength [**186**]; however, unlike [**186**] we do not require that the gradients of $\mathbf{f} : \mathbb{R}^{np} \to \mathbb{R}$ are universally bounded.

In the same vein as the previous subsection, we define the following Lyapunov function that we will use thoughout our convergence analysis to track the progress of the NEAR-DGD$^+$ iterates,

$$\mathcal{L}_\infty(\mathbf{y}) = \mathbf{f}(\mathbf{M}\mathbf{y}). \tag{3.1.8}$$

Assumption 3.1.14. ***(KL Lyapunov function (NEAR-DGD$^+$))*** *The Lyapunov function $\mathcal{L}_\infty : \mathbb{R}^{np} \to \mathbb{R}$ is a KL function.*

We begin our analysis by demonstrating that the $\mathbf{x}_k$ iterates of NEAR-DGD$^+$ generated by (1.0.5) achieve consensus with a linear rate.

Lemma 3.1.15. ***(Bounded distance to consensus (NEAR-DGD$^+$))*** *Suppose that $t(k) = k$ in (1.0.5) and let $A_k := (\mathbf{Z}^k - \mathbf{M})\mathbf{y}_k = (I - \mathbf{M})\mathbf{x}_k$ be the distance to consensus, and $B_k = \mathbf{Z}^k\mathbf{y}_k - \mathbf{u}^\star$, where $\mathbf{u}^\star \in crit_{\mathbf{f}}$ which is guaranteed to be non-empty by Assumption 3.1.3. Then if the steplength in (1.0.6) satisfies $\alpha < (\beta^{-1} - 1)/L$, where β is the second biggest eigenvalue of $\mathbf{W}$, the distance to consensus decays with linear rate, i.e.*

$$\|A_k\| \leq \mu^k C, \text{ where } \mu = \beta(1 + \alpha L) < 1 \text{ and}$$

$$C = \max\left\{ \frac{\|A_0\| + \alpha L \|B_0\|}{1 + \alpha L}, \|B_0\| + \frac{\|\mathbf{u}^\star\|}{\alpha L} \right\}.$$

Proof. Combining (1.0.5) and (1.0.6) at the k^{th} iteration with $t(k) = k$ yields

$$\mathbf{y}_k = \mathbf{Z}^{k-1}\mathbf{y}_{k-1} - \alpha \nabla \mathbf{f}(\mathbf{Z}^{k-1}\mathbf{y}_{k-1}). \tag{3.1.9}$$

We multiply Eq. (3.1.9) with $\mathbf{Z}^k - \mathbf{M}$ and take the norm on both sides to obtain

$$\begin{aligned} \|A_k\| &\leq \|(\mathbf{Z}^k - \mathbf{M})A_{k-1}\| + \alpha\|(\mathbf{Z}^k - \mathbf{M})\nabla\mathbf{f}(\mathbf{Z}^{k-1}\mathbf{y}_{k-1})\| \\ &\leq \beta^k\|A_{k-1}\| + \alpha\beta^k L\|B_{k-1}\|, \end{aligned} \tag{3.1.10}$$

where we used the triangle inequality and the fact that $\mathbf{ZM} = \mathbf{M}$ to get the first inequality, and applied Assumptions 3.1.1 and 3.1.2 to get the second inequality.

For the quantity B_k, Eq. (3.1.9) yields

$$\begin{aligned} \|B_k\| &\leq \|\mathbf{Z}^k B_{k-1}\| + \alpha\|\mathbf{Z}^k\nabla\mathbf{f}(\mathbf{Z}^{k-1}\mathbf{y}_{k-1})\| \\ &\quad + \|(\mathbf{Z}^k - I)\mathbf{u}^\star\| \leq (1+\alpha L)\|B_{k-1}\| + \|\mathbf{u}^\star\|, \end{aligned}$$

where we added and subtracted $\mathbf{Z}^k\mathbf{u}^\star$ and applied the triangle inequality to get the first inequality, and obtained the second inequality from Assumptions 3.1.1 and 3.1.2.

Let $R := \|B_0\| + (\alpha L)^{-1}\|\mathbf{u}^\star\|$. Applying the preceding relation recursively yields

$$\begin{aligned} \|B_k\| &\leq (1+\alpha L)^k\|B_0\| + \|\mathbf{u}^\star\|\sum_{j=0}^{k-1}(1+\alpha L)^j \\ &= (1+\alpha L)^k\|B_0\| + \frac{((1+\alpha L)^k - 1)\|\mathbf{u}^\star\|}{\alpha L} \leq (1+\alpha L)^k R. \end{aligned}$$

Next, we substitute the relation above in (3.1.10) to obtain

$$\|A_k\| \leq \beta^k\|A_{k-1}\| + \alpha\beta^k(1+\alpha L)^{k-1}LR.$$

We will now prove the claim of this lemma by induction. The claim holds trivially for $k = 1$. Assuming $\|A_{k-1}\| \leq \beta^{k-1}(1+\alpha L)^{k-1}C$, the preceding relation yields

$$\|A_k\| \leq \big(\beta(1+\alpha)\big)^k \left(\frac{\beta^{k-1}+\alpha L}{1+\alpha L}\right) C.$$

Since $\beta < 1$, the claim holds for all $k \geq 1$. □

Next, we utilize the diminishing distance to consensus established in Lemma 3.1.15 to show that the sequence of the Lyapunov function values $\{\mathcal{L}_\infty(\mathbf{y}_k)\}$ is non-increasing.

Lemma 3.1.16. ***(Sufficient descent (NEAR-DGD$^+$))*** *For NEAR-DGD$^+$, let the steplength in* (1.0.6) *satisfy*

$$\alpha < \min\left\{\frac{1+\sqrt{5}}{2L}, \frac{\beta^{-1}-1}{L}\right\}$$

Then the following relations hold for $k \geq 0$

$$\mathcal{L}_\infty(\mathbf{y}_{k+1}) \leq \mathcal{L}_\infty(\mathbf{y}_k) + \mu^{2k}C_1 - C_2\|\nabla \mathbf{f}(\mathbf{M}\mathbf{y}_k)\|^2_{\mathbf{M}} \tag{3.1.11a}$$

$$\mathcal{L}_\infty(\mathbf{y}_{k+1}) \leq \mathcal{L}_\infty(\mathbf{y}_k) + \mu^{2k}C_1 - C_3\|\nabla \mathbf{f}(\mathbf{x}_k)\|^2_{\mathbf{M}}, \tag{3.1.11b}$$

where $C_1 = \frac{\alpha L^2 C^2}{2}(1+\alpha L)$, $\mu \in (0,1)$ and $C > 0$ are defined in Lemma 3.1.15, $C_2 = \frac{\alpha}{2}(1+\alpha L - \alpha^2 L^2)$ and $C_3 = \frac{\alpha}{2}\left(1 - \frac{\alpha^2 L^2}{1+\alpha L}\right)$.

Proof. Let $\nabla_{1,k} = \nabla \mathbf{f}(\mathbf{M}\mathbf{y}_k)$ and $\nabla_{2,k} = \nabla \mathbf{f}(\mathbf{x}_k)$. Substituting (1.0.6) in (3.1.8) for the $(k+1)^{th}$ iteration we obtain

$$\begin{aligned}
\mathcal{L}_\infty(\mathbf{y}_{k+1}) &= \mathbf{f}(\mathbf{M}\mathbf{y}_k - \alpha \mathbf{M}\nabla \mathbf{f}(\mathbf{x}_k)) \\
&\leq \mathbf{f}(\mathbf{M}\mathbf{y}_k) - \alpha \langle \nabla_{1,k}, \nabla_{2,k} \rangle_{\mathbf{M}} + \frac{\alpha^2 L}{2} \|\nabla_{2,k}\|_{\mathbf{M}}^2 \\
&\leq \mathcal{L}_\infty(\mathbf{y}_k) + \frac{\alpha}{2}(1+\alpha L)\|\nabla_{1,k} - \nabla_{2,k}\|_{\mathbf{M}}^2 - \frac{\alpha}{2}\|\nabla_{2,k}\|_{\mathbf{M}}^2 \\
&\quad - \frac{\alpha}{2}(1+\alpha L)\|\nabla_{1,k}\|_{\mathbf{M}}^2 + \alpha^2 L \|\nabla_{1,k}\|_{\mathbf{M}} \|\nabla_{2,k}\|_{\mathbf{M}} \\
&= \mathcal{L}_\infty(\mathbf{y}_k) + \frac{\alpha}{2}(1+\alpha L)\|\nabla_{1,k} - \nabla_{2,k}\|_{\mathbf{M}}^2 \\
&\quad - \frac{\alpha}{2}\|\nabla_{1,k}\|_{\mathbf{M}}^2 \left(\frac{\|\nabla_{2,k}\|_{\mathbf{M}}^2}{\|\nabla_{1,k}\|_{\mathbf{M}}^2} - 2\alpha L \frac{\|\nabla_{2,k}\|_{\mathbf{M}}}{\|\nabla_{1,k}\|_{\mathbf{M}}} + (1+\alpha L) \right) \\
&= \mathcal{L}_\infty(\mathbf{y}_k) + \frac{\alpha}{2}(1+\alpha L)\|\nabla_{1,k} - \nabla_{2,k}\|_{\mathbf{M}}^2 \\
&\quad - \frac{\alpha}{2}\|\nabla_{2,k}\|_{\mathbf{M}}^2 \left((1+\alpha L) \frac{\|\nabla_{1,k}\|_{\mathbf{M}}^2}{\|\nabla_{2,k}\|_{\mathbf{M}}^2} - 2\alpha L \frac{\|\nabla_{1,k}\|_{\mathbf{M}}}{\|\nabla_{2,k}\|_{\mathbf{M}}} + 1 \right).
\end{aligned}$$

We first notice that $\|\nabla_{1,k} - \nabla_{2,k}\|_{\mathbf{M}}^2 \leq L^2 \|A_k\|^2 \leq \mu^{2k} L^2 C^2$ by Assumption 3.1.2 and Lemma 3.1.15. Moreover, consider the 2^{nd} degree polynomials $P_1(z) = z^2 - 2\alpha L z + (1+\alpha L)$ and $P_2(z) = (1+\alpha L)z^2 - 2\alpha L z + 1$. If $\alpha < \frac{1+\sqrt{5}}{2L}$, then $4\alpha^2 L^2 - 4(1+\alpha L) < 0$ and $P_1(z), P_2(z) > 0$ for all $z \in \mathbb{R}$ with $\min_z P_1(z) = 2\alpha^{-1} C_2$ and $\min_z P_2(z) = 2\alpha^{-1} C_3$. Applying the definitions of C_1, C_2 and C_3 to the preceding relation yields the final result of this lemma. □

Lemma 3.1.16 combined with Assumption 3.1.3 implies that the average iterates of NEAR-DGD$^+$ form a compact set. Given that distance to consensus diminishes by Lemma 3.1.15, we conclude that the $\mathbf{x}_k$ iterates of NEAR-DGD$^+$ also belong to a compact set for all $k \geq 0$. This result is formally stated in the following Lemma.

Lemma 3.1.17. ***(Boundedness (NEAR-DGD$^+$))*** *Let $\{\mathbf{x}_k\}$ be the sequence of iterates generated by Eq. (1.0.5) with $t(k) = k$ and suppose the steplength in Eq. (1.0.6) satisfies $\alpha < \min\{(1+\sqrt{5})/(2L), (\beta^{-1}-1)/L\}$. Then there exists positive constant B_x^+ such that $\|\mathbf{x}_k\| \leq B_x^+$ for all $k \geq 0$.*

Proof. The triangle inequality yields

$$\begin{aligned}\|\mathbf{x}_k\| &\leq \|\mathbf{x}_k\|_{\mathbf{M}} + \|(I-\mathbf{M})\mathbf{x}_k\| \\ &\leq \|\mathbf{x}_k\|_{\mathbf{M}} + \mu^k C \\ &\leq \|\mathbf{x}_k\|_{\mathbf{M}} + C,\end{aligned}$$

where we invoked Lemma 3.1.15 to obtain the second inequality.

Moreover, Lemma 3.1.16 implies that the sequence $\{\mathcal{L}_\infty(\mathbf{y}_k)\} = \{\mathbf{f}(\mathbf{M}\mathbf{y}_k)\} = \{\mathbf{f}(\mathbf{M}\mathbf{x}_k)\}$ is bounded, and hence by Assumption 3.1.3 the norm $\|\mathbf{x}_k\|_{\mathbf{M}}$ is bounded for all $k \geq 0$. We conclude that a constant $B_x^+ > 0$ exists such that $\|\mathbf{x}_k\| \leq B_x^+$ for all $k \geq 0$. □

In the next Theorem we show that the sequence of iterates of NEAR-DGD$^+$ have at least one subsequence that converges to a critical point of $\mathcal{L}_\infty$. This fact combined with Lemma 3.1.15 implies that the local iterates $x_{i,k}$ of such convergent subsequence of $\{\mathbf{x}_k\}$ converge to critical points of the function f or Problem 1.0.1.

Theorem 3.1.18. ***(Subsequence convergence (NEAR-DGD$^+$))*** *Let $\{\mathbf{x}_k\}$ be the sequence of iterates generated by Eq (1.0.5) with $t(k) = k$ and suppose the steplength in Eq. (1.0.6) satisfies $\alpha < \min\{(1+\sqrt{5})/(2L), (\beta^{-1}-1)/L\}$. Then if the set of critical points of $f(x) = \sum_{i=1}^n f_i(x)$ is non-empty, $\{\mathbf{x}_k\}$ has a subsequence converging to a point $\mathbf{x}^\infty = \bar{x}^\infty \otimes 1_n \in \mathbb{R}^{np}$ where $\bar{x}^\infty \in critf$.*

Proof. By Lemma 3.1.17, the sequence $\{\mathbf{x}_k\}$ is bounded, and thus has at least one subsequence converging to some point $\mathbf{x}^\infty$. By Eq. (3.1.11b) of Lemma 3.1.16 and the convergence of $\{\mathcal{L}_\infty(\mathbf{y}_k)\}$ we have $\|\nabla \mathbf{f}(\mathbf{x}_k)\|_{\mathbf{M}} \to 0$, which yields $\mathbf{M}\nabla \mathbf{f}(\mathbf{x}^\infty) = (n^{-1}\sum_{i=1}^n \nabla f_i(x_i^\infty)) \otimes 1_n \to 0$. Finally, Lemma 3.1.15 guarantees that

$$x_{j,k} \to n^{-1}\sum_{i=1}^n x_i^\infty = \bar{x}^\infty,$$

for all $j \in \{1, ..., n\}$. □

Theorem 3.1.18 does not guarantee the convergence of the iterates of NEAR-DGD$^+$; to establish this result, we first prove the following intermediate Lemma on the distance between two consecutive iterates of the NEAR-DGD$^+$ method under Assumption 3.1.14.

Lemma 3.1.19. ***(Bounded difference (NEAR-DGD$^+$))*** *Suppose that for some index $k \geq 0$, the point $\mathbf{y}_k$ satisfies the KL inequality with respect to some $\mathbf{y}^\star$. Then there exists a positive constant Q such that*

$$\|\mathbf{x}_{k+1} - \mathbf{x}_k\| \leq \frac{\alpha}{C_2}(\phi(l_k) - \phi(l_{k+1})) + \mu^k Q$$

where $l_k = \mathcal{L}_\infty(\mathbf{y}_k) - \mathcal{L}_\infty(\mathbf{y}^\star)$, $C_2 > 0$ is defined in Lemma 3.1.16 and $\mu \in (0,1)$ is defined in Lemma 3.1.15.

Proof. Let $\mathbf{d}_x = \mathbf{x}_{k+1} - \mathbf{x}_k$ and recall that $A_k = (I - \mathbf{M})\mathbf{x}_k$. The triangle inequality yields

$$\begin{aligned} \|\mathbf{d}_x\| &\leq \|\mathbf{d}_x\|_{\mathbf{M}} + \|A_{k+1}\| + \|A_k\| \\ &\leq \alpha\|\nabla \mathbf{f}(\mathbf{x}_k)\|_{\mathbf{M}} + \mu^k(\mu+1)C. \end{aligned} \tag{3.1.12}$$

where we applied Lemma 3.1.15 to get the last inequality.

We proceed by deriving a bound for the first term on the right-hand side of the equation above. Eq. (3.1.11a) of Lemma 3.1.16 directly yields

$$\|\nabla \mathbf{f}(\mathbf{M}\mathbf{y}_k)\|_{\mathbf{M}}^2 \leq C_2^{-1} \left(\mathcal{L}_\infty(\mathbf{y}_k) - \mathcal{L}_\infty(\mathbf{y}_{k+1})\right) + \mu^{2k} C_1 C_2^{-1}.$$

We multiply the preceding relation with $\phi'(l_k) > 0$ to obtain

$$\begin{aligned}\phi'(l_k)\|\nabla \mathbf{f}(\mathbf{M}\mathbf{y}_k)\|_{\mathbf{M}}^2 &\leq -C_2^{-1}\phi'(l_k)(l_{k+1} - l_k) + \mu^{2k} C_1 C_2^{-1}\phi'(l_k) \\ &\leq C_2^{-1}(\phi(l_k) - \phi(l_{k+1})) + \mu^{2k} C_1 C_2^{-1}\phi'(l_k),\end{aligned}$$

where the last inequality follows from the concavity of ϕ.

Note that we can re-write the KL inequality for $\mathcal{L}_\infty$ as $\phi'(l_k)\|\nabla \mathbf{f}(\mathbf{M}\mathbf{y}_k)\|_{\mathbf{M}} \geq 1$. Combining this with the inequality above yields

$$\|\nabla \mathbf{f}(\mathbf{M}\mathbf{y}_k)\|_{\mathbf{M}} \leq C_2^{-1}(\phi(l_k) - \phi(l_{k+1})) + \mu^{2k} C_1 C_2^{-1}\phi'(l_k).$$

Moreover, the following inequality follows from Lemma 3.1.15 and Assumption 3.1.2

$$\begin{aligned}\|\nabla \mathbf{f}(\mathbf{x}_k)\|_{\mathbf{M}} &\leq \|\nabla \mathbf{f}(\mathbf{M}\mathbf{y}_k)\|_{\mathbf{M}} + \|\nabla \mathbf{f}(\mathbf{x}_k) - \nabla \mathbf{f}(\mathbf{M}\mathbf{y}_k)\|_{\mathbf{M}} \\ &\leq \|\nabla \mathbf{f}(\mathbf{M}\mathbf{y}_k)\|_{\mathbf{M}} + \mu^k CL.\end{aligned}$$

Combining the two preceding relations yields

$$\|\nabla \mathbf{f}(\mathbf{x}_k)\|_{\mathbf{M}} \leq C_2^{-1}(\phi(l_k) - \phi(l_{k+1})) + \mu^{2k} C_1 C_2^{-1}\phi'(l_k) + \mu^k CL.$$

Since ϕ is continuously differentiable and l_k is bounded by Lemma 3.1.16, there exists constant B_ϕ such that $\phi'(l_k) < B_\phi$ for all $k \geq 0$. Substituting everything in Eq. (3.1.12) and setting $Q = \mu C_1 C_2^{-1} B_\phi + CL + (\mu + 1)C$ yields the final result. $\square$

We are now ready to combine Lemmas 3.1.4 and 3.1.19 in order to establish the global convergence of the $\{\mathbf{x}_k\}$ iterates of NEAR-DGD$^+$ in the following Theorem.

Theorem 3.1.20. ***(Global convergence (NEAR-DGD$^+$))*** *Let $\{\mathbf{x}_k\}$ be the sequence of NEAR-DGD$^+$ iterates produced by (1.0.5) with $t(k) = k$ and suppose the steplength in Eq. (1.0.6) satisfies $\alpha < \min\{(1+\sqrt{5})/(2L), (\beta^{-1}-1)/L\}$. Moreover, let $\mathbf{x}^\infty$ be a limit point of a convergent subsequence of $\{\mathbf{x}_k\}$ as defined in Theorem 3.1.18. Then under Assumption 3.1.14 the following statements hold: i) there exists an index $k_0 \in \mathbb{N}^+$ such that the KL inequality with respect to $\mathbf{x}^\infty$ holds for all $k \geq k_0$, and ii) the sequence $\{\mathbf{x}_k\}$ converges to $\mathbf{x}^\infty$.*

Proof. We first note that $\mathcal{L}_\infty(\mathbf{y}_k) = \mathcal{L}_\infty(\mathbf{x}_k)$ for all $k \geq 0$ by the construction of $\mathcal{L}_\infty$. Lemma 3.1.16 implies that the sequence $\{\mathcal{L}_\infty(\mathbf{x}_k)\}$ is non-increasing, and therefore $\mathcal{L}_\infty(\mathbf{x}^\infty) \leq \mathcal{L}_\infty(\mathbf{x}_k)$ for all $k \geq 0$. Moreover, by Lemma 3.1.19 the following inequality holds for all $\mathbf{x}_k$ that satisfy the KL inequality with respect to $\mathbf{x}^\infty$,

$$\|\mathbf{x}_{k+1} - \mathbf{x}_k\| \leq \alpha C_2^{-1}(\phi(l_k) - \phi(l_{k+1})) + \mu^k Q,$$

where $l_k = \mathcal{L}_\infty(\mathbf{x}_k) - \mathcal{L}_\infty(\mathbf{x}^\infty)$, $\mu \in (0,1)$ and $C_2, Q > 0$.

If Assumption 3.1.5 holds, then the objects U and η in Def. 1 exist and by the continuity of ϕ, it is possible to find an index k_0 satisfying the following relations,

$$\alpha C_2^{-1}\phi\left(\mathcal{L}_\infty(\mathbf{x}_{k_0}) - \mathcal{L}_\infty(\mathbf{x}^\infty)\right) + \|\mathbf{x}_{k_0} - \mathbf{x}^\infty\| + \frac{\mu^{k_0}Q}{1-\mu} < r,$$

$$\mathcal{L}_\infty(\mathbf{x}_k) \in [\mathcal{L}_\infty(\mathbf{x}^\infty), \mathcal{L}_\infty(\mathbf{x}^\infty) + \eta), \quad \forall k \geq k_0,$$

where $\mathcal{B}(\mathbf{x}^\infty, r) \subset U$.

The global convergence of NEAR-DGD$^+$ follows from applying Lemma 3.1.4 on the sequence $\{\mathbf{x}_k\}_{k\geq k_0}$ with $h = \mathcal{L}_\infty$, $c = \alpha C_2^{-1}$, $\{\nu_k\} = \{\mu^{k+k_0}Q\}$ and $\mathbf{x}^\star = \mathbf{x}^\infty$. Since $\mathbf{x}^\infty$ is the limit point of a subsequence of $\{\mathbf{x}_k\}$ and $\{\mathbf{x}_k\}$ is convergent, we conclude that $\{\mathbf{x}_k\} \to \mathbf{x}^\infty$. □

We have completed our analysis of the first-order convergence properties of NEAR-DGD^t and NEAR-DGD$^+$, the two instances of NEAR-DGD under examination. In the next subsection we provide second order guarantees for the same two variants.

3.1.3. Second order guarantees

In this subsection, we provide second order guarantees for the two variants of the NEAR-DGD method (1.0.5), (1.0.6) we studied in Subsection 3.1.2, namely NEAR-DGD^t ($t(k) = t$ in Eq. (1.0.5) for some $t \in \mathbb{N}^+$) and NEAR-DGD$^+$ ($t(k) = k$ in Eq. (1.0.5)). Specifically, using recent results stemming from dynamical systems theory, we will prove that those two variants almost surely avoid strict saddles when initialized randomly. We begin by listing a number of relevant assumptions, definitions and theoretical results.

Assumption 3.1.21. ***(Differentiability)*** *The function* $\mathbf{f}$ *is* $\mathcal{C}^2$.

We note that Assumption 3.1.21 implies that the Lyapunov functions (3.1.2) and (3.1.8) are also $\mathcal{C}^2$.

Definition 2. ***(Differential of a mapping)*** *[Ch. 3, [**1**]] The differential of a mapping $g : \mathcal{X} \to \mathcal{X}$, denoted as $Dg(x)$, is a linear operator from $\mathcal{T}(x) \to \mathcal{T}(g(x))$, where $\mathcal{T}(x)$ is the tangent space of $\mathcal{X}$ at point x. Given a curve γ in $\mathcal{X}$ with $\gamma(0) = x$ and $\frac{d\gamma}{dt}(0) = v \in \mathcal{T}(x)$, the linear operator is defined as $Dg(x)v = \frac{d(g\circ\gamma)}{dt}(0) \in \mathcal{T}(g(x))$. The determinant of the linear operator $\det(Dg(x))$ is the determinant of the matrix representing $Dg(x)$ with respect to an arbitrary basis.*

Definitions 3-4, Theorems 3.1.22-3.1.23 and Corollaries 3.1.25-3.1.25 are adapted from [**85**].

Definition 3. ***(Unstable fixed points)*** *The set of unstable fixed points $\mathcal{A}_g^\star$ of a mapping $g : \mathcal{X} \to \mathcal{X}$ is defined as $\mathcal{A}_g^\star = \{x \in \mathcal{X} : g(x) = x, \max_i |\lambda_i(Dg(x))| > 1\}$.*

Definition 4. ***(Strict saddles)*** *The set of strict saddles $\mathcal{X}^\star$ of a function $f : \mathcal{X} \to \mathbb{R}$ is defined as $\mathcal{X}^\star = \{x^\star \in \mathcal{X} : \nabla f(x^\star) = 0, \lambda_1(\nabla^2 f(x^\star)) < 0\}$.*

Theorem 3.1.22. ***(Stable Center Manifold Theorem)*** *[Theorem 1, [**85**]] Let $x^\star$ be a fixed point for the C^r local diffeomorphism $g : \mathcal{X} \to \mathcal{X}$. Suppose that $\mathcal{X} = \mathcal{X}_s \oplus \mathcal{X}_u$, where $\mathcal{X}_s$ is the span of the eigenvectors corresponding to eigenvalues of magnitude less than or equal to one of $Dg(x^\star)$, and $\mathcal{X}_u$ is the span of the eigenvectors corresponding to eigenvalues of magnitude greater than one of $Dg(x^\star)$. Then there exists a C^r embedded disk W_{loc}^{cs} that is tangent to $\mathcal{X}_s$ at $x^\star$ called the local stable center manifold. Moreover, there exists a neighborhood B of $x^\star$, such that $g(W_{loc}^{cs}) \cap B \subset W_{loc}^{cs}$, and $\cap_\infty^{k=0} g^{-k}(B) \subset W_{loc}^{cs}$.*

Theorem 3.1.23. ***(Non-convergence to unstable fixed points)*** *[Theorem 2, [**85**]] Let g be a $\mathcal{C}^1$ mapping from $\mathcal{X} \to \mathcal{X}$ and $\det(Dg(x)) \neq 0$ for all $x \in \mathcal{X}$. Then the set of initial points that converge to an unstable fixed point has measure zero, i.e., $\mu(\{x_0 : \lim x_k \in \mathcal{A}_g^\star\}) = 0$, where $\mathcal{A}_g^\star$ is the set of unstable fixed points of g.*

Corollary 3.1.24. ***(Non-convergence to saddle points)*** *[Corollary 1, [**85**]] Under the same conditions as Theorem 3.1.23, and in addition assume $\mathcal{X}^\star \subset \mathcal{A}_g^\star$ where $\mathcal{X}^\star$ the set of strict saddles of a function f, then $\mu(W_g) = 0$ where $W_g = \{x_0 : \lim_k g^k(x_0) \in \mathcal{X}^\star\}$ and $g^k(x_0)$ is the k-fold composition of the mapping g (k steps) starting from initial point x_0.*

Corollary 3.1.25. ***(Gradient descent converges to minimizers)*** *[Corollary 2, [**85**]] Let $g(x_k) = x_k - \alpha\nabla f(x_k)$ be the gradient descent algorithm, where $f \in \mathcal{C}^2$ and $\|\nabla^2 f(x)\|_2 \leq L$. Then if $\alpha < 1/L$, the stable set of the strict saddle points has measure zero, meaning $\mu(W_g) = 0$.*

We can express the k^{th} iterate of NEAR-DGD as a mapping applied either on the $\mathbf{x}_k$ or the $\mathbf{y}_k$ iterates produced by Equations (1.0.5) and (1.0.6) respectively as follows,

$$g_k^x(\mathbf{x}) = \mathbf{Z}^{t(k)}(\mathbf{x} - \alpha\nabla\mathbf{f}(\mathbf{x})) \tag{3.1.13a}$$

$$g_k^y(\mathbf{y}) = \mathbf{Z}^{t(k)}\mathbf{y} - \alpha\nabla\mathbf{f}(\mathbf{Z}^{t(k)}\mathbf{y}). \tag{3.1.13b}$$

with $Dg_k^y(\mathbf{y}) = \mathbf{Z}^{t(k)}\left(I - \alpha\nabla^2\mathbf{f}(\mathbf{Z}^{t(k)}\mathbf{y})\right)$ and $Dg_k^x(\mathbf{x}) = \mathbf{Z}^{t(k)}(I - \alpha\nabla^2\mathbf{f}(\mathbf{x}))$.

To to utilize Theorem 3.1.22 for establishing the non-convergence of NEAR-DGD to strict saddles, we first need to confirm that the mappings in Eq. (3.1.13) are diffeomorphisms. In the following Lemma we show that this is indeed the case for any fixed value of the sequence $\{t(k)\}$.

Lemma 3.1.26. ***(Diffeomorphism)*** *Let the steplength in Eq.* (3.1.13a) *and Eq.* (3.1.13b) *satisfy* $\alpha < 1/L$. *Then the mappings* $g_k^x, g_k^y : \mathbb{R}^{np} \to \mathbb{R}^{np}$ *are diffeomorphisms for any positive integer value value* $t(k)$.

Proof. It suffices to show that $\det(Dg_k^y(\mathbf{y})) \neq 0$ for all $\mathbf{y} \in \mathbb{R}^{np}$ and $\det(Dg_k^x(\mathbf{x})) \neq 0$ for all $\mathbf{x} \in \mathbb{R}^{np}$. For some vector $\mathbf{v} \in \mathbb{R}^{np}$, let $\lambda_i(\nabla^2\mathbf{f}(\mathbf{v}))$ be the eigenvalues of the Hessian $\nabla^2\mathbf{f}(\mathbf{v})$. Assumption 3.1.2 implies that $\lambda_i(\nabla^2\mathbf{f}(\mathbf{v})) \leq L$ for all $i \in \{1, ..., np\}$. Moreover, the determinants of both Dg_k^x and Dg_k^y can be decomposed in the form $\det\left(Dg_k(\cdot)\right) = \det(\mathbf{Z}^{t(k)}) \det(I - \alpha\nabla^2\mathbf{f}(\mathbf{v})) = \left(\prod_i \lambda_i^{t(k)}(\mathbf{Z})\right) (\prod_i (1 - \alpha\lambda_i(\nabla^2\mathbf{f}(\mathbf{v}))))$. Thus, $\det\left(Dg_k(\cdot)\right) > 0$ by the positive-definiteness of $\mathbf{Z}$ and $\alpha < 1/L$. □

We continue our analysis by proving that the NEAR-DGD^t variant almost surely avoids the strict saddles of the Lyapunov function (3.1.2).

Theorem 3.1.27. ***(Convergence to 2nd order stationary points (NEAR-DGD^t))*** *Let* $\{\mathbf{y}_k\}$ *be the sequence of iterates generated by NEAR-DGD^t under steplength* $\alpha < 1/L$. *Moreover, let us define the set of unstable fixed points* $\mathcal{A}_g^\star$ *of NEAR-DGD^t and the set of strict saddles* $\mathcal{Y}^\star$ *of the Lyapunov function* (3.1.2) *following Def. 3 and 4, respectively. Then if the Lyapunov function* $\mathcal{L}_t$ (3.1.2) *satisfies the strict saddle property (i.e. its critical points are either minima or strict saddles), the sequence* $\{\mathbf{y}_k\}$ *converges almost surely to a* 2^{nd} *order stationary point of* $\mathcal{L}_t$.

Proof. We first observe that NEAR-DGD^t is expressed as the mapping $g_t^y : \mathbb{R}^{np} \to \mathbb{R}^{np}$ for fixed $t \in \mathbb{N}^+$, which is a diffeomorphism by Lemma 3.1.26. Hence, to prove almost sure avoidance of strict saddles with Theorem 3.1.23 it suffices to show that $\mathcal{Y}^\star \subset \mathcal{A}_g^\star$.

Every critical point $\mathbf{y}^\star$ of (3.1.2) satisfies $\nabla \mathcal{L}_t(\mathbf{y}^\star) = \mathbf{0}$, i.e.

$$\mathbf{Z}^t \nabla \mathbf{f}(\mathbf{Z}^t \mathbf{y}^\star) + \frac{1}{\alpha} \mathbf{Z}^t \mathbf{y}^\star - \frac{1}{\alpha} \mathbf{Z}^{2t} \mathbf{y}^\star = 0.$$

Since $\mathbf{Z}$ is positive-definite and by thus non-singular, we can multiply both sides of the equality above with $\alpha \mathbf{Z}^{-t}$ and re-arrange the resulting terms to obtain $\mathbf{y}^\star = g_t^y(\mathbf{y}^\star)$, which confirms that $\mathbf{y}^\star$ is a fixed point of NEAR-DGD^t.

The Hessian of $\mathcal{L}_t$ (3.1.2) at $\mathbf{y}^\star$ is given by,

$$\begin{aligned} \nabla^2 \mathcal{L}_t(\mathbf{y}^\star) &= \mathbf{Z}^t \nabla^2 \mathbf{f}(\mathbf{Z}^t \mathbf{y}^\star) \mathbf{Z}^t + \frac{1}{\alpha} \mathbf{Z}^t (I - \mathbf{Z}^t) \\ &= \frac{1}{\alpha} \left(I - D g_t^y(\mathbf{y}^\star) \right) \mathbf{Z}^t. \end{aligned} \tag{3.1.14}$$

We define the matrix $\mathbf{P} := \alpha \mathbf{Z}^{-\frac{t}{2}} \nabla^2 \mathcal{L}_t(\mathbf{y}^\star) \mathbf{Z}^{-\frac{t}{2}}$. Using the positive-definiteness of $\mathbf{Z}$, we obtain from (3.1.14)

$$I - D g_t^y(\mathbf{y}^\star) = \alpha \nabla^2 \mathcal{L}_t(\mathbf{y}^\star) \mathbf{Z}^{-t} = \mathbf{Z}^{\frac{t}{2}} \mathbf{P} \mathbf{Z}^{-\frac{t}{2}},$$

which implies that $(I - D g_t^y(\mathbf{y}^\star))$ and $\mathbf{P}$ are similar matrices and have identical spectrums. Moreover, the matrix $\mathbf{Z}^{-\frac{t}{2}}$ is symmetric by Assumption 3.1.1. Hence, $\mathbf{P}$ and $(\alpha \nabla^2 \mathcal{L}_t(\mathbf{y}^\star))$ are congruent and by Sylvester's law of inertia [Theorem 4.5.8, [**69**]] they have the same number of negative eigenvalues. Given that $\nabla^2 \mathcal{L}_t(\mathbf{y}^\star)$ has at least one negative eigenvalue

by Def. 4, we conclude that so does $\mathbf{P}$ and there exists index i such that $1-\lambda_i(Dg_t^y(\mathbf{y}^\star)) < 0$ or $\lambda_i(Dg_t^y(\mathbf{y}^\star)) > 1$. Applying Corollary 3.1.24 completes the proof. □

We conclude our convergence analysis by demonstrating that the NEAR-DGD$^+$ variant locally avoids the strict saddles of the function $f : \mathbb{R}^p \to \mathbb{R}$ of Problem 1.0.1 almost surely providing that the steplength α in Eq. (1.0.6) is chosen within an appropriate range.

Theorem 3.1.28. ***(Convergence to SOS (NEAR-DGD$^+$))*** *Suppose that the function $f(x) = \sum_{i=1}^n f_i(x)$ of Problem 1.0.1 has a non-empty set of minimizers and satisfies the strict saddle property. Moreover, suppose that f is $\mathcal{C}^2$ and has L_f-Lipschitz continuous gradients. Then under steplength satisfying*

$$\alpha < \min\left\{\frac{1+\sqrt{5}}{2L}, \frac{\beta^{-1}-1}{L}, \frac{n}{L_f}\right\}$$

in Eq. (1.0.6)*, the sequences of the local $\{x_{i,k}\}$ iterates of the NEAR-DGD$^+$ method generated by Eq.* (1.0.5) *with $t(k) = k$ almost surely converge to a minimizer of f.*

Proof. First, we observe that NEAR-DGD$^+$ can be viewed as a successive application of the mappings

$$g_k^x(\mathbf{x}) = \mathbf{Z}^k(\mathbf{x} - \alpha\nabla\mathbf{f}(\mathbf{x})), \quad k = 1, 2, \ldots.$$

Let $v_1, ..., v_n$ be the eigenvectors of $\mathbf{W}$ with corresponding eigenvalues $0 < \lambda_1 \leq ... < \lambda_n$, where $v_n = 1_n$ and $\lambda_n = 1$. Any vector $u \in \mathbb{R}^n$ can be written as $u = \sum_{i=1}^n a_i v_i$ for some real coefficients $a_i \in \mathbb{R}$ with $a_n = \frac{1}{n}1_n^T u$. Multiplying u with $\mathbf{W}^k$ and taking the limit

$k \to \infty$ yields,

$$\lim_{k\to\infty} \mathbf{W}^k u = \lim_{k\to\infty} \left(\sum_{i=1}^{n} a_i \mathbf{W}^k v_i \right)$$
$$= \lim_{k\to\infty} \left(\sum_{i=1}^{n} a_i \lambda_i^k v_i \right) = \left(\frac{1}{n} 1_n^T u \right) 1_n,$$

due to $\lambda_i^k \to 0$ for $i \neq n$.

Hence, successive applications of $g_k^x(\mathbf{x})$ as $k \to \infty$ converge to the mapping g_∞^x given by

$$g_\infty^x(\mathbf{x}) = \mathbf{M}(\mathbf{x} - \alpha \nabla \mathbf{f}(\mathbf{x})).$$

The mapping g_∞^x is not a diffeomorphism due to the positive semi-definiteness of $\mathbf{M}$. However, by Lemma 3.1.15 we have $\|(I - \mathbf{M})\mathbf{x}_k\| \to 0$, and applying g_∞^x to $\bar{\mathbf{x}}^\infty = \mathbf{M}\mathbf{x}^\infty$ yields

$$g_\infty^x(\bar{\mathbf{x}}^\infty) = \mathbf{M}\bar{\mathbf{x}}^\infty - \alpha \mathbf{M} \nabla \mathbf{f}(\bar{\mathbf{x}}^\infty) = \bar{\mathbf{x}}^\infty - \alpha \mathbf{M} \nabla \mathbf{f}(\bar{\mathbf{x}}^\infty).$$

Let $\bar{x}^\infty = n^{-1} \sum_{i=1}^n x_i^\infty \in \mathbb{R}^p$ so that $\mathbf{M}\bar{\mathbf{x}}^\infty = \bar{x}^\infty \otimes 1_n$. The preceding relation implies that the mapping induced by g_∞^x on $\bar{x}^\infty$ is equivalent to the standard centralized gradient descent method with steplength αn^{-1}, i.e.

$$g_\infty^x \left(\begin{bmatrix} \bar{x}^\infty \\ \bar{x}^\infty \\ \vdots \\ \bar{x}^\infty \end{bmatrix} \right) = \begin{bmatrix} \bar{x}^\infty \\ \bar{x}^\infty \\ \vdots \\ \bar{x}^\infty \end{bmatrix} - \frac{\alpha}{n} \begin{bmatrix} \sum_{i=1}^n \nabla f_i(\bar{x}^\infty) \\ \sum_{i=1}^n \nabla f_i(\bar{x}^\infty) \\ \vdots \\ \sum_{i=1}^n \nabla f_i(\bar{x}^\infty) \end{bmatrix}$$
$$= (\bar{x}^\infty - \frac{\alpha}{n} \nabla f(\bar{x}^\infty)) \otimes 1_n.$$

Applying Corollary 3.1.25 completes the proof. □

3.2. Numerical Results

3.2.1. Quadratic problem

We evaluate the empirical performance of NEAR-DGD on the following regularized quadratic problem,

$$\min_{x\in\mathbb{R}^p} f(x) = \frac{1}{2}\sum_{i=1}^{n}\left(\|x\|_{Q^i}^2\right) + \frac{1}{4}\|x\|_{D_I}^4,$$

where $I \in \{1, ..., p\}$ is some positive index and $Q^i \in \mathbb{R}^{p\times p}$ and $D_I \in \mathbb{R}^{p\times p}$ are diagonal matrices constructed as follows: $q_{jj}^i < 0$ if $j = I$ and $q_{jj}^i > 0$ otherwise, and $D_I = c \cdot e_I e_I'$, where $c > 0$ is a constant and e_I is the indicator vector for the I^{th} element. It is easy to check that f has a unique saddle point at $x = \mathbf{0}$ and two minima at $x^\star = \pm\frac{1}{c}\left(\sqrt{\sum_{i=1}^n -q_{II}^i}\right)e_I$. We can distribute this problem to n nodes by setting $f_i(x) = \frac{1}{2}\|x\|_{Q^i}^2 + \frac{1}{4n}\|x\|_{D_I}^4$. Moreover, each f_i has Lipschitz gradients in any compact subset of $\mathbb{R}^p$.

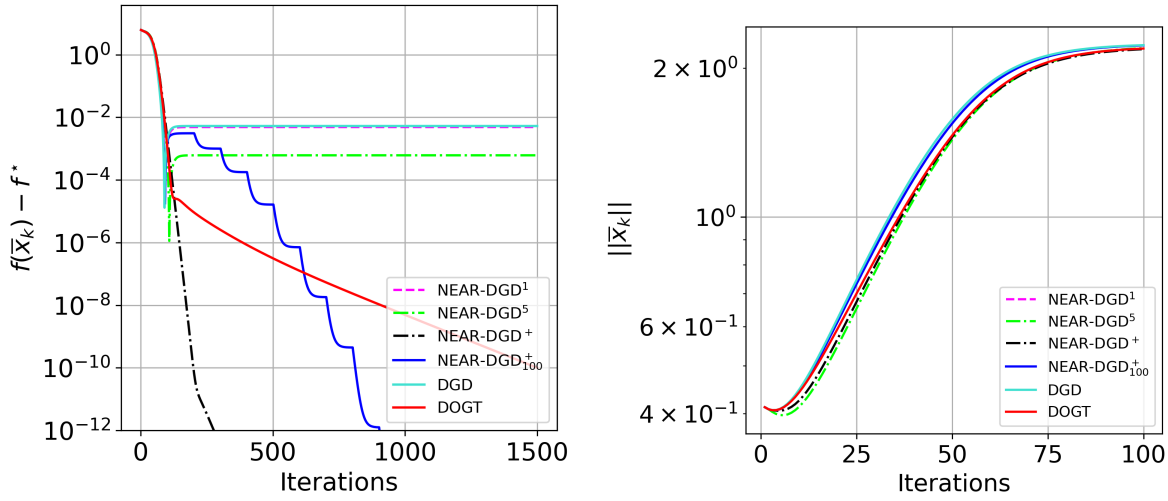
We set $p = I = 4$ for the purposes of our numerical experiment. The matrices Q^i were constructed randomly with $q_{jj}^i \in (-1, 0)$ for $j = I$ and $q_{jj}^i \in (0, 1)$ otherwise, and the parameter c of matrix D_I was set to 1. We allocated each f_i to a unique node in a network of size $n = 12$ with ring graph topology. We tested 6 methods in total, including DGD [**111, 186**], DOGT (with doubly stochastic consensus matrix) [**40**], and 4 variants of the NEAR-DGD method: *i*) NEAR-DGD^1 (one consensus round per gradient evaluation), *ii*) NEAR-DGD^5 (5 consensus rounds per gradient evaluation), *iii*) a variant of NEAR-DGD where the sequence of consensus rounds increases by 1 at every iteration, and to which we will refer as NEAR-DGD$^+$, and *iv*) a practical variant of NEAR-DGD$^+$,

where starting from one consensus round/iteration, we double the number of consensus rounds every 100 gradient evaluations. We will refer to this last modification as NEAR-DGD$^{+}_{100}$. All algorithms were initialized from the same randomly chosen point in the interval $[-1, 1]^{np}$. We manually tuned the steplength to $\alpha = 10^{-1}$ to achieve the fastest possible convergence rates, and used the same value for all methods for fairness.

In Fig. 3.1, we plot the objective function error $f(\bar{x}_k) - f^\star$ where $f^\star = f(x^\star)$ (Fig. 3.1a) and the distance $\|\bar{x}_k\|$ of the average iterates to the saddle point $x = \mathbf{0}$ (Fig. 3.1b) versus the number of iterations/gradient evaluations for all methods. In Fig. 3.1a, we observe that convergence accuracy increases with the value of the parameter t of NEAR-DGD^t, as predicted by our theoretical results. NEAR-DGD^1 performs comparably to DGD, while the two variants of NEAR-DGD paired with increasing sequences of consensus rounds per iteration, i.e. NEAR-DGD$^+$ and NEAR-DGD$^+_{100}$, achieve exact convergence to the optimal value with faster rates compared to DOGT. All methods successfully escape the saddle point of f with approximately the same speed (Fig. 3.1b). We noticed that the trends in Fig. 3.1b were very sensitive to small changes in problem parameters and the selection of initial point.

In Fig. 3.2, we plot the objective function error $f(\bar{x}_k) - f^\star$ versus the cumulative application cost (per node) for all methods, where we calculated the cost per iteration using the framework proposed in [**13**],

$$\text{Cost} = c_c \times \#\text{Communications} + c_g \times \#\text{Computations},$$



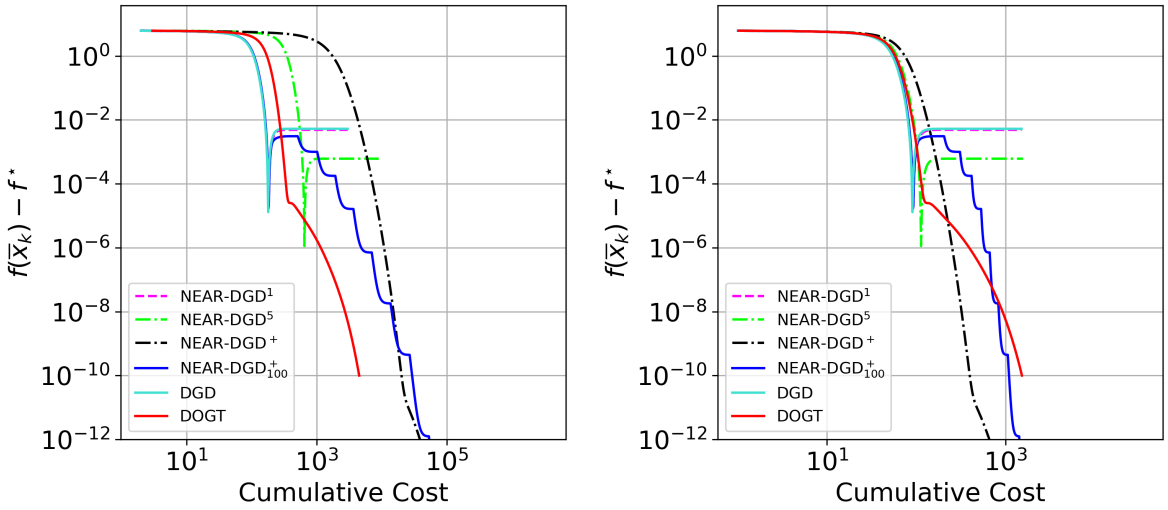
(a) Objective function error (b) Distance to $x = \mathbf{0}$ (saddle)

Figure 3.1. Distance to $f^\star$ (left) and to saddle point (right)

where c_c and c_g are constants representing the application-specific costs of one communication and one computation operation, respectively. In Fig. 3.2a, the costs of communication and computation are equal ($c_c = c_g$) and DOGT outperforms NEAR-DGD$^+$ and NEAR-DGD$^+_{100}$ since it requires only two communication rounds per gradient evaluation to achieve exact convergence. Conversely, in Fig. 3.2b, the cost of communication is relatively low compared to the cost of computation ($c_c = 10^{-2} c_g$). In this case, NEAR-DGD$^+$ converges to the optimal value faster than the remaining methods in terms of total application cost.

3.2.2. Neural Networks

We conclude the Numerical Results section of this chapter by assessing the performance of NEAR-DGD on the classification of the MNIST dataset [**42**] with a feed-forward Neural



(a) $c_g = 1$, $c_c = 1$ (b) $c_g = 1$, $c_c = 10^{-2}$

Figure 3.2. Objective function error as a function of cumulative application cost (per node)

Network (NN) trained in a decentralized manner. We report results for the following methods: *i*) NEAR-DGD^1, *ii*) NEAR-DGD^2, *iii*) DGD [**111, 186**] and *iv*) DOGT [**40**] with a doubly-stochastic matrix. We note that methods (i) and (iii) have the same iteration cost as they perform the same amount of computation and communication at every iteration, and the same is true for methods (ii) and (iv). For the purposes of our experiment, we implemented a 2-hidden layer NN in Python with layer dimensions $(784, 128, 64, 10)$ and with cross-entropy as the loss function. We used the sigmoid activation function for the first two layers and the softmax activation function for the output layer.

The training set for MNIST contains $6 \cdot 10^4$ samples in total, which we shuffled and evenly distributed among n agents connected in a network with ring graph topology. All nodes computed full gradients using all $6 \cdot 10^4/n$ samples at their disposal at every iteration. The weights and biases of each layer were randomly initialized at the same point

for all methods $(i) - (iv)$, and grid search was used to find the maximum learning rate α that allows each method to successfully converge. At every iteration of each algorithm we computed the average model, i.e. the model whose parameters are the averages of the local parameters at each node, and tracked the following performance-related metrics:

(1) the value of the loss function for the average model calculated in a forward pass using all $6 \cdot 10^4$ training samples as input;

(2) the testing accuracy of the average model using all 10^4 samples in the testing set as input; and

(3) the consensus violation, i.e. the sum of the distances (l_2 norms) of the local models to the average model over all nodes and model parameters.

Finally, we repeated the experiment for two different network sizes, $n = 10$ (Fig. 3.3) and $n = 30$ (Fig. 3.4).

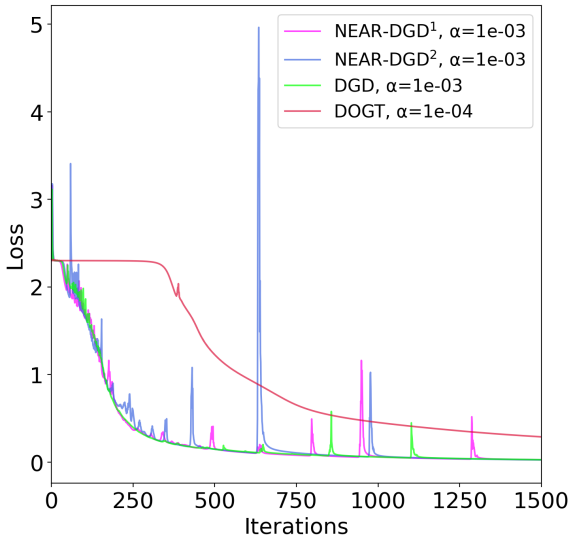
The values of the metrics (1), (2) and (3) for network size $n = 10$ and methods (i)-(iv) are plotted in Figures 3.3a, 3.3b and 3.3c, respectively. The two variants of NEAR-DGD^t and DGD were able to reach convergence with learning rate $\alpha = 10^{-3}$, while DOGT required a smaller learning rate ($\alpha = 10^{-4}$) and as a result is the slowest method to converge, although it achieves the smallest consensus error out of all methods (Fig. 3.3c). DGD, NEAR-DGD^1 and NEAR-DGD^2 perform equally well with respect to loss function value and testing accuracy (Figures 3.3a and 3.3b), however NEAR-DGD^2 outperforms DGD and NEAR-DGD^1 in the consensus violation metric (3.3c), demonstrating the advantage of performing additional consensus rounds and confirming our theoretical analysis.

Our results for metrics (1), (2) and (3), network size $n = 30$ and methods (i)-(iv) are plotted in Figures 3.4a, 3.4b and 3.4c. All methods converged with learning rate $\alpha = 10^{-3}$

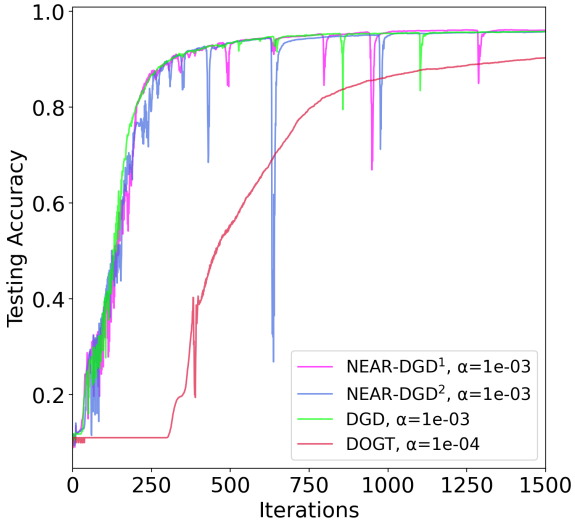
in this instance, with NEAR-DGD^2 converging the fastest out of all methods and DOGT the slowest (Figures 3.4a and 3.4b). In terms of agreement between nodes, NEAR-DGD^2 and DOGT achieve the lowest consensus error (Fig. 3.4c), with NEAR-DGD^2 yielding a smoother curve. DGD and NEAR-DGD^1 performed comparably in all cases.

3.3. Summary

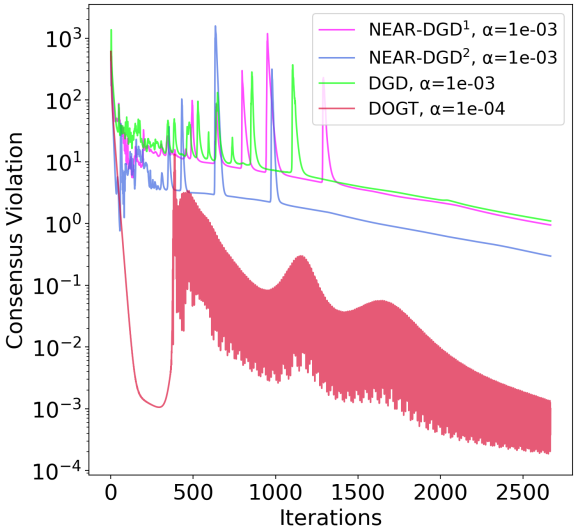
NEAR-DGD [**13**] is a distributed first order method that permits adjusting the amounts of computation and communication carried out at each iteration to balance convergence accuracy and total application cost. We have extended to the nonconvex setting the analysis of two variants of NEAR-DGD: *i*) NEAR-DGD^t, which performs a fixed number of communication rounds at every iteration controlled by the parameter t, and *ii*) NEAR-DGD$^+$, a time-varying instance of NEAR-DGD that increases the number of consensus rounds executed by 1 at every iteration. Given a connected, undirected network with general topology, and the relatively mild assumptions of function coercivity, Lipschitz gradient continuity and satisfaction of the Kurdyka-Łojasiewicz (KL) property in the entire domain, we have shown that NEAR-DGD^t converges to the set of critical points of a custom Lyapunov function which approaches the set of first order stationary points of the original problem as t increases, and that NEAR-DGD$^+$ converges to first order stationary points of the original problem while its iterates achieve consensus exponentially fast. Moreover, using recent results from dynamical systems theory, we were able to establish almost sure avoidance of strict saddles for both variants. Our numerical results confirm our theoretical analysis and demonstrate that NEAR-DGD can perform favorably against state-of-the-art methods for nonconvex problems.



(a) Loss function value (average model)

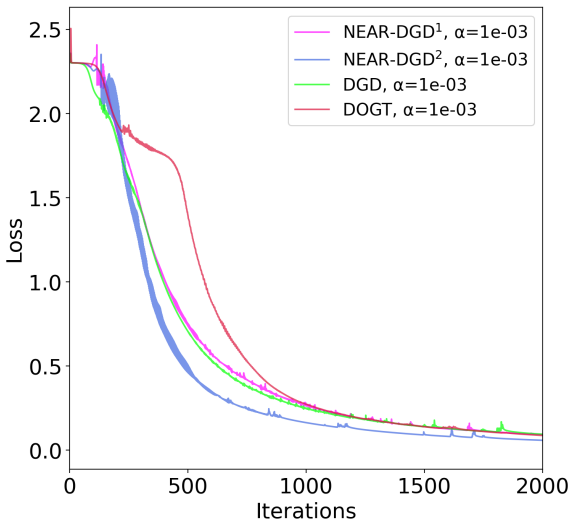


(b) Testing accuracy (average model)

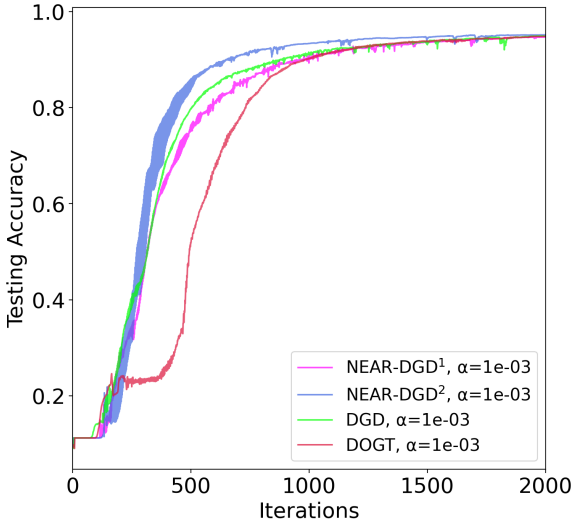


(c) Consensus violation

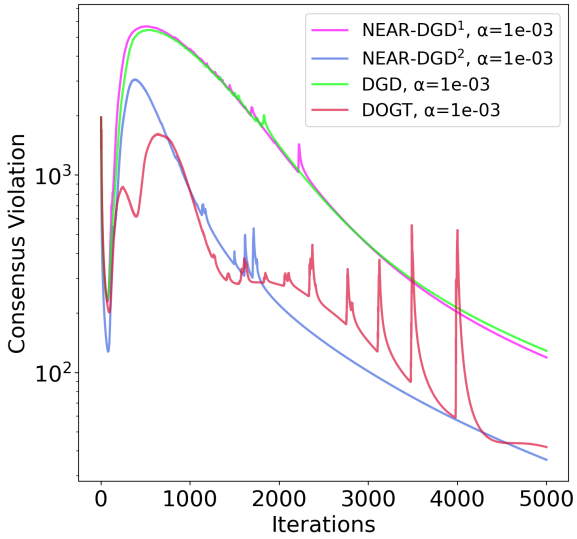
Figure 3.3. Performance of distributed optimization methods for a 2-hidden layer NN classifying the MNIST dataset, network size $N = 10$



(a) Loss function value (average model)



(b) Testing accuracy (average model)



(c) Consensus violation

Figure 3.4. Performance of distributed optimization methods for a 2-hidden layer NN classifying the MNIST dataset, network size $N = 30$

CHAPTER 4

Asynchronous Distributed Rendezvous With Probabilistic Guarantees

4.1. Algorithm Development

For the rest of this Chapter, we focus on the case of fixed, undirected network topologies, implicitly assuming connectivity maintenance. We adopt point models for the agents, and do not consider interagent and agent-obstacle collisions. Moreover, we assume that all agents satisfy the following conditions:

C1: Agents have access to a global, fixed coordinate system;

C2: Each agent can store a number of packets equal to the size of its neighborhood;

C3: Agents can instantaneously evaluate gradients, read stored packets and broadcast messages over the network;

C4: Agents are not equipped with sensors capable of detecting the positions of other agents; conversely, this information can only be obtained via the communication channel.

Each agent $i \in \mathcal{V}$ is initially located at position $x_{i,0} \in \mathbb{R}^p$ with its direction $g_{i,0}$ set at $g_{i,0} = 0$ to enforce stillness. Moreover, each agent arbitrarily initializes the stored positions $x^b_{j,0}$ of its neighbors for all pairs $(i, j) \in \mathcal{E}$ (eg. $x^b_{j,0}$ can be set equal to zero or the actual position of agent j if available or any other value). All agents know in advance the following global parameters: a velocity parameter $c > 0$ which serves as a scaling factor for all agents' velocities and a parameter $\alpha > 0$ that controls the accuracy

of the rendezvous. In line with existing works utilizing Poisson clocks for the analysis of asynchronous distributed algorithms [**133, 97**], we define each local agent activation, i.e. each arrival of a local Poisson clock, to be an iteration count $k = 1, 2, ...$ of the algorithm. We assume that only a single agent can be active at each iteration count.

If agent i is active at the k^{th} iteration count then it performs the following actions: i) it immediately senses its current position $x_{i,k}$ and broadcasts it to its neighbors, effectively setting $x^b_{i,k} = x_{i,k}$; and ii) it reads from its buffer the outdated values $x^b_{j,k-1}$ for $(i,j) \in \mathcal{E}$ (if there are multiple values for $x^b_{j,k-1}$ in its buffer, we assume agent i can deduce which is the most recent one) and sets its velocity (controller input) equal to $c \cdot g_{i,k}$ where $g_{i,k}$ is computed using the equation below

$$g_{i,k} = x^w_{i,k} - \alpha \nabla f_i(x^w_{i,k}) - x_{i,k}, \tag{4.1.1}$$

where $x^w_{i,k} = w_{ii}x_{i,k} + \sum_{j \neq i} w_{ij}x^b_{j,k-1}$ and $w_{i,j}$ is the element in the i^{th} row and j^{th} column of a matrix $W \in \mathbb{R}^{n \times n}$ satisfying the following assumption.

Assumption 4.1.1. ***(Consensus matrix)*** *The matrix $W \in \mathbb{R}^{n \times n}$ has the following properties: i) symmetry; ii) double stochasticity; iii) $w_{ij} > 0$ if and only if $(i,j) \in \mathcal{E}$ or $i = j$ and $w_{ij} = 0$ otherwise; and iv) positive-definiteness.*

Conversely, if agent i is inactive during the k^{th} iteration count, it continues to move in the most recently calculated direction (or remains still if it has never been activated) and inactively listens for incoming messages from neighbors. In the absence of collisions and other disturbances, it follows that regardless of the state of agent i (active or inactive),

its position at k^{th} iteration will be

$$x_{i,k} = x_{i,k-1} + c \cdot \delta_t(k) \cdot g_{i,k-1},$$

where $\delta_t(k)$ is the elapsed (continuous) time between the k^{th} and the $(k-1)^{th}$ iteration, $x_{i,k-1}$ is the position of agent i at the $(k-1)^{th}$ iteration, and $g_{i,k-1}$ is the direction of agent i at the $(k-1)^{th}$ iteration. The entire procedure is summarized in Algorithm 2.

Algorithm 2: Asynchronous distributed rendezvous for continuously moving agents at node i

Initialization: position $x_{i,0} \in \mathbb{R}$, neighbor positions $x^b_{j,0}$ for $(i,j) \in \mathcal{E}$, direction $g_{i,0} = 0$, velocity parameter $c > 0$, proximity parameter $\alpha > 0$;
for $k = 1, 2, ...$ **do**
  if *i is active at k* **then**
    sense current position $x_{i,k} = x_{i,k-1} + c \cdot \delta_t(k) \cdot g_{i,k-1}$;
    broadcast $x_{i,k}$ to neighbors, i.e. set $x^b_{i,k} = x_{i,k}$;
    read received neighbor positions $x^b_{j,k-1}$ for all $(i,j) \in \mathcal{E}$;
    calculate new direction $g_{i,k}$ using Eq. (4.1.1);
    set velocity equal to $c \cdot g_{i,k}$;
  else
    continue moving in the direction $g_{i,k-1}$;
    inactively listen for messages from neighbors;
  end
end

For simplicity, we assume that the space dimension p is equal to 1 for the remainder of this Chapter. We note, however, that the analysis can be directly extended to any value of p. We define the function $F : \mathbb{R}^n \to \mathbb{R}$

$$F(X) = \sum_{i=1}^{n} f_i(x_i),$$

where $X = [x_1, ..., x_n]' \in \mathbb{R}^n$ is the column-wise concatenation of the variables x_i for $i \in \mathcal{V}$. We adopt the following standard assumptions on the function F.

Assumption 4.1.2. ***(Lipschitz gradients)*** *The function F has L-Lipschitz continuous gradients.*

Assumption 4.1.3. ***(Strong convexity)*** *The function F is μ-strongly convex.*

Moreover, let $\Phi_k = diag(\phi_{11,k}, ..., \phi_{nn,k}) \in \mathbb{R}^{n\times n}$ be a diagonal selection matrix, such that $\phi_{ii,k} = 1$ if agent i is active at iteration k and $\phi_{jj,k} = 0$ for $j \neq i$. We then can compactly express the iterates of Algorithm 2 as

$$X_k = X_{k-1} + c\delta_t(k)(X^w_{k-1} - \alpha\nabla F(X^w_{k-1}) - X^b_{k-1}) \quad (position) \tag{4.1.2}$$

$$X^b_k = \Phi_k X_k + (I - \Phi_k)X^b_{k-1} \quad (buffer) \tag{4.1.3}$$

$$X^w_k = \Phi_k(W_d X_k + (W - W_d)X^b_{k-1}) + (I - \Phi_k)X^w_{k-1} \quad (consensus), \tag{4.1.4}$$

where $W_d = diag(w_{11}, ..., w_{nn}) \in \mathbb{R}^{n\times n}$ is the diagonal of the consensus matrix W and $X_k = [x_{1,k}, ..., x_{n,k}]' \in \mathbb{R}^n$, $X^b_k = [x^b_{1,k}, ..., x^b_{n,k}]' \in \mathbb{R}^n$ and $X^w_k = [x^w_{1,k}, ..., x^w_{n,k}]' \in \mathbb{R}^n$ are the column-wise concatenations of the position variables $x_{i,k}$, the stored positions $x^b_{i,k}$ known to neighbors and the "consensual" positions $x^w_{i,k}$ at iteration k, respectively. We note that the variables $x^w_{i,k}$ are calculated only during active states to update the directions $g_{i,k}$ using Eq. (4.1.1).

In the next section, we derive the convergence properties of Algorithm 2. Namely, we demonstrate that agents converge to an arbitrarily small neighborhood of the optimal solution $x^\star$ of Problem 1.0.1 while achieving approximate rendezvous.

4.2. Convergence Analysis

We begin this section by defining the following function that will play a key role in our analysis,

$$F_\alpha(X) = F(WX) + \frac{1}{2\alpha}\|X\|^2_{W-W^2}, \tag{4.2.1}$$

where $\|\cdot\|$ denotes the l_2-norm, i.e. $\|X\|^2 = \sum_{i=1}^n x_i^2$ for $X = [x_1, ..., x_n]' \in \mathbb{R}^n$.

The function F_α is the sum of a strongly convex and a convex function and therefore strongly convex. We will demonstrate that the iterates of Algorithm 2 converge to the unique minimizer $X^\star$ of F_α.

We also define the distance metric $\Delta_\alpha(X, Y) = X - Y - \alpha(\nabla F(X) - \nabla F(Y))$ for any two vectors $X, Y \in \mathbb{R}^n$ and the following quantities we will invoke throughout our analysis

$$\begin{aligned} \mathcal{E}_k^\star &= X_k - X^\star \quad (\textit{distance to optimality}) \\ \mathcal{E}_k^w &= WX_k^b - X_k^w \quad (\textit{consensus error}) \\ \mathcal{E}_k^b &= X_k - X_k^b \quad (\textit{buffer error}). \end{aligned} \tag{4.2.2}$$

Using the notation above, we can re-write Equations (4.1.2)-(4.1.4) as

$$X_k = X_{k-1} + c\delta_t(k)Q_{k-1} \tag{4.2.3}$$

$$X_k^b = X_{k-1}^b + \Phi_k\mathcal{E}_{k-1}^b + c\delta_t(k)\Phi_k Q_{k-1} \tag{4.2.4}$$

$$X_k^w = X_{k-1}^w + c\delta_t(k)\Phi_k W_d Q_{k-1} + \Phi_k W_d\mathcal{E}_{k-1}^b + \Phi_k\mathcal{E}_{k-1}^w, \tag{4.2.5}$$

where $Q_k = -\alpha W^{-1}\nabla F_\alpha(X_k) + \mathcal{E}_k^b + \Delta_\alpha(WX_k^b, WX_k) + \Delta_\alpha(X_k^w, WX_k^b)$. In the next subsection, we state a number of preliminary results necessary for our main analysis.

4.2.1. Preliminaries

The function F_α (4.2.1) is the sum of functions with Lipschitz continuous gradients, and therefore also has Lipschitz continuous gradients. We explicitly calculate the Lipschitz constant of F_α in the following Lemma.

Lemma 4.2.1. ***(Lipschitz gradients)*** *The function $F_\alpha : \mathbb{R}^n \to \mathbb{R}$ has L_α-Lipschitz continuous gradients, where $L_\alpha = L + \alpha^{-1}(1-\beta)$ and β is the smallest eigenvalue of the consensus matrix W.*

Proof. For any pair $X, Y \in \mathbb{R}^n$ we have

$$\begin{aligned}\|\nabla F_\alpha(X) - \nabla F_\alpha(Y)\| &= \|W(\nabla F(WX) - \nabla F(WY) + \alpha^{-1}(I-W)(X-Y))\| \\ &\leq L\|X-Y\| + \alpha^{-1}\|(I-W)(X-Y)\|,\end{aligned}$$

where we used the non-expansiveness of W twice and applied the triangle inequality to get the second inequality.

Observing that the maximum eigenvalue of $I - W$ is $1 - \beta$ concludes the proof. □

Due to the randomness of the Poisson clocks associated with each agent, in order to derive convergence guarantees for Algorithm 2 it is necessary to examine the expected values of the errors in Eq. (4.2.2). We calculate the expectations for various quantities of interest that emerge in the analysis of Algorithm 2 in the next Lemma.

Lemma 4.2.2. ***(Expectations)*** *Recall that λ_i is the Poisson clock parameter of agent i and consider the (continuous) time $\delta_t(k)$ elapsed between the k^{th} and $(k-1)^{th}$ global clock ticks and the selection matrix $\Phi_k = diag(\phi_{11,k}, ..., \phi_{nn,k})$ at the k^{th} global clock tick. The following relations hold for all $k \geq 1$ and deterministic vectors $X \in \mathbb{R}^n$,*

$$\mathbb{E}[c\delta_t(k)|\mathcal{F}_{k-1}] = \tilde{c}, \quad \mathbb{E}[c^2\delta_t^2(k)|\mathcal{F}_{k-1}] = 2\tilde{c}^2, \tag{4.2.6}$$

$$\mathbb{E}[\|\Phi_k X\|^2|\mathcal{F}_{k-1}] \leq \Pi\|X\|^2, \quad \mathbb{E}[\|(I-\Phi_k)X\|^2|\mathcal{F}_{k-1}] \leq (1-\pi)\|X\|^2, \tag{4.2.7}$$

(4.2.8)

$$\mathbb{E}[c^2\delta_t(k)^2\|\Phi_k X\|^2|\mathcal{F}_{k-1}] \leq 2\tilde{c}^2\Pi\|X\|^2, \mathbb{E}[c^2\delta_t(k)^2\|(I-\Phi_k)X\|^2|\mathcal{F}_{k-1}] \leq 2\tilde{c}^2(1-\pi)\|X\|^2,$$

where $\mathcal{F}_{k-1}$ is the sigma-algebra containing the history of the method up to and including the $(k-1)^{th}$ iteration, $\tilde{c} = c\left(\sum_{i=1}^n \lambda_i\right)^{-1}$, $\Pi = \max_j \lambda_j \left(\sum_{i=1}^n \lambda_i\right)^{-1}$ is the maximum activation probability among agents and $\pi = \min_j \lambda_j \left(\sum_{i=1}^n \lambda_i\right)^{-1}$ is the minimum activation probability among agents.

Proof. Eq. (4.2.6) follows directly from $\delta_t(k)$ being an exponential random variable with parameter $\sum_i \lambda_i$. Let $p_i = \lambda_i \left(\sum_i \lambda_i\right)^{-1}$ be the activation probability of agent i. To prove Eq. (4.2.7), we observe that $\Phi_k^2 = \Phi_k$ and hence for any deterministic $X = [x_1, ..., x_n]' \in \mathbb{R}^n$ we have

$$\mathbb{E}[\|\Phi_k X\|^2|\mathcal{F}_{k-1}] = \sum_{i=1}^n \mathbb{E}[\phi_{ii,k}|\mathcal{F}_{k-1}]x_i^2 = \sum_{i=1}^n p_i x_i^2 \leq \Pi\|X\|^2,$$

and

$$\mathbb{E}[\|(I-\Phi_k)X\|^2|\mathcal{F}_{k-1}] = \mathbb{E}[\|X\|^2 - \|\Phi_k X\|^2|\mathcal{F}_{k-1}] = \sum_{i=1}^{n}(1-p_i)x_i^2 \leq (1-\pi)\|X\|^2,$$

proving Eq. (4.2.7).

We now prove Eq. (4.2.8). Let t_i be an exponential random variable denoting the time between interarrivals for the i^{th} agent, i.e. t_i is an exponential random variable with parameter λ_i. If node i is inactive at the k^{th} global clock tick, the quantity $\phi_{ii,k}\delta_t(k)$ satisfies

$$\mathcal{P}(\phi_{ii,k}\delta_t(k) = 0|\mathcal{F}_{k-1}) = \mathcal{P}(\phi_{ii,k} = 0|\mathcal{F}_{k-1}) = 1-p_i.$$

If agent i is active at the k^{th} global clock tick, then for any positive scalar τ we have

$$\mathcal{P}(0 < \phi_{ii,k}\delta_t(k) \leq \tau|\mathcal{F}_{k-1}) = \mathcal{P}(t_i = \min\{t_1,...,t_n\}, t_i \leq \tau|\mathcal{F}_{k-1}).$$

Let ψ_i be the pdf of t_i and let $\bar{\lambda} = \sum_i \lambda_i$. The preceding relationship yields

$$\begin{aligned}
\mathcal{P}(t_i = \min\{t_1,...,t_n\}, t_i \leq \tau|\mathcal{F}_{k-1}) &= \int_0^\tau \psi_i(\tau-s)\prod_{j\neq i}\mathcal{P}(t_j > \tau - s)ds \\
&= \int_0^\tau \lambda_i e^{-\lambda_i(\tau-s)}\prod_{j\neq i} e^{-\lambda_j(\tau-s)}ds \\
&= \lambda_i e^{-\bar{\lambda}\tau}\int_0^\tau e^{\bar{\lambda}s}ds = p_i\left(1-e^{-\bar{\lambda}\tau}\right).
\end{aligned}$$

Hence the pdf of $\phi_{ii,k}\delta_t(k)$ conditioned on $\mathcal{F}_{k-1}$ and $\tau > 0$ is

$$\frac{d\left(p_i(1-e^{-\bar{\lambda}\tau})\right)}{d\tau} = \lambda_i e^{-\bar{\lambda}\tau},$$

which yields the expectation

$$\begin{aligned}\mathbb{E}[\phi_{ii,k}\delta_t(k)^2|\mathcal{F}_{k-1}] &= \lambda_i \int_0^\infty \tau^2 e^{-\bar{\lambda}\tau} d\tau \\ &= -\lambda_i \left(\frac{\tau^2}{\bar{\lambda}} e^{-\bar{\lambda}\tau} + \frac{2}{\bar{\lambda}^3}(\bar{\lambda}\tau e^{-\bar{\lambda}\tau} + e^{-\bar{\lambda}\tau})\right)\Bigg|_0^\infty = \frac{2\lambda_i}{\bar{\lambda}^3} = 2p_i\bar{\lambda}^{-2}.\end{aligned}$$

Hence, for any deterministic vector $X \in \mathbb{R}^n$ we have

$$\mathbb{E}[c^2\delta_t(k)^2\|\Phi_k X\|^2|\mathcal{F}_{k-1}] = c^2\sum_{i=1}^p \mathbb{E}[\delta_t(k)^2\phi_{ii,k}|\mathcal{F}_{k-1}]x_i^2 = 2\tilde{c}^2\sum_{i=1}^p p_i x_i^2 \leq 2\tilde{c}^2\Pi\|X\|^2,$$

and

$$\begin{aligned}\mathbb{E}[c^2\delta_t(k)^2\|(I-\Phi_k)X\|^2|\mathcal{F}_{k-1}] &= \mathbb{E}[c^2\delta_t(k)^2(\|X\|^2 - \|\Phi_k X\|^2)|\mathcal{F}_{k-1}] \\ &= c^2\sum_{i=1}^p \mathbb{E}[\delta_t(k)^2 - \delta_t(k)^2\phi_{ii,k}|\mathcal{F}_{k-1}]x_i^2 \\ &= 2\tilde{c}^2\sum_{i=1}^p (1-p_i)x_i^2 \leq 2\tilde{c}^2(1-\pi)\|X\|^2,\end{aligned}$$

thus proving Eq. (4.2.8). □

The following Theorem has been adapted from [**164**].

Theorem 4.2.3. ***(Conditional form of Jensen's inequality)****[***164***], Ch. 7, Theorem 10] Let $(\Omega, \mathcal{G}, P)$ be a probability space and let g be a convex function over $(-\infty, +\infty)$ and X be a random variable such that X and $g(X)$ have finite expectations. If $\mathcal{B}$ is any*

sub-sigma field of $\mathcal{G}$*, then*

$$g(\mathbb{E}[X|\mathcal{B}]) \leq \mathbb{E}[g(X)|\mathcal{B}] \quad a.s.$$

The following result is a variation of the standard convergence result for the iterates of the centralized gradient descent method (eg. [[**115**], Theorem 2.1.15]), and we will employ it in our main analysis to determine a suitable range for the proximity parameter α.

Lemma 4.2.4. ***(Contraction)*** *Under Assumptions 4.1.2 and 4.1.3, consider the difference* $\Delta_\alpha(X,Y) = X - Y - \alpha(\nabla F(X) - \nabla F(Y))$ *for any two vectors* $X, Y \in \mathbb{R}^n$*. Then if* $\alpha < 2/(\mu + L)$*, the following inequality holds*

$$\|\Delta_\alpha(X,Y)\|^2 \leq \gamma^2 \|X - Y\|^2,$$

where $\gamma^2 = 1 - \frac{2\alpha\mu L}{\mu + L}$*.*

Proof. Taking the squared norm of $\Delta_\alpha(X,Y)$ yields

$$\begin{aligned}\|\Delta_\alpha(X,Y)\|^2 &= \|X - Y\|^2 + \alpha^2\|\nabla F(X) - \nabla F(Y)\|^2 - 2\alpha\langle X - Y, \nabla F(X) - \nabla F(Y)\rangle \\ &\leq \|X - Y\|^2 + \alpha^2\|\nabla F(X) - \nabla F(Y)\|^2 \\ &\quad - \frac{2\alpha\mu L}{\mu + L}\|X - Y\|^2 - \frac{2\alpha}{\mu + L}\|\nabla F(X) - \nabla F(Y)\|^2,\end{aligned}$$

where we used [[**115**], Theorem 2.1.12] to bound the inner product in the first equality.

Observing that the coefficient of the term $\|\nabla F(X) - \nabla F(Y)\|^2$ is strictly negative concludes the proof. □

We conclude this subsection with the following Lemma adapted from [**125**]. We will use Lemma 4.2.5 in the next subsection to show that Algorithm 2 acts like a contractive operator on the expectations of the errors listed in Eq. (4.2.2).

Lemma 4.2.5. *[[**125**], Lemma 5] Let $S = [s_{ij}] \in \mathbb{R}^{3\times 3}$ be a nonnegative, irreducible matrix with $s_{ii} < \lambda_*$ for some $\lambda_* > 0$ for $i = 1, 2, 3$. Then $\rho(S) < \lambda_\star$ iff $det(\lambda_* I - S) > 0$.*

4.2.2. Main analysis

We begin our analysis by deriving an upper bound for the displacement (up to a scaling factor) Q_{k-1} between two consecutive iterates X_k and X_{k-1} in Eq. (4.2.3) with respect to the errors $\mathcal{E}^\star, \mathcal{E}^b$ and $\mathcal{E}^w$ defined in Eq. (4.2.2). As a result of Lemma 4.2.6, we obtain a range for the proximity parameter α of Algorithm 2.

Lemma 4.2.6. ***(Displacement)*** *Let $Q_k = -\alpha W^{-1}\nabla F_\alpha(X_k) + \mathcal{E}_k^b + \Delta_\alpha(WX_k^b, WX_k) + \Delta_\alpha(X_k^w, WX_k^b)$ and suppose that the parameter α satisfies $\alpha < 2/(\mu + L)$. Then for all $k \geq 1$ we have*

$$\|Q_k\| \leq \alpha\beta^{-1}L_\alpha\|\mathcal{E}_k^\star\| + (1+\gamma)\|\mathcal{E}_k^b\| + \gamma\|\mathcal{E}_k^w\|,$$

where β is the smallest eigenvalue of the consensus matrix W, L_α is defined in Lemma (4.2.1) *and $\gamma = \sqrt{1 - \frac{2\alpha\mu L}{\mu + L}}$ where L and μ are defined in Assumptions 4.1.2 and 4.1.3, respectively.*

Proof. Taking the norm of Q_k and applying the triangle inequality yields

$$\|Q_k\| \leq \alpha\|W^{-1}\nabla F_\alpha(X_k)\| + \|\mathcal{E}_k^b\| + \|\Delta_\alpha(WX_k^b, WX_k)\| + \|\Delta_\alpha(X_k^w, WX_k^b)\|$$

$$\leq \alpha\beta^{-1}\|\nabla F_\alpha(X_k)\| + \|\mathcal{E}_k^b\| + \|\Delta_\alpha(WX_k^b, WX_k)\| + \|\Delta_\alpha(X_k^w, WX_k^b)\|,$$

where the last inequality holds due to the spectral properties of W.

Applying the Lipschitz gradient continuity of F_α and Lemma 4.2.1 on the first term of the preceding relation and Lemma 4.2.4 on the last two terms concludes the proof. □

Next, we derive a bound on the expected distance to optimality with respect to the expected errors defined in Eq. (4.2.2) at the previous iteration. We derive similar results for the buffer error and the consensus error defined in Eq. (4.2.2) in Lemmas 4.2.8 and 4.2.9, respectively.

Lemma 4.2.7. ***(Distance to optimality)*** *Suppose that the parameter α in* (4.1.2) *satisfies $\alpha < 2/(\mu+L)$. Then the expected norm of the distance to optimality $\mathcal{E}_k^\star = X_k - X^\star$ satisfies the following relation for all $k \geq 1$*

$$\mathbb{E}[\|\mathcal{E}_k^\star\|] \leq (\sqrt{1 + 2\tilde{c}^2 - 2\tilde{c}} + \tilde{c}\gamma)\mathbb{E}[\|\mathcal{E}_{k-1}^\star\|] + \tilde{c}(1+\gamma)\mathbb{E}[\|\mathcal{E}_{k-1}^b\|] + \tilde{c}\gamma\mathbb{E}[\|\mathcal{E}_{k-1}^w\|],$$

where $\tilde{c} = c\left(\sum_i \lambda_i\right)^{-1}$ and $\gamma = \sqrt{1 - \frac{2\alpha\mu L}{\mu+L}}$.

Proof. Subtracting $X^\star$ from (4.2.3) and taking the norm on both sides yields

(4.2.9)

$$\begin{aligned}\|\mathcal{E}_k^\star\| &\leq \|\mathcal{E}_{k-1}^\star - \alpha c\delta_t(k) W^{-1}\nabla F_\alpha(X_{k-1})\| + c\delta_t(k)\|\mathcal{E}_{k-1}^b\| \\ &\quad + c\delta_t(k)\|\Delta_\alpha(WX_{k-1}^b, WX_{k-1})\| + c\delta_t(k)\|\Delta_\alpha(X_{k-1}^w, WX_{k-1}^b)\| \\ &\leq \|\mathcal{E}_{k-1}^\star - \alpha c\delta_t(k) W^{-1}\nabla F_\alpha(X_{k-1})\| + c\delta_t(k)(1+\gamma)\|\mathcal{E}_{k-1}^b\| + c\delta_t(k)\gamma\|\mathcal{E}_{k-1}^w\|,\end{aligned}$$

where we applied Lemma 4.2.4 to get the last inequality.

For the first term in (4.2.9) we have

$$\begin{aligned}\|\mathcal{E}_{k-1}^\star - \alpha c\delta_t(k) W^{-1}\nabla F_\alpha(X_{k-1})\| &= \|\mathcal{E}_{k-1}^\star + c\delta_t(k)(WX_{k-1} - X_{k-1} - \alpha\nabla F(WX_{k-1}))\| \\ &= \|(1 - c\delta_t(k))(\mathcal{E}_{k-1}^\star) + c\delta_t(k)\Delta_\alpha(WX_{k-1}, WX^\star)\| \\ &\leq (|1 - c\delta_t(k)| + c\delta_t(k)\gamma)\|\mathcal{E}_{k-1}^\star\|,\end{aligned}$$

where the second equality follows from the optimality of $X^\star$ and we obtain the last inequality by applying the triangle inequality and Lemma 4.2.4.

Substituting the preceding relation back in (4.2.9) yields

$$\|\mathcal{E}_k^\star\| \leq (|1 - c\delta_t(k)| + c\delta_t(k)\gamma)\|\mathcal{E}_{k-1}^\star\| + c\delta_t(k)(1+\gamma)\|\mathcal{E}_{k-1}^b\| + c\delta_t(k)\gamma\|\mathcal{E}_{k-1}^w\|.$$

We take the expectation conditional on $\mathcal{F}_{k-1}$ on both sides of the preceding relation and apply Eq. (4.2.6) of Lemma 4.2.2 to obtain

$$\mathbb{E}[\|\mathcal{E}_k^\star\||\mathcal{F}_{k-1}] \leq (\mathbb{E}[|1 - c\delta_t(k)||\mathcal{F}_{k-1}] + \tilde{c}\gamma)\|\mathcal{E}_{k-1}^\star\| + \tilde{c}(1+\gamma)\|\mathcal{E}_{k-1}^b\| + \tilde{c}\gamma\|\mathcal{E}_{k-1}^w\|.$$

Eq. (4.2.6) of Lemma 4.2.2 yields $\mathbb{E}[(1-c\delta_t(k))^2|\mathcal{F}_{k-1}] = \mathbb{E}[1+c^2\delta_t^2(k)-2c\delta_t(k)|\mathcal{F}_{k-1}] = 1+2\tilde{c}^2-2\tilde{c}$, and thus by Theorem 4.2.3 applied on the convex function $g(x) = -\sqrt{x}$ we have $\mathbb{E}[|1-c\delta_t(k)||\mathcal{F}_{k-1}] \leq \sqrt{1+2\tilde{c}^2-2\tilde{c}}$. Substituting this bound in the preceding relation and taking the total expectation on both sides completes the proof. □

Lemma 4.2.8. ***(Buffer error)*** *Under parameter $\alpha < 2/(\mu+L)$ in (4.1.2), the following relation holds for the buffer error $\mathcal{E}_k^b = X_k - X_k^b$ for all $k \geq 1$*

$$\mathbb{E}[\|\mathcal{E}_k^b\|] \leq \sqrt{1-\pi}\big(1+\tilde{c}\sqrt{2}(1+\gamma)\big)\mathbb{E}[\|\mathcal{E}_{k-1}^b\|] + \alpha\beta^{-1}\tilde{c}\sqrt{2(1-\pi)}L_\alpha\mathbb{E}[\|\mathcal{E}_k^\star\|] + \tilde{c}\sqrt{2(1-\pi)}\gamma\mathbb{E}[\|\mathcal{E}_k^w\|],$$

where $\pi = \min_j \lambda_j \left(\sum_i \lambda_i\right)^{-1}$ is the minimum activation probability among agents, $\tilde{c} = c\left(\sum_i \lambda_i\right)^{-1}$, $\gamma = \sqrt{1-\frac{2\alpha\mu L}{\mu+L}}$, β is the smallest eigenvalue of the consensus matrix W and L_α is defined in Lemma 4.2.1.

Proof. Subtracting (4.2.4) from (4.2.3) yields

$$\mathcal{E}_k^b = (I-\Phi_k)\mathcal{E}_{k-1}^b + c\delta_t(k)(I-\Phi_k)Q_{k-1}.$$

After taking the norm on both sides of the preceding relation and applying the triangle inequality, we obtain

$$\|\mathcal{E}_k^b\| \leq \|(I-\Phi_k)\mathcal{E}_{k-1}^b\| + c\delta_t(k)\|(I-\Phi_k)Q_{k-1}\|.$$

Taking the expectation conditional on $\mathcal{F}_{k-1}$ on both sides of the relation above and invoking Eq. (4.2.7) and (4.2.8) of Lemma 4.2.2 and Theorem 4.2.3 applied on the convex

function $g(x) = -\sqrt{x}$ yields,

$$\begin{aligned}\mathbb{E}[\|\mathcal{E}_k^b\||\mathcal{F}_{k-1}] &\leq \sqrt{1-\pi}\|\mathcal{E}_{k-1}^b\| + \tilde{c}\sqrt{2(1-\pi)}\|Q_{k-1}\| \\ &\leq \sqrt{1-\pi}\|\mathcal{E}_{k-1}^b\| + \tilde{c}\sqrt{2(1-\pi)}\big(\alpha\beta^{-1}L_\alpha\|\mathcal{E}_k^\star\| + (1+\gamma)\|\mathcal{E}_k^b\| + \gamma\|\mathcal{E}_k^w\|\big),\end{aligned}$$

where we applied Lemma 4.2.6 to get the last inequality.

Taking the total expectation on both sides of the preceding relation concludes the proof. □

Lemma 4.2.9. ***(Consensus error)*** *Under parameter $\alpha < 2/(\mu + L)$ in (4.1.2), the following relation holds for the consensus error $\mathcal{E}_k^w = WX_k^b - X_k^w$ for all $k \geq 1$*

$$\begin{aligned}\mathbb{E}[\|\mathcal{E}_k^w\|] \leq (\sqrt{1-\pi} + \tilde{c}(1-m)\sqrt{2\Pi}\gamma)\mathbb{E}[\|\mathcal{E}_k^w\|] \\ + (1-m)\sqrt{\Pi}(1+\tilde{c}\sqrt{2}(1+\gamma))\mathbb{E}[\|\mathcal{E}_{k-1}^b\|] + \alpha\beta^{-1}\tilde{c}(1-m)\sqrt{2\Pi}L_\alpha\mathbb{E}[\|\mathcal{E}_{k-1}^\star\|],\end{aligned}$$

where $\pi = \min_j \lambda_j \left(\sum_i \lambda_i\right)^{-1}$ is the minimum activation probability among agents, m is the smallest diagonal element of the consensus matrix W, $\Pi = \max_j \lambda_j \left(\sum_i \lambda_i\right)^{-1}$ is the maximum activation probability among agents, $\tilde{c} = c\left(\sum_i \lambda_i\right)^{-1}$, $\gamma = \sqrt{1 - \frac{2\alpha\mu L}{\mu + L}}$, β is the smallest eigenvalue of the consensus matrix W and L_α is defined in Lemma 4.2.1.

Proof. Multiplying Eq. (4.2.4) with W and subtracting Eq. (4.2.5) yields

$$\mathcal{E}_k^w = (I - \Phi_k)\mathcal{E}_{k-1}^w + (W - W_d)\Phi_k\mathcal{E}_{k-1}^b + c\delta_t(k)(W - W_d)\Phi_k Q_{k-1},$$

where we used the fact that Φ_k and W_d are diagonal and their multiplication is commutative.

Taking the norm on both sides of the preceding relation and applying the triangle inequality yields

$$\begin{aligned}\|\mathcal{E}_k^w\| &\leq \|(I-\Phi_k)\mathcal{E}_k^w\| + \|(W-W_d)\Phi_k\mathcal{E}_{k-1}^b\| + c\delta_t(k)\|(W-W_d)\Phi_k Q_{k-1}\| \\ &\leq \|(I-\Phi_k)\mathcal{E}_k^w\| + (1-m)\|\Phi_k\mathcal{E}_{k-1}^b\| + c\delta_t(k)(1-m)\|\Phi_k Q_{k-1}\|,\end{aligned}$$

where the last inequality follows from the fact that the spectral radius of a non-negative matrix is bounded by its maximum row sum [[**69**], Lemma 8.1.21].

We take the expectation conditional on $\mathcal{F}_{k-1}$ on both sides of the relation above and apply Eq. (4.2.7) and Eq. (4.2.8) of Lemma 4.2.2 in conjuction with Theorem 4.2.3 applied for the convex function $g(x) = -\sqrt{x}$ to obtain

$$\begin{aligned}\mathbb{E}[\|\mathcal{E}_k^w\||\mathcal{F}_{k-1}] &\leq \sqrt{1-\pi}\|\mathcal{E}_k^w\| + (1-m)\sqrt{\Pi}\|\mathcal{E}_{k-1}^b\| + \tilde{c}(1-m)\sqrt{2\Pi}\|Q_{k-1}\| \\ &\leq (\sqrt{1-\pi} + \tilde{c}(1-m)\sqrt{2\Pi}\gamma)\|\mathcal{E}_k^w\| + (1-m)\sqrt{\Pi}(1+\tilde{c}\sqrt{2}(1+\gamma))\|\mathcal{E}_{k-1}^b\| \\ &\quad + \alpha\beta^{-1}\tilde{c}(1-m)\sqrt{2\Pi}L_\alpha\|\mathcal{E}_{k-1}^\star\|,\end{aligned}$$

where we invoked Lemma 4.2.6 to get the last inequality.

Taking the total expectation on both sides of the preceding relation completes the proof. □

We conclude our theoretical results with the following Theorem which proves the convergence of Algorithm 2.

Theorem 4.2.10. ***(Convergence)*** *Under Assumptions 4.1.1-4.1.3, suppose that* $\alpha < 2/(\mu + L)$ *in Eq.* (4.1.2) *and that the quantity* $\tilde{c} = c\left(\sum_i \lambda_i\right)^{-1}$ *satisfies*

$$\tilde{c} < \min\left\{\frac{2(1-\gamma)}{2-\gamma^2}, \frac{1-\sqrt{1-\pi}}{(1+\gamma)\sqrt{2(1-\pi)}}, \frac{1-\sqrt{1-\pi}}{(1-m)\gamma\sqrt{2\Pi}}\right\}, \tag{4.2.10}$$

where $\gamma = \sqrt{1 - \frac{2\alpha\mu L}{\mu+L}}$, $\pi = \min_j \lambda_j \left(\sum_i \lambda_i\right)^{-1}$ *is the minimum activation probability among agents,* $\Pi = \max_j \lambda_j \left(\sum_i \lambda_i\right)^{-1}$ *is the maximum activation probability among agents and* m *is the smallest diagonal element of the consensus matrix* W*.*

Then there exist positive constants C *and* B *and a scalar* $\rho \in (0,1)$ *such that if* $\tilde{c} < C$*, the distance to optimality* $\mathcal{E}_k^\star = X_k - X^\star$ *where* $X^\star = \arg\min_X F_\alpha(X)$*, the buffer error* $\mathcal{E}_k^b = X_k - X_k^b$ *and the consensus error* $\mathcal{E}_k^w = WX_k^b - X_k^w$ *satisfy*

$$\mathbb{E}[\|\mathcal{E}_k^\star\|] \leq \rho^k B, \quad \mathbb{E}[\|\mathcal{E}_k^b\|] \leq \rho^k B, \quad \mathbb{E}[\|\mathcal{E}_k^w\|] \leq \rho^k B.$$

Proof. Combining Lemmas 4.2.7-4.2.9 we construct the following system of linear inequalities

$$\begin{bmatrix} \mathbb{E}[\|\mathcal{E}_k^\star\|] \\ \mathbb{E}[\|\mathcal{E}_k^b\|] \\ \mathbb{E}[\|\mathcal{E}_k^w\|] \end{bmatrix} \leq M \begin{bmatrix} \mathbb{E}[\|\mathcal{E}_{k-1}^\star\|] \\ \mathbb{E}[\|\mathcal{E}_{k-1}^b\|] \\ \mathbb{E}[\|\mathcal{E}_{k-1}^w\|] \end{bmatrix},$$

where the matrix $M \in \mathbb{R}^{3\times 3}$ is given by

$$M = \begin{bmatrix} \sqrt{1+2\tilde{c}^2-2\tilde{c}} + \tilde{c}\gamma & \tilde{c}(1+\gamma) & \tilde{c}\gamma \\ \alpha\beta^{-1}\tilde{c}\sqrt{2(1-\pi)}L_\alpha & \sqrt{1-\pi}\big(1+\tilde{c}\sqrt{2}(1+\gamma)\big) & \tilde{c}\sqrt{2(1-\pi)}\gamma \\ \alpha\beta^{-1}\tilde{c}(1-m)\sqrt{2\Pi}L_\alpha & (1-m)\sqrt{\Pi}(1+\tilde{c}\sqrt{2}(1+\gamma)) & \sqrt{1-\pi}+\tilde{c}(1-m)\sqrt{2\Pi}\gamma \end{bmatrix}.$$

We will show that the spectral radius $\rho(M)$ of M satisfies $\rho(M) < 1$ for small enough $\tilde{c}$. The determinant of $I - M$ is

$$\begin{aligned} det(I - M) = (1 - m_{11})\Big((1 - m_{22})(1 - m_{33}) - m_{23}m_{32}\Big) \\ - m_{12}\Big(m_{21}(1 - m_{33}) + m_{23}m_{31}\Big) \\ - m_{13}\Big(m_{21}m_{32} + (1 - m_{22})m_{31}\Big). \end{aligned}$$

We first note that all elements of M are positive and that $m_{ii} < 1$ for $i = 1, 2, 3$ due to Eq. 4.2.10 . Moreover, it is easy to verify that the following relation holds for any positive scalar u

$$1 - u \leq \sqrt{1 + 2u^2 - 2u} \leq 1 - u + u^2,$$

and thus we can construct a lower bound L_B for $det(I - M)$ as follows,

$$\begin{aligned} \det(I - M) \geq L_B = m_L(1 - m_{22})(1 - m_{33}) - m_U m_{23}m_{32} \\ - m_{12}\Big(m_{21}(1 - m_{33}) + m_{23}m_{31}\Big) \\ - m_{13}\Big(m_{21}m_{32} + (1 - m_{22})m_{31}\Big), \end{aligned}$$

where $m_U = \tilde{c}(1 - \gamma)$ and $m_L = \tilde{c}(1 - \gamma - \tilde{c})$.

We observe that the quantity $\tilde{c}^{-1}L_B$ is a 3^{rd} degree polynomial of $\tilde{c}$, i.e. it can be written in the form $\tilde{c}^{-1}L_B = P(\tilde{c}) = a_0 + a_1\tilde{c} + a_2\tilde{c}^2 + a_3\tilde{c}^3$, where the coefficients a_0 and

a_3 are given by

$$a_0 = (1-\gamma)(1+\sqrt{1-\pi})^2 > 0$$

$$a_3 = -2\gamma(1-m)(1+\gamma)\sqrt{\Pi(1-\pi)} < 0.$$

To prove $\rho(M) < 1$ using Lemma 4.2.5, it suffices to guarantee that $P(\tilde{c}) > 0$ in the range of $\tilde{c}$. Due to $a_3 < 0$, there exists a positive scalar x such that $P(x) < 0$. Moreover, we have $P(0) = a_0 > 0$ and $P(\tilde{c})$ is continuous in the interval $[0, x]$. Hence, by the Intermediate Value Theorem the polynomial $P(\tilde{c})$ has at least one root r in the interval $(0, x)$ such that $P(\tilde{c}) > 0$ in $[0, r)$, and a range of small enough values of $\tilde{c}$ such that both $\tilde{c} < r$ and Eq. 4.2.10 is satisfied is guaranteed to exist (we note that the closed form of this range can be calculated with the cubic root formula). Hence, $det(I - M) \geq L_B = \tilde{c}^{-1}P(\tilde{c}) > 0$ and $\rho(M) < 1$ by Lemma 4.2.5, which implies that for all $k \geq 1$ we have

$$\begin{bmatrix} \mathbb{E}[\|\mathcal{E}_k^\star\|] \\ \mathbb{E}[\|\mathcal{E}_k^b\|] \\ \mathbb{E}[\|\mathcal{E}_k^w\|] \end{bmatrix} \leq M^k \begin{bmatrix} \|\mathcal{E}_0^\star\| \\ \|\mathcal{E}_0^b\| \\ \|\mathcal{E}_0^w\| \end{bmatrix},$$

or for $\mathcal{E}_k$ being any of $\mathcal{E}_k^\star, \mathcal{E}_k^b, \mathcal{E}_k^w$

$$\mathbb{E}[\|\mathcal{E}_k\|] \leq (\rho(M))^k \sqrt{\|\mathcal{E}_0^\star\|^2 + \|\mathcal{E}_0^b\|^2 + \|\mathcal{E}_0^w\|^2}.$$

Setting $\rho = \rho(M)$ and $B = \sqrt{\|\mathcal{E}_0^\star\|^2 + \|\mathcal{E}_0^b\|^2 + \|\mathcal{E}_0^w\|^2}$ concludes the proof. □

We note that Theorem 4.2.10 implies both component-wise convergence to zero in the mean sense for the errors $\mathcal{E}_k^\star$, $\mathcal{E}_k^b$ and $\mathcal{E}_k^w$ due to $\|v\|_1 \leq \sqrt{n}\|v\|$ for all vectors $v \in \mathbb{R}^n$, and

convergence in probability to zero for their norms by the Markov inequality, i.e. for $\mathcal{E}_k$ being any of the errors $\mathcal{E}_k^\star$, $\mathcal{E}_k^b$ and $\mathcal{E}_k^w$ and all $\epsilon > 0$ we have

$$\mathcal{P}\left(\|\mathcal{E}_k\| > \epsilon\right) \leq \frac{\rho^k B}{\epsilon} \to 0.$$

4.3. Numerical Results

We evaluated the performance of Algorithm 2 on a 2-dimensional quadratic problem

$$\min_{x\in\mathbb{R}^2} \sum_{i=1}^{n} \left(\frac{1}{2}\|x\|_{Q^i}^2 + b^i x\right), \tag{4.3.1}$$

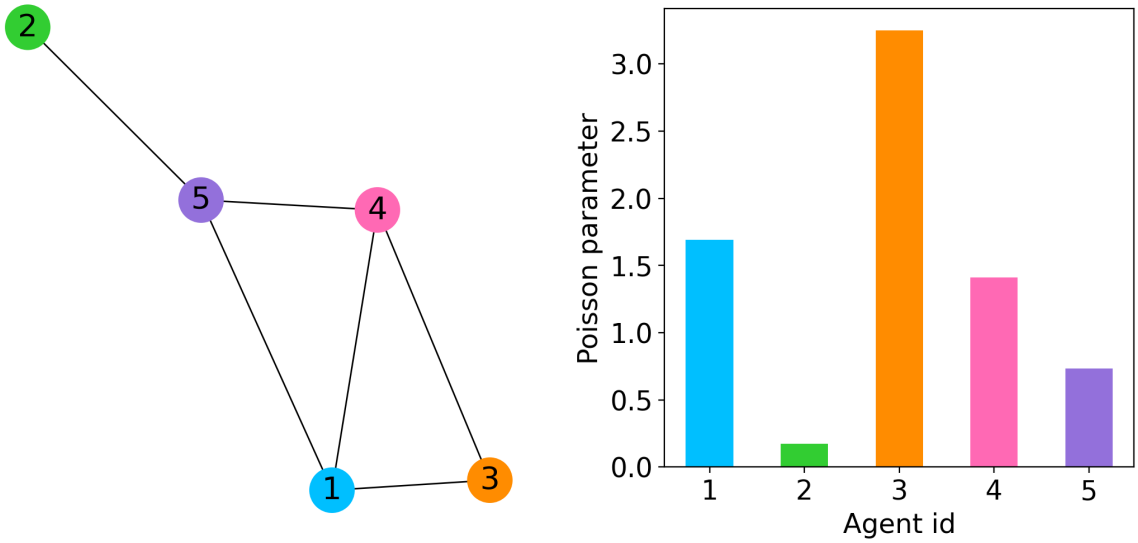
where the function $f_i(x) = \frac{1}{2}\|x\|_{Q^i}^2 + b^i x$ is assigned to agent $i \in \mathcal{V}$.

Each vector $b^i \in \mathbb{R}^2$ for $i = 1, ..., n$ was randomly initialized in the interval $[-1, 1]$. The matrices $Q^i \in \mathbb{R}^{2\times 2}$ were generated as follows: for each agent $i \in \mathcal{V}$ we generated a random orthonormal matrix $O^i \in \mathbb{R}^{2\times 2}$ and set $Q^i = \xi^i \cdot O^i \cdot diag([1, \kappa]') \cdot (O^i)^{-1}$, where $\xi^i \in (0, 1)$ is a random seed unique to each agent and $\kappa = L/\mu = 10^2$ is the global condition number of Problem 4.3.1. We opted for a network of size $n = 5$ with random graph topology (Erdős–Rényi with edge probability 0.5) shown in Fig. 4.1a.

To construct the local Poisson clocks, we randomly initialized the rates λ_i (arrivals per second) for each agent by sampling the positive side of the standard normal distribution, resulting in the values shown in Fig. 4.1b. Let $t_{i,j}$ be the time of the j^{th} activation of agent i; then $t_{i,j} = t_{i,j-1} + s_i$, where s_i is a random sample from an exponential distribution with parameter λ_i. The global clock was created by merging and sorting the activation times $t_{i,j}$ for all $i \in \mathcal{V}$ and values of j. We terminated the experiment after $T = 2 \cdot 10^3$

seconds. The agents were randomly initialized on the 2-dimensional plane within the interval $[-15, 15]$ (in meters). We set $\alpha = 2/(\mu + L)$ and $c = 5 \cdot 10^{-2}$ in Eq. (4.1.2).

The results of a typical run of the experiment are shown in Figures 4.2 and 4.3. In Fig. 4.2a, we plot the following quantities over the entire duration of the experiment: *i*) the distance between the solution $x^\star \in \mathbb{R}^2$ of Problem 1.0.1 and the average position $\bar{x}_t \in \mathbb{R}^2$ at time $t \in [0, T]$, i.e. $\bar{x}_t = n^{-1} \sum_{i=1}^n x_{i,t}$ where $x_{i,t} \in \mathbb{R}^2$ is the position of agent $i \in \mathcal{V}$ at time t (solid blue line); *ii*) the distance to rendezvous over time, i.e. the average $\frac{1}{n} \sum_{i=1}^n \|\bar{x}_t - x_{i,t}\|^2$ of the distances between the local positions $x_{i,t}$ and $\bar{x}_t$ (dashed orange line); and *iii*) the gradient norm $\|\nabla F_\alpha(X_t)\|$ of the function F_α 4.2.1 where $X_t = [x'_{1,t}, ..., x'_{2,t}]' \in \mathbb{R}^{2n}$ is the column-wise concatenation of the local positions $x_{i,t}$. Our numerical results confirm our theoretical analysis, namely that the system-wide iterates $X_t \in \mathbb{R}^{np}$ of Algorithm 2 converge to the minimum of F_α with linear rate, while the local iterates $x_{i,t} \in \mathbb{R}^p$ converge to a neighborhood of the optimal solution $x^\star$ of Problem 1.0.1 while achieving approximate rendezvous. In Fig. 4.2b we plot the norms of each agent's velocity for $t \in [0, 200]$, i.e. $\|c \cdot g_{i,t}\|$. We observe in Fig. 4.2b that agents naturally decrease their velocities as they approach rendezvous and the optimal solution $x^\star$ of Problem 1.0.1; moreover, agents that are activated more frequently have smoother velocity curves. Finally, to facilitate the visual interpretation of our results, we have plotted snapshots of the trajectories of all agents for time instances $t = \{0, 45.24, 114.61, 294.07\}$ in Figure 4.3. The agents are color-coded as in Fig. 4.1, and the solution $x^\star$ of Problem 1.0.1 is plotted with a red "x" marker.

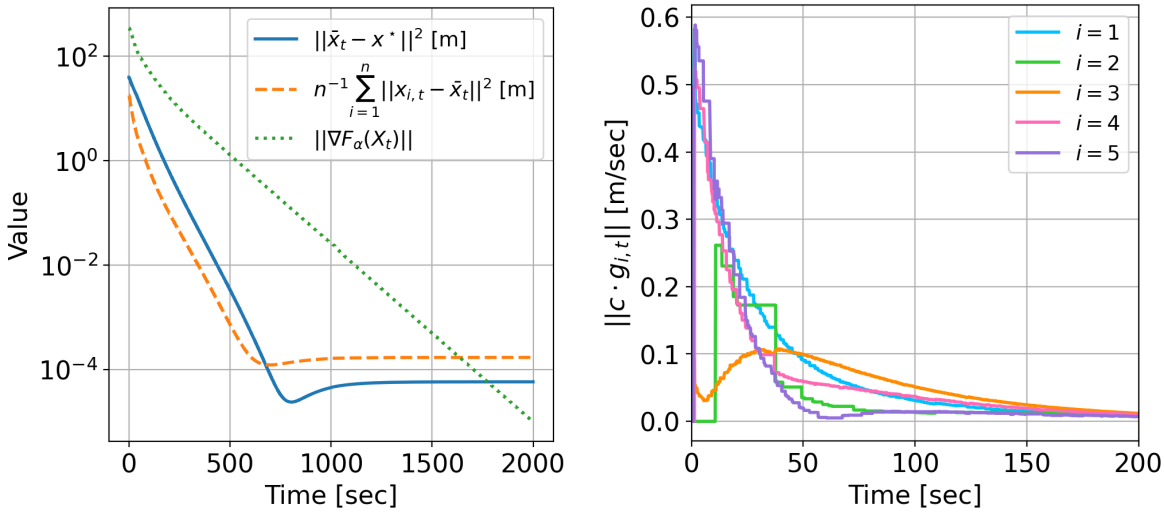


(a) Network topology (b) Poisson parameters λ_i

Figure 4.1. Network topology (left) and Poisson parameter values λ_i for agents $i = 1, 2, 3, 4, 5$ (right).

4.4. Summary

We considered a generalized consensus optimization version of the multi-agent rendezvous problem, where each agent is associated with a local cost function and the optimal rendezvous point minimizes the sum of the local cost functions. In our setting, agents randomly and independently alternate between two non-overlapping states: i) an active state, where they can sense their current positions, broadcast messages to their neighbors, access their local buffers where outdated information from their neighbors is stored, and adjust their velocities; and ii) an inactive state, where they continue to move towards the direction they calculated in their most recent active state and passively listen for messages. We proposed a fully asynchronous distributed algorithm for reaching rendezvous over fixed, undirected networks of mobile agents that is robust to outdated

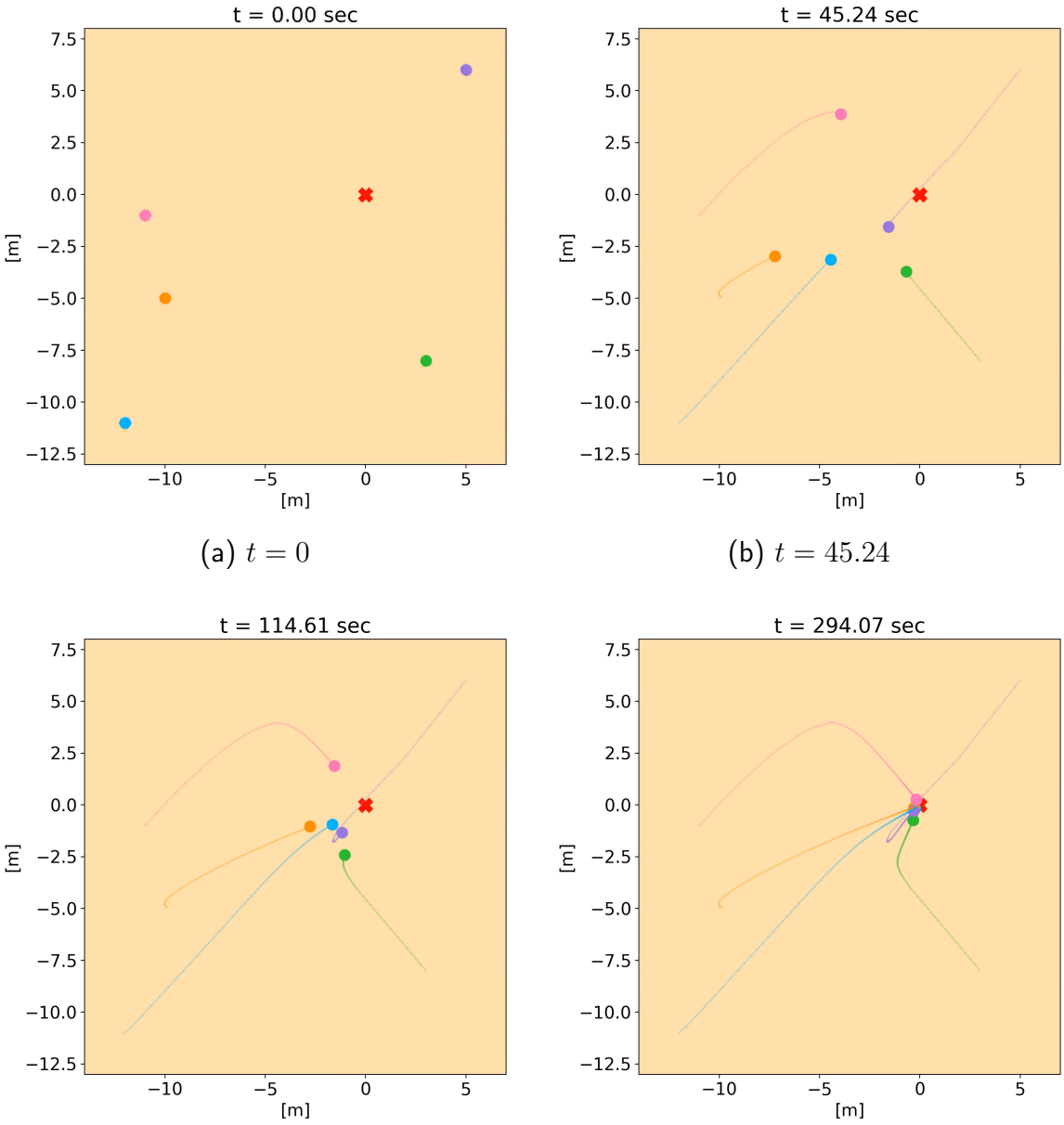


(a) Distances to optimality and to rendezvous and gradient norm of F_α

(b) Agent velocities

Figure 4.2. Distance of average position $\bar{x}_t = n^{-1}\sum_{i=1}^{n} x_{i,t}$ at time $t \in [0, T]$ to the solution $x^\star$ of Problem 1.0.1 (left, solid blue line), distance to rendezvous (left, dashed orange line) and gradient norm $\nabla F_\alpha(X_t)$ where $X_t = [x'_{1,t}, ..., x_{5,t}]'$ (left, dotted green line) and velocity norms for agents $i = 1, 2, 3, 4, 5$ in the interval $t \in [0, 200]$ seconds (right).

information and erroneous displacements caused by inactive states, and provided probabilistic guarantees for its convergence; namely we have shown that under appropriate selection of parameters, our algorithm converges in the mean sense to an arbitrarily small neighborhood of the optimal rendezvous point while achieving approximate rendezvous. Our numerical simulations have confirmed our theoretical findings.



(a) $t = 0$ (b) $t = 45.24$

(c) $t = 114.61$ (d) $t = 294.07$

Figure 4.3. Trajectory snapshots for time instances $t = 0$ (top left), $t = 45.24$ (top right), $t = 114.61$ (bottom left) and $t = 294.07$ (bottom right). The agents $i = 1, 2, 3, 4, 5$ are color-coded as in Fig. 4.1 and the optimal solution $x^\star$ of Problem 1.0.1 is plotted with a red "x" marker.

References

[1] Absil, P.-A., Mahony, R., and Sepulchre, R. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, NJ, 2008.

[2] Acevedo, J. J., Arrue, B. C., Maza, I., and Ollero, A. A decentralized algorithm for area surveillance missions using a team of aerial robots with different sensing capabilities. In *2014 IEEE International Conference on Robotics and Automation (ICRA)* (2014), pp. 4735–4740.

[3] Alistarh, D., Grubic, D., Li, J., Tomioka, R., and Vojnovic, M. Qsgd: Communication-efficient sgd via gradient quantization and encoding. In *Advances in Neural Information Processing Systems* (2017), pp. 1709–1720.

[4] Alistarh, D., Hoefler, T., Johansson, M., Khirirat, S., Konstantinov, N., and Renggli, C. The convergence of sparsified gradient methods. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (Red Hook, NY, USA, 2018), NIPS'18, Curran Associates Inc., p. 5977–5987.

[5] Alpcan, T., and Bauckhage, C. A distributed machine learning framework. In *Proceedings of the 48h IEEE Conference on Decision and Control (CDC) held jointly with 2009 28th Chinese Control Conference* (Dec 2009), pp. 2546–2551.

[6] Ando, H., Oasa, Y., Suzuki, I., and Yamashita, M. Distributed memoryless point convergence algorithm for mobile robots with limited visibility. *IEEE transactions on robotics and automation 15*, 5 (1999), 818–828.

[7] Attouch, H., and Bolte, J. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Math. Program. 116*, 1–2 (Jan. 2009), 5–16.

[8] Attouch, H., Bolte, J., Redont, P., and Soubeyran, A. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the kurdyka-łojasiewicz inequality. *Math. Oper. Res. 35*, 2 (May 2010), 438–457.

[9] Attouch, H., Bolte, J., and Svaiter, B. F. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods. *Mathematical Programming 137*, 1-2 (Feb. 2013), 91–129.

[10] Aysal, T., Coates, M., and Rabbat, M. Distributed Average Consensus With Dithered Quantization. *IEEE Transactions on Signal Processing 56*, 10 (Oct. 2008), 4905–4918.

[11] Bach, F., and Moulines, E. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Proceedings of the 24th International Conference on Neural Information Processing Systems* (USA, 2011), NIPS'11, Curran Associates Inc., pp. 451–459.

[12] Beard, R. W., and Stepanyan, V. Information consensus in distributed multiple vehicle coordinated control. In *42nd IEEE International Conference on Decision and Control (IEEE Cat. No.03CH37475)* (Dec 2003), vol. 2, pp. 2029–2034 Vol.2.

[13] Berahas, A. S., Bollapragada, R., Keskar, N. S., and Wei, E. Balancing Communication and Computation in Distributed Optimization. *IEEE Transactions on Automatic Control 64*, 8 (Aug. 2019), 3141–3155.

[14] Berahas, A. S., Bollapragada, R., and Wei, E. On the convergence of nested decentralized gradient methods with multiple consensus and gradient steps, 2020.

[15] Berahas, A. S., Iakovidou, C., and Wei, E. Nested distributed gradient methods with adaptive quantized communication, 2019.

[16] Bertsekas, D. Nonlinear programming. *Athena Scientific 48* (01 1995).

[17] Bertsekas, D. P., and Tsitsiklis, J. N. *Parallel and Distributed Computation: Numerical Methods*. Prentice-Hall, Inc., USA, 1989.

[18] Bianchi, P., Hachem, W., and Franck, I. A stochastic coordinate descent primal-dual algorithm and applications. In *2014 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)* (Sep. 2014), pp. 1–6.

[19] Bianchi, P., and Jakubowicz, J. Convergence of a multi-agent projected stochastic gradient algorithm for non-convex optimization. *IEEE Transactions on Automatic Control 58*, 2 (2013), 391–405.

[20] Bijral, A. S. Data Dependent Convergence for Distributed Stochastic Optimization. *arXiv:1608.08337 [cs, math, stat]* (Aug. 2016). arXiv: 1608.08337.

[21] Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, V., Kiddon, C., Konečný, J., Mazzocchi, S., McMahan, H. B., Overveldt, T. V., Petrou, D., Ramage, D., and Roselander, J. Towards federated learning at scale: System design, 2019.

[22] Bottou, L. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010* (Heidelberg, 2010), Y. Lechevallier and G. Saporta, Eds., Physica-Verlag HD, pp. 177–186.

[23] Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Foundations and Trends in Machine Learning 3*, 1 (2010), 1–122.

[24] Bullo, F., Cortés, J., and Martinez, S. *Distributed Control of Robotic Networks: a Mathematical Approach to Motion Coordination Algorithms*. Princeton University Press, 2009.

[25] Caicedo-Nunez, C., and Zefran, M. Consensus-based rendezvous. In *2008 IEEE International Conference on Control Applications* (2008), IEEE, pp. 1031–1036.

[26] Caicedo-Nunez, C., and Zefran, M. Probabilistic guarantees for rendezvous under noisy measurements. In *2009 American Control Conference* (2009), IEEE, pp. 5180–5185.

[27] Cao, M., Morse, A. S., and Anderson, B. D. O. Reaching a consensus in a dynamically changing environment: A graphical approach. *SIAM journal on control and optimization 47*, 2 (2008), 575–600.

[28] Cao, Y., Stuart, D., Ren, W., and Meng, Z. Distributed containment control for multiple autonomous vehicles with double-integrator dynamics: Algorithms and experiments. *IEEE Transactions on Control Systems Technology 19*, 4 (July 2011), 929–938.

[29] Cao, Y., Yu, W., Ren, W., and Chen, G. An overview of recent progress in the study of distributed multi-agent coordination. *IEEE Transactions on Industrial Informatics 9*, 1 (Feb 2013), 427–438.

[30] Cardona, G. A., and Calderon, J. M. Robot swarm navigation and victim detection using rendezvous consensus in search and rescue operations. *Applied sciences 9*, 8 (2019), 1702–.

[31] Chang, T.-H., Hong, M., Wai, H.-T., Zhang, X., and Lu, S. Distributed Learning in the Nonconvex World: From batch data to streaming and beyond. *IEEE Signal Processing Magazine 37*, 3 (May 2020), 26–38.

[32] Charron-Bost, B., and Lambein-Monette, P. Randomization and quantization for average consensus. *arXiv:1804.10919 [cs]* (Apr. 2018). arXiv: 1804.10919.

[33] Chatzipanagiotis, N., and Zavlanos, M. M. A Distributed Algorithm for Convex Constrained Optimization Under Noise. *IEEE Transactions on Automatic Control 61*, 9 (Sept. 2016), 2496–2511.

[34] Chen, A., and Ozdaglar, A. A fast distributed proximal-gradient method. pp. 601–608.

[35] Chen, J., and Sayed, A. H. Diffusion adaptation strategies for distributed optimization and learning over networks. *IEEE Transactions on Signal Processing 60*, 8 (Aug 2012), 4289–4305.

[36] Cohen, R., and Peleg, D. Convergence properties of the gravitational algorithm in asynchronous robot systems. *SIAM journal on computing 34*, 6 (2005), 1516–1528.

[37] Conte, G., and Pennesi, P. The rendezvous problem with discontinuous control policies. *IEEE transactions on automatic control 55*, 1 (2010), 279–283.

[38] Cortes, J., Martinez, S., and Bullo, F. Robust rendezvous for mobile autonomous agents via proximity graphs in arbitrary dimensions. *IEEE transactions on automatic control 51*, 8 (2006), 1289–1298.

[39] Czyzowicz, J., Gasieniec, L., Gorry, T., Kranakis, E., Martin, R., and Pajak, D. Evacuating robots via unknown exit in a disk. In *Distributed Computing* (Berlin, Heidelberg, 2014), F. Kuhn, Ed., Springer Berlin Heidelberg, pp. 122–136.

[40] Daneshmand, A., Scutari, G., and Kungurtsev, V. Second-order guarantees of distributed gradient algorithms. *SIAM Journal on Optimization 30*, 4 (2020), 3029–3068.

[41] Dean, J., Corrado, G., Monga, R., Chen, K., Devin, M., Mao, M., aurelio Ranzato, M., Senior, A., Tucker, P., Yang, K., Le, Q. V., and Ng, A. Y. Large scale distributed deep networks. In *Advances in Neural*

Information Processing Systems 25, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1223–1231.

[42] Deng, L. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine 29*, 6 (2012), 141–142.

[43] Di Lorenzo, P., and Scutari, G. Next: In-network nonconvex optimization. *IEEE Transactions on Signal and Information Processing over Networks 2*, 2 (2016), 120–136.

[44] Dimakis, A. G., Kar, S., Moura, J. M. F., Rabbat, M. G., and Scaglione, A. Gossip algorithms for distributed signal processing. *Proceedings of the IEEE 98*, 11 (Nov 2010), 1847–1864.

[45] Dimakis, A. G., Sarwate, A. D., and Wainwright, M. J. Geographic Gossip: Efficient Aggregation for Sensor Networks. *Proceedings of International Conference on Information Processing in Sensor Networks* (2006), 69–76.

[46] Doan, T. T., Maguluri, S. T., and Romberg, J. Convergence rates of distributed gradient methods under random quantization: A stochastic approximation approach. *IEEE Transactions on Automatic Control* (2020), 1–1.

[47] Doan, T. T., Maguluri, S. T., and Romberg, J. Fast convergence rates of distributed subgradient methods with adaptive quantization. *IEEE Transactions on Automatic Control* (2020), 1–1.

[48] Dong, Y., and Xu, S. Rendezvous with connectivity preservation problem of linear multiagent systems via parallel event-triggered control strategies. *IEEE transactions on cybernetics 52*, 5 (2022), 2725–2734.

[49] Dua, D., and Graff, C. UCI machine learning repository, 2017.

[50] Duchi, J., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research 12* (07 2011), 2121–2159.

[51] Duchi, J. C., Agarwal, A., and Wainwright, M. J. Dual averaging for distributed optimization: Convergence analysis and network scaling. *IEEE Transactions on Automatic Control 57*, 3 (March 2012), 592–606.

[52] Eisen, M., Mokhtari, A., and Ribeiro, A. Decentralized quasi-newton methods. *IEEE Transactions on Signal Processing PP* (04 2016).

[53] El Chamie, M., Liu, J., and Başar, T. Design and analysis of distributed averaging with quantized communication. *IEEE Transactions on Automatic Control 61*, 12 (Dec 2016), 3870–3884.

[54] El Chamie, M., Neglia, G., and Avrachenkov, K. Reducing communication overhead for average consensus. In *2013 IFIP Networking Conference* (May 2013), pp. 1–9.

[55] Erdős, P., and Rényi, A. On random graphs. *Publicationes Mathematicae 6* (1959), 290–297.

[56] Falcao, D. M., Wu, F. F., and Murphy, L. Parallel and distributed state estimation. *IEEE Transactions on Power Systems 10*, 2 (May 1995), 724–730.

[57] Fallah, A., Gurbuzbalaban, M., Ozdaglar, A., Simsekli, U., and Zhu, L. Robust distributed accelerated stochastic gradient methods for multi-agent networks, 2019.

[58] Fang, L., and Antsaklis, P. Asynchronous consensus protocols using nonlinear paracontractions theory. *IEEE transactions on automatic control 53*, 10 (2008), 2351–2355.

[59] Feng Zhao, Jaewon Shin, and Reich, J. Information-driven dynamic sensor collaboration. *IEEE Signal Processing Magazine 19*, 2 (March 2002), 61–72.

[60] Fercoq, O., Qu, Z., Richtárik, P., and Takáč, M. Fast distributed coordinate descent for non-strongly convex losses. In *2014 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)* (Sep. 2014), pp. 1–6.

[61] Giannakis, G. B., Kekatos, V., Gatsis, N., Kim, S., Zhu, H., and Wollenberg, B. F. Monitoring and optimization for power grids: A signal processing perspective. *IEEE Signal Processing Magazine 30*, 5 (Sep. 2013), 107–128.

[62] Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., and He, K. Accurate, large minibatch sgd: Training imagenet in 1 hour.

[63] Hajinezhad, D., Hong, M., and Garcia, A. ZONE: Zeroth-Order Nonconvex Multiagent Optimization Over Networks. *IEEE Transactions on Automatic Control 64*, 10 (Oct. 2019), 3995–4010.

[64] Hong, M. A Distributed, Asynchronous, and Incremental Algorithm for Nonconvex Optimization: An ADMM Approach. *IEEE TRANSACTIONS ON CONTROL OF NETWORK SYSTEMS 5*, 3 (2018), 11.

[65] Hong, M., and Chang, T.-H. Stochastic Proximal Gradient Consensus Over Random Networks. *IEEE Transactions on Signal Processing 65*, 11 (June 2017), 2933–2948.

[66] Hong, M., Razaviyayn, M., and Lee, J. Gradient primal-dual algorithm converges to second-order stationary solution for nonconvex distributed optimization over networks. In *Proceedings of the 35th International Conference on Machine Learning* (Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018), J. Dy and A. Krause, Eds., vol. 80 of *Proceedings of Machine Learning Research*, PMLR, pp. 2009–2018.

[67] Hong, M., Razaviyayn, M., Luo, Z., and Pang, J. A unified algorithmic framework for block-structured optimization involving big data: With applications in machine learning and signal processing. *IEEE Signal Processing Magazine 33*, 1 (2016), 57–77.

[68] Hong, M., Zeng, S., Zhang, J., and Sun, H. On the Divergence of Decentralized Non-Convex Optimization. *arXiv:2006.11662 [cs, math]* (June 2020). arXiv: 2006.11662.

[69] Horn, R. A., and Johnson, C. R. *Matrix analysis*, 2nd ed ed. Cambridge University Press, Cambridge ; New York, 2012.

[70] Iakovidou, C., and Wei, E. Nested distributed gradient methods with stochastic computation errors. *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)* (2019), 339–346.

[71] Irisarri, G., Wang, X., Tong, J., and Mokhtari, S. Maximum loadability of power systems using interior point nonlinear optimization method. *IEEE transactions on Power Systems 12*, 1 (1997), 162–172.

[72] Iutzeler, F., Bianchi, P., Ciblat, P., and Hachem, W. Asynchronous distributed optimization using a randomized alternating direction method of multipliers. In *52nd IEEE Conference on Decision and Control* (Dec 2013), pp. 3671–3676.

[73] Jadbabaie, A., Ozdaglar, A., and Zargham, M. 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* (Dec 2009), pp. 2736–2741.

[74] Jakovetic, D., Xavier, J., and Moura, J. M. F. Fast Distributed Gradient Methods. *IEEE Transactions on Automatic Control 59*, 5 (2014), 1131–1146.

[75] Johansson, B., Rabi, M., and Johansson, M. A simple peer-to-peer algorithm for distributed optimization in sensor networks. 4705–4710.

[76] Kar, S., and Moura, J. M. F. Distributed Consensus Algorithms in Sensor Networks With Imperfect Communication: Link Failures and Channel Noise. *IEEE Transactions on Signal Processing 57*, 1 (2009), 355–369.

[77] Kashyap, A., Basar, T., and Srikant, R. Quantized Consensus. In *2006 IEEE International Symposium on Information Theory* (Seattle, WA, July 2006), IEEE, pp. 635–639.

[78] Kekatos, V., and Giannakis, G. B. Distributed robust power system state estimation. *IEEE Transactions on Power Systems 28*, 2 (2013), 1617–1626.

[79] Kia, S. S., Cortés, J., and Martínez, S. Distributed convex optimization via continuous-time coordination algorithms with discrete-time communication. *Automatica 55* (2015), 254 – 264.

[80] Kingma, D., and Ba, J. Adam: A method for stochastic optimization. *International Conference on Learning Representations* (12 2014).

[81] Konečný, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., and Bacon, D. Federated learning: Strategies for improving communication efficiency, 2016.

[82] Lan, G., Lee, S., and Zhou, Y. Communication-efficient algorithms for decentralized and stochastic optimization. *Mathematical Programming 180*, 1 (Mar. 2020), 237–284.

[83] Lavei, J., Rantzer, A., and Low, S. Power flow optimization using positive quadratic programming. *IFAC Proceedings Volumes 44*, 1 (2011), 10481–10486.

[84] Lee, C.-S., Michelusi, N., and Scutari, G. Finite rate quantized distributed optimization with geometric convergence. In *2018 52nd Asilomar Conference on Signals, Systems, and Computers* (2018), IEEE, pp. 1876–1880.

[85] Lee, J. D., Panageas, I., Piliouras, G., Simchowitz, M., Jordan, M. I., and Recht, B. First-order methods almost always avoid strict saddle points. *Mathematical Programming 176*, 1 (July 2019), 311–337.

[86] Li, B., Page, B. R., Hoffman, J., Moridian, B., and Mahmoudian, N. Rendezvous planning for multiple auvs with mobile charging stations in dynamic currents. *IEEE Robotics and Automation Letters 4*, 2 (2019), 1653–1660.

[87] Li, H., Zheng, L., Wang, Z., Yan, Y., Feng, L., and Guo, J. S-diging: A stochastic gradient tracking algorithm for distributed optimization. *IEEE Transactions on Emerging Topics in Computational Intelligence* (2020), 1–13.

[88] Li, J., Chen, G., Wu, Z., and He, X. Distributed subgradient method for multi-agent optimization with quantized communication: J. LI *ET AL. Mathematical Methods in the Applied Sciences 40*, 4 (Mar. 2017), 1201–1213.

[89] Li, M., Zhang, T., Chen, Y., and Smola, A. J. Efficient mini-batch training for stochastic optimization. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY, USA, 2014), KDD '14, Association for Computing Machinery, p. 661–670.

[90] Lian, X., Zhang, C., Zhang, H., Hsieh, C.-J., Zhang, W., and Liu, J. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. In *Advances in Neural Information Processing Systems* (2017), I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30, Curran Associates, Inc., pp. 5330–5340.

[91] Lian, X., Zhang, W., Zhang, C., and Liu, J. Asynchronous Decentralized Parallel Stochastic Gradient Descent. *arXiv:1710.06952 [cs, math, stat]* (Oct. 2017). arXiv: 1710.06952.

[92] Lin, J., Morse, A. S., and Anderson, B. D. O. The multi-agent rendezvous problem. part 1: The synchronous case. *SIAM journal on control and optimization 46*, 6 (2007), 2096–2119.

[93] Lin, J., Morse, A. S., and Anderson, B. D. O. The multi-agent rendezvous problem. part 2: The asynchronous case. *SIAM journal on control and optimization 46*, 6 (2007), 2120–2147.

[94] Ling, Q., and Tian, Z. Decentralized sparse signal recovery for compressive sleeping wireless sensor networks. *IEEE Transactions on Signal Processing 58*, 7 (July 2010), 3816–3827.

[95] Lorenzo, P. D., and Scutari, G. Next: In-network nonconvex optimization. *IEEE Transactions on Signal and Information Processing over Networks 2*, 2 (2016), 120–136.

[96] Low, S. H. Convex relaxation of optimal power flow—part i: Formulations and equivalence. *IEEE Transactions on Control of Network Systems 1*, 1 (2014), 15–27.

[97] Mansoori, F., and Wei, E. Superlinearly convergent asynchronous distributed network newton method. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)* (Dec 2017), pp. 2874–2879.

[98] Mansoori, F., and Wei, E. A flexible framework of first-order primal-dual algorithms for distributed optimization, 2019.

[99] Mansoori, F., and Wei, E. A general framework of exact primal-dual first order algorithms for distributed optimization, 2019.

[100] Mathew, N., Smith, S. L., and Waslander, S. L. Multirobot rendezvous planning for recharging in persistent tasks. *IEEE transactions on robotics 31*, 1 (2015), 128–142.

[101] McMahan, H. B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data, 2016.

[102] Minghui Zhu, and Martinez, S. On the convergence time of distributed quantized averaging algorithms. In *2008 47th IEEE Conference on Decision and Control* (Cancun, Mexico, 2008), IEEE, pp. 3971–3976.

[103] Mokhtari, A., Ling, Q., and Ribeiro, A. Network newton distributed optimization methods. *IEEE Transactions on Signal Processing 65*, 1 (Jan 2017), 146–161.

[104] Mokhtari, A., and Ribeiro, A. Dsa: Decentralized double stochastic averaging gradient algorithm. *J. Mach. Learn. Res. 17*, 1 (Jan. 2016), 2165–2199.

[105] Molzahn, D. K., Dörfler, F., Sandberg, H., Low, S. H., Chakrabarti, S., Baldick, R., and Lavaei, J. A survey of distributed optimization and control algorithms for electric power systems. *IEEE Transactions on Smart Grid 8*, 6 (Nov 2017), 2941–2962.

[106] Morral, G., Bianchi, P., and Fort, G. Success and Failure of Adaptation-Diffusion Algorithms With Decaying Step Size in Multiagent Networks. *IEEE Transactions on Signal Processing 65*, 11 (June 2017), 2798–2813.

[107] Mota, J., Xavier, J., Aguiar, P., and Püschel, M. D-admm: A communication-efficient distributed algorithm for separable optimization. *IEEE Transactions on Signal Processing 61* (02 2012).

[108] Mu, B., Zhang, K., Xiao, F., and Shi, Y. Event-based rendezvous control for a group of robots with asynchronous periodic detection and communication time delays. *IEEE transactions on cybernetics 49*, 7 (2019), 2642–2651.

[109] Nedic, A., Olshevsky, A., Ozdaglar, A., and Tsitsiklis, J. On Distributed Averaging Algorithms and Quantization Effects. *IEEE Transactions on Automatic Control 54*, 11 (Nov. 2009), 2506–2517.

[110] Nedic, A., Olshevsky, A., Ozdaglar, A., and Tsitsiklis, J. N. Distributed subgradient methods and quantization effects. In *2008 47th IEEE Conference on Decision and Control* (Cancun, Mexico, 2008), IEEE, pp. 4177–4184.

[111] Nedić, A., and Ozdaglar, A. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control 54*, 1 (Jan 2009), 48 –61.

[112] Nedic, A., Ozdaglar, A., and Parrilo, P. Constrained Consensus and Optimization in Multi-Agent Networks. *IEEE Transactions on Automatic Control 55*, 4 (Apr. 2010), 922–938.

[113] Nedich, A., Olshevsky, A., and Shi, A. Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM Journal on Optimization 27*, 4 (1 2017), 2597–2633.

[114] Nedić, A., and Olshevsky, A. Stochastic gradient-push for strongly convex functions on time-varying directed graphs. *IEEE Transactions on Automatic Control 61*, 12 (2016), 3936–3947.

[115] Nesterov, Y. Introductory Lectures on Convex Programming Volume I: Basic course. 212.

[116] Oh, H., Kim, S., Shin, H.-S., White, B. A., Tsourdos, A., and Rabbath, C. A. Rendezvous and standoff target tracking guidance using differential geometry. *Journal of intelligent & robotic systems 69*, 1-4 (2012), 389–405.

[117] Oh, S., Zelinsky, A., Taylor, K., et al. Autonomous battery recharging for indoor mobile robots. In *Proceedings of the australian conference on robotics and automation* (2000).

[118] Olfati-Saber, R., Fax, J. A., and Murray, R. M. Consensus and cooperation in networked multi-agent systems. *Proceedings of the IEEE 95*, 1 (2007), 215–233.

[119] Parasuraman, R., Kim, J., Luo, S., and Min, B.-C. Multipoint rendezvous in multirobot systems. *IEEE transactions on cybernetics 50*, 1 (2020), 310–323.

[120] Parker, C., Morrison, I., and Sutanto, D. Application of an optimisation method for determining the reactive margin from voltage collapse in reactive power planning. *IEEE Transactions on Power Systems 11*, 3 (1996), 1473–1481.

[121] Peng, Z., Xu, Y., Yan, M., and Yin, W. Arock: an algorithmic framework for asynchronous parallel coordinate updates. *SIAM Journal on Scientific Computing 38*, 5 (2016), A2851–A2879.

[122] Predd, J. B., Kulkarni, S. B., and Poor, H. V. Distributed learning in wireless sensor networks. *IEEE Signal Processing Magazine 23*, 4 (July 2006), 56–69.

[123] Predd, J. B., Kulkarni, S. R., and Poor, H. V. A collaborative training algorithm for distributed learning. *IEEE Transactions on Information Theory 55*, 4 (April 2009), 1856–1871.

[124] Pu, S., and Garcia, A. A Flocking-based Approach for Distributed Stochastic Optimization. *arXiv:1709.07085 [math]* (Sept. 2017). arXiv: 1709.07085.

[125] Pu, S., and Nedić, A. Distributed stochastic gradient tracking methods. *Mathematical programming 187*, 1-2 (2020), 409–457.

[126] Pu, S., Olshevsky, A., and Paschalidis, I. C. Asymptotic network independence in distributed stochastic optimization for machine learning: Examining distributed and centralized stochastic gradient descent. *IEEE Signal Processing Magazine 37*, 3 (2020), 114–122.

[127] Pu, S., Olshevsky, A., and Paschalidis, I. C. A sharp estimate on the transient time of distributed stochastic gradient descent, 2020.

[128] Pu, Y., Zeilinger, M. N., and Jones, C. N. Quantization Design for Distributed Optimization. *IEEE Transactions on Automatic Control 62*, 5 (May 2017), 2107–2120.

[129] Purchala, K., Meeus, L., Van Dommelen, D., and Belmans, R. Usefulness of dc power flow for active power flow analysis. In *Power Engineering Society General Meeting, 2005. IEEE* (2005), IEEE, pp. 454–459.

[130] Qu, G., and Li, N. Harnessing smoothness to accelerate distributed optimization. *IEEE Transactions on Control of Network Systems PP* (04 2017), 1–1.

[131] Rabbat, M., and Nowak, R. Distributed optimization in sensor networks. In *Proceedings of the 3rd international symposium on Information processing in sensor networks* (2004), ACM, pp. 20–27.

[132] Rabbat, M., and Nowak, R. Quantized incremental algorithms for distributed optimization. *IEEE Journal on Selected Areas in Communications 23*, 4 (Apr. 2005), 798–808.

[133] Ram, S. S., Nedić, A., and Veeravalli, V. V. Asynchronous gossip algorithms for stochastic optimization. In *Decision and Control, 2009 held jointly with the 2009 28th Chinese Control Conference. CDC/CCC 2009. Proceedings of the 48th IEEE Conference on* (2009), IEEE, pp. 3581–3586.

[134] Ram, S. S., Nedić, A., and Veeravalli, V. V. Asynchronous gossip algorithm for stochastic optimization: Constant stepsize analysis. In *Recent Advances in Optimization and its Applications in Engineering*. Springer, 2010, pp. 51–60.

[135] Reisizadeh, A., Mokhtari, A., Hassani, H., and Pedarsani, R. An exact quantized decentralized gradient descent algorithm. *IEEE Transactions on Signal Processing 67*, 19 (2019), 4934–4947.

[136] Ren, W., Beard, R. W., and Atkins, E. M. Information consensus in multivehicle cooperative control. *IEEE Control Systems Magazine 27*, 2 (April 2007), 71–82.

[137] Richtárik, P., and Takáč, M. Parallel coordinate descent methods for big data optimization. *Mathematical Programming 156*, 1-2 (2016), 433–484.

[138] Roy, N., and Dudek, G. Collaborative robot exploration and rendezvous: Algorithms, performance bounds and observations. *Autonomous Robots 11*, 2 (2001), 117–136.

[139] Rubenstein, M., Ahler, C., and Nagpal, R. Kilobot: A low cost scalable robot system for collective behaviors. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on* (2012), IEEE, pp. 3293–3298.

[140] Rucco, A., Sujit, P. B., Aguiar, A. P., Borges de Sousa, J., and Lobo Pereira, F. Optimal rendezvous trajectory for unmanned aerial-ground vehicles. *IEEE transactions on aerospace and electronic systems 54*, 2 (2018), 834–847.

[141] Sayed, A. Diffusion adaptation over networks. *Academic Press Library in Signal Processing 3* (05 2012).

[142] Schizas, I. D., Ribeiro, A., and Giannakis, G. B. Consensus in ad hoc wsns with noisy links—part i: Distributed estimation of deterministic signals. *IEEE Transactions on Signal Processing 56*, 1 (Jan 2008), 350–364.

[143] Scutari, G., and Sun, Y. Distributed nonconvex constrained optimization over time-varying digraphs. *Math. Program. 176*, 1–2 (July 2019), 497–544.

[144] Shamir, O., and Srebro, N. Distributed stochastic optimization and learning. In *2014 52nd Annual Allerton Conference on Communication, Control, and Computing (Allerton)* (Monticello, IL, USA, Sept. 2014), IEEE, pp. 850–857.

[145] Shen, Z., Mokhtari, A., Zhou, T., Zhao, P., and Qian, H. Towards more efficient stochastic decentralized learning: Faster convergence and sparse communication. In *Proceedings of the 35th International Conference on Machine Learning* (Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018), J. Dy and A. Krause, Eds., vol. 80 of *Proceedings of Machine Learning Research*, PMLR, pp. 4624–4633.

[146] Shi, W., Ling, Q., Wu, G., and Yin, W. Extra: An exact first-order algorithm for decentralized consensus optimization. *SIAM Journal on Optimization 25*, 2 (2015), 944–966.

[147] Smith, V., Forte, S., Ma, C., Takáč, M., Jordan, M. I., and Jaggi, M. CoCoA: A General Framework for Communication-Efficient Distributed Optimization. 49.

[148] Spiridonoff, A., Olshevsky, A., and Paschalidis, I. C. Robust asynchronous stochastic gradient-push: Asymptotically optimal and network-independent performance for strongly convex functions. *Journal of Machine Learning Research 21*, 58 (2020), 1–47.

[149] Srivastava, K., and Nedic, A. Distributed Asynchronous Constrained Stochastic Optimization. *IEEE Journal of Selected Topics in Signal Processing 5*, 4 (Aug. 2011), 772–790.

[150] Su, H., Wang, X., and Chen, G. Rendezvous of multiple mobile agents with preserved network connectivity. *Systems & control letters 59*, 5 (2010), 313–322.

[151] Sun, A. X., Phan, D. T., and Ghosh, S. Fully decentralized ac optimal power flow algorithms. In *Power and Energy Society General Meeting (PES), 2013 IEEE* (2013), IEEE, pp. 1–5.

[152] Sun, H., and Hong, M. Distributed non-convex first-order optimization and information processing: Lower complexity bounds and rate optimal algorithms. In *2018 52nd Asilomar Conference on Signals, Systems, and Computers* (2018), pp. 38–42.

[153] Sun, H., Lu, S., and Hong, M. Improving the sample and communication complexity for decentralized non-convex optimization: Joint gradient estimation and tracking. In *Proceedings of the 37th International Conference on Machine Learning* (13–18 Jul 2020), H. D. III and A. Singh, Eds., vol. 119 of *Proceedings of Machine Learning Research*, PMLR, pp. 9217–9228.

[154] Sundhar Ram, S., Nedić, A., and Veeravalli, V. V. Distributed Stochastic Subgradient Projection Algorithms for Convex Optimization. *Journal of Optimization Theory and Applications 147*, 3 (Dec. 2010), 516–545.

[155] Swenson, B., Murray, R., Kar, S., and Poor, H. V. Distributed stochastic gradient descent: Nonconvexity, nonsmoothness, and convergence to local minima. *arXiv:2003.02818 [math.OC]* (Aug. 2020).

[156] Swenson, B., Murray, R., Poor, H. V., and Kar, S. Distributed gradient flow: Nonsmoothness, nonconvexity, and saddle point evasion. *arXiv:2008.05387 [math.OC]* (Aug. 2020).

[157] Tang, H., Lian, X., Yan, M., Zhang, C., and Liu, J. d^2: Decentralized training over decentralized data. In *Proceedings of the 35th International Conference on Machine Learning* (Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018), J. Dy and A. Krause, Eds., vol. 80 of *Proceedings of Machine Learning Research*, PMLR, pp. 4848–4856.

[158] Tang, Y., and Li, N. Distributed zero-order algorithms for nonconvex multi-agent optimization. In *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)* (2019), pp. 781–786.

[159] Tatarenko, T., and Touri, B. Non-convex distributed optimization. *IEEE Transactions on Automatic Control 62*, 8 (2017), 3744–3757.

[160] Towfic, Z. J., and Sayed, A. H. Adaptive Penalty-Based Distributed Stochastic Convex Optimization. *IEEE Transactions on Signal Processing 62*, 15 (Aug. 2014), 3924–3938.

[161] Tsianos, K. I., Lawlor, S., and Rabbat, M. G. Consensus-based distributed optimization: Practical issues and applications in large-scale machine learning. In

2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton) (Oct 2012), pp. 1543–1550.

[162] Tsitsiklis, J. *Problems in Decentralized Decision Making and Computation*. Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, 1984.

[163] Tsitsiklis, J., Bertsekas, D., and Athans, M. Distributed asynchronous deterministic and stochastic gradient optimization algorithms. *IEEE Transactions on Automatic Control 31*, 9 (Sep. 1986), 803–812.

[164] Tucker, H. G. *A graduate course in probability*. Academic Press, 1967.

[165] Vogels, T., Karimireddy, S. P., and Jaggi, M. Powersgd: Practical low-rank gradient compression for distributed optimization. In *Advances in Neural Information Processing Systems* (2019), H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32, Curran Associates, Inc., pp. 14259–14268.

[166] Wangni, J., Wang, J., Liu, J., and Zhang, T. Gradient sparsification for communication-efficient distributed optimization. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (Red Hook, NY, USA, 2018), NIPS'18, Curran Associates Inc., p. 1306–1316.

[167] Wei, E., and Ozdaglar, A. Distributed alternating direction method of multipliers. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)* (Dec 2012), pp. 5445–5450.

[168] Wei, E., and Ozdaglar, A. On the o(1/k) convergence of asynchronous distributed alternating direction method of multipliers. *2013 IEEE Global Conference on Signal and Information Processing, GlobalSIP 2013 - Proceedings* (07 2013).

[169] Wei, E., Ozdaglar, A., and Jadbabaie, A. A distributed newton method for network utility maximization–i: Algorithm. *IEEE Transactions on Automatic Control 58*, 9 (Sep. 2013), 2162–2175.

[170] Welford, B. P. Note on a method for calculating corrected sums of squares and products. *Technometrics 4*, 3 (1962), 419–420.

[171] Xiao, F., and Wang, L. Asynchronous rendezvous analysis via set-valued consensus theory. *SIAM journal on control and optimization 50*, 1 (2012), 196–221.

[172] Xiao, L., Boyd, S., and Kim, S.-J. Distributed average consensus with least-mean-square deviation. *Journal of Parallel and Distributed Computing 67*, 1 (2007), 33–46.

[173] Xiao, L., Boyd, S., and Lall, S. A scheme for robust distributed sensor fusion based on average consensus. In *Proceedings of the 4th international symposium on Information processing in sensor networks* (2005), IEEE Press, p. 9.

[174] Xiao, Y., Bandi, C., and Wei, E. Supply function equilibrium in power markets: Mesh networks. In *Signal and Information Processing (GlobalSIP), 2016 IEEE Global Conference on* (2016), IEEE, pp. 861–865.

[175] Xin, R., Khan, U. A., and Kar, S. Variance-reduced decentralized stochastic optimization with accelerated convergence. *IEEE Transactions on Signal Processing 68* (2020), 6255–6271.

[176] Xu, J., Tian, Y., Sun, Y., and Scutari, G. Accelerated primal-dual algorithms for distributed smooth convex optimization over networks, 2019.

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

[178] Yang, T. Trading computation for communication: Distributed stochastic dual coordinate ascent. In *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2013, pp. 629–637.

[179] Yi, P., and Hong, Y. Quantized Subgradient Algorithm and Data-Rate Analysis for Distributed Optimization. *IEEE Transactions on Control of Network Systems 1*, 4 (Dec. 2014), 380–392.

[180] Yi, P., Hong, Y., and Liu, F. Initialization-free distributed algorithms for optimal resource allocation with feasibility constraints and application to economic dispatch of power systems. *Automatica 74* (2016), 259 – 269.

[181] Yordanova, V., and Griffiths, H. Synchronous rendezvous technique for multi-vehicle mine countermeasure operations. In *OCEANS 2015 - MTS/IEEE Washington* (2015), pp. 1–6.

[182] Yuan, D., Xu, S., Zhao, H., and Rong, L. Distributed dual averaging method for multi-agent optimization with quantized communication. *Systems & Control Letters 61*, 11 (Nov. 2012), 1053–1061.

[183] Yuan, K., Xu, W., and Ling, Q. Can primal methods outperform primal-dual methods in decentralized dynamic optimization? *IEEE Transactions on Signal Processing 68* (2020), 4466–4480.

[184] Yuan, K., Ying, B., Zhao, X., and Sayed, A. H. Exact diffusion for distributed optimization and learning—part i: Algorithm development. *IEEE Transactions on Signal Processing 67*, 3 (Feb 2019), 708–723.

[185] Zegers, F. M., Guralnik, D. P., and Dixon, W. E. Event/self-triggered multi-agent system rendezvous with graph maintenance. In *2021 60th IEEE Conference on Decision and Control (CDC)* (2021), pp. 1886–1891.

[186] Zeng, J., and Yin, W. On nonconvex decentralized gradient descent. *IEEE Transactions on Signal Processing 66*, 11 (2018), 2834–2848.

[187] Zhou, K., and Roumeliotis, S. I. Multirobot active target tracking with combinations of relative observations. *IEEE Transactions on Robotics 27*, 4 (Aug 2011), 678–695.

[188] Zhu, M., and Martinez, S. On distributed convex optimization under inequality and equality constraints. *IEEE Transactions on Automatic Control 57*, 1 (Jan 2012), 151–164.

[189] Zhu, S., Soh, Y. C., and Xie, L. Distributed parameter estimation with quantized communication via running average. *IEEE Transactions on Signal Processing 63*, 17 (2015), 4634–4646.

[190] Zinkevich, M., Weimer, M., Smola, A., and Li, L. Parallelized stochastic gradient descent. vol. 23, pp. 2595–2603.